

TURING TEST AND CONVERSATION

A THESIS

SUBMITTED TO THE DEPARTMENT OF COMPUTER
ENGINEERING AND INFORMATION SCIENCE
AND THE INSTITUTE OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By

Ayşe Pınar Saygın

July, 1999

6

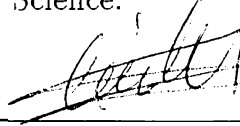
341

·529

1999

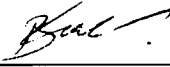
B⁰49041

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.



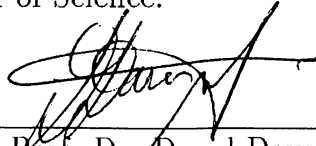
Asst. Prof. Dr. İlyas Çiçekli(Principal Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.



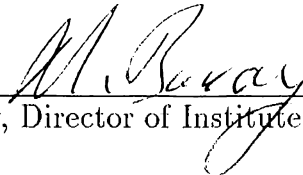
Asst. Prof. Dr. Bilge Say

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.



Asst. Prof. Dr. David Davenport

Approved for the Institute of Engineering and Science:



Prof. Dr. Mehmet Baray, Director of Institute of Engineering and Science

ABSTRACT

TURING TEST AND CONVERSATION

Ayşe Pınar Saygın

M.S. in Computer Engineering and Information Science

Supervisor: Asst. Prof. Dr. İlyas Çiçekli

July, 1999

The Turing Test is one of the most disputed topics in Artificial Intelligence, Philosophy of Mind and Cognitive Science. It has been proposed 50 years ago, as a method to determine whether machines can think or not. It embodies important philosophical issues, as well as computational ones. Moreover, because of its characteristics, it requires interdisciplinary attention. The Turing Test posits that, to be granted intelligence, a computer should imitate human conversational behavior so well that it should be indistinguishable from a real human being. From this, it follows that conversation is a crucial concept in its study. Surprisingly, focusing on conversation in relation to the Turing Test has not been a prevailing approach in previous research. This thesis first provides a thorough and deep review of the 50 years of the Turing Test. Philosophical arguments, computational concerns, and repercussions in other disciplines are all discussed. Furthermore, this thesis studies the Turing Test as a special kind of conversation. In doing so, the relationship between existing theories of conversation and human-computer communication is explored. In particular, Grice's cooperative principle and conversational maxims are concentrated on. Viewing the Turing Test as conversation and computers as language users have significant effects on the way we look at Artificial Intelligence, and on communication in general.

Key words: Turing Test, Artificial Intelligence, Conversational maxims, Cooperative Principle, Pragmatics, Natural Language Conversation Systems, Chatterbots, Conversation Analysis, Cognitive Science, Philosophy of Language, Computational Linguistics

ÖZET

TURING TESTİ VE KONUŞMA

Ayşe Pınar Saygın

Bilgisayar ve Enformatik Mühendisliği, Yüksek Lisans

Danışman: Yrd. Doç. Dr. İlyas Çiçekli

Temmuz, 1999

Turing Testi Yapay Zeka, Dil Felsefesi ve Bilişsel Bilimler alanlarında çok tartışılan konulardan biridir. 50 yıl önce, makinelerin düşünüp düşünmediğini ölçmek için kullanılacak bir test olarak öne sürülmüştür. Bünyesinde, hem felsefe hem de bilgisayar bilimi açısından önemli olan kavramları barındırır. Ayrıca, kendine has özelliklerinden dolayı, disiplinler arası bir yaklaşım gerektirmektedir. Turing Testi'ne göre bir bilgisayara zeki diyebilmemiz için, onun insan konuşma davranışlarını gerçek bir insandan ayırdedilemeyecek kadar iyi taklit edebilmesi gerekir. Buradan da görülebileceği gibi, konuşma, Turing Testi'nin çok önemli bir parçasıdır. Ama şaşırtıcı bir şekilde, testle ilgili önceki yorumlar konuya bu açıdan yaklaşmamaktadır. Bu tez, öncelikle Turing Testi'nin geniş ve derin bir incelemesini sunmaktadır. Felsefi tartışmalara, pratik gelişmelere, ve konunun diğer bilimlerde yarattığı yankılara yer verilmiştir. Ayrıca, Turing Testi bir çeşit konuşma olarak ele alınmaktadır. Halen varolan konuşma teorileri ile bilgisayar-insan iletişimi arasındaki ilişki incelenmiştir. Özellikle Grice'in işbirliği ilkesi ve konuşma ilkeleri üzerine yoğunlaşmıştır. Turing Testi'ni bir çeşit konuşma olarak, ve bilgisayarları dil kullanıcıları olarak görmek, hem Yapay Zeka'ya, hem de genel olarak iletişime bakış açımız üzerinde büyük etkiye sahiptir.

Anahtar kelimeler: Turing Testi, Yapay Zeka, Konuşma İlkeleri, İşbirliği İlkesi, Edimbilim, Doğal Dil Konuşma Sistemleri, Otomatik Gevezeler, Konuşma İncelemesi, Bilişsel Bilimler, Dil Felsefesi, Bilgisayarlı Dilbilim

To my mother, my father, my sister, and Musti...

ACKNOWLEDGMENTS

I would like to express my gratitude to my supervisor Dr. İlyas Çiçekli for his guidance, kind support and motivation during this study.

I would also like to thank Dr. David Davenport and Dr. Bilge Say for the valuable comments they made on this thesis.

Dr. Giray Uraz from Hacettepe University has been very helpful in carrying out the preliminary open-ended questionnaires. I am also thankful to friends and family, notably Hülya Saygın, Funda Saygın, Bilge Say, Oytun Öztürk, Emel Aydın, Will Turner, Stephen Wilson, Gülayşe İnce and Evrim Dener, for their help in conducting the surveys.

My mother has directly and indirectly contributed to this thesis in more ways than can be listed here. I could never pay her back the assistance and caring she has provided. I am indebted to my father for encouraging and supporting me in all of my “scientific” endeavours and helping me maintain my childhood curiosity about the world. My sister has been a best friend to me during the last few years. Without her organizational skills, I could still be struggling among heaps of paper, trying to process the survey data for this thesis. I would also like to thank my grandmother for her prayers and to Musti for taking care of our family.

Everything is possible with a little help from friends. I am especially thankful to Emel for her endless patience and understanding, Boğa for his companionship and support, Yücel for cheering me up and being my “partner”, Tuba for being an excellent officemate, Tamer for setting an example to us all by being the honest and hardworking person he is, Okyay for the enjoyable discussions and for “otomatik gevezeler”, Aysel and Mustafa for the laughs, Esin, Deniz, Hüseyin, Will, Evrim, Nihan, Stephen and Sinan for their friendship, and last but not least, Reyhan for having been an angel all her life.

This thesis, for several reasons, would not be possible without Bilge Say, Haldun Özaktaş and Nihan Özyürek.

I wish to thank Oytun Öztürk, for everything...

Contents

1	Introduction	1
1.1	The Turing Test: A Misfit in Artificial Intelligence	1
1.2	Conversation: A Misfit in Linguistics	2
1.3	Turing Test as Conversation	3
1.4	The Organization of This Thesis	4
2	Turing Test	6
2.1	Introduction	6
2.2	Turing’s ‘Computing Machinery and Intelligence’	8
2.2.1	The Imitation Game	8
2.2.2	Contrary Views and Turing’s Replies	15
2.2.3	Learning Machines	18
2.2.4	Turing’s Predictions	20
2.3	From the Imitation Game to the Turing Test: The 60’s and the 70’s	22
2.3.1	Rocks that Imitate and All-purpose Vacuum Cleaners . .	22

2.3.2	The TT as Science Fiction	24
2.3.3	Anthropomorphism and the TT	26
2.3.4	The TT Interpreted Inductively	26
2.4	In and Out of the Armchair: The 80's and the 90's	29
2.4.1	Behaviorism and Ned Block	30
2.4.2	The Chinese Room	37
2.4.3	Consciousness and the TT	38
2.4.4	Alternative Versions of the TT and Their Repercussions	40
2.4.5	Subcognition and Robert French	48
2.4.6	Getting Real	54
2.5	TT in the Social Sciences	56
2.5.1	Sociological Aspects	56
2.5.2	On Gender	59
2.5.3	Artificial Paranoia	60
2.6	Chatbots	62
2.6.1	The Loebner Contest	62
2.6.2	Tricks of the Trade	65
2.6.3	What Else Should be Done?	72
2.7	Discussions and Conclusion	74
3	A Pragmatic Look At the Turing Test	79
3.1	Pragmatics and Conversation	80

3.1.1	Pragmatics and Why We Care About It	81
3.1.2	The Cooperative Principle and the Conversational Maxims	84
3.1.3	Implicature	86
3.1.4	Some Issues	94
3.2	Empirical Study	95
3.2.1	On Methodology and Choices of Methodology	96
3.2.2	Aims	97
3.2.3	Design	98
3.2.4	The Conversations	102
3.2.5	The Results	109
3.2.6	Discussions	116
3.2.7	On Bias	123
3.3	On Human-Computer Conversation	127
3.3.1	Cooperation as a Special Case of Intentionality	127
3.3.2	Cooperation Revisited: Practical Concerns in General Human-Computer Communication	129
3.3.3	The TT Situation	131
3.3.4	Knowing vs. Not Knowing	133
3.3.5	Implicature vs. Condemnation	134
3.3.6	Cooperation Revisited: The TT Situation	137

4.1 Turing Test: 50 Years Later	140
4.2 Turing Test and Pragmatics	143
4.3 Turing Test and Conversation Planning	146
4.4 A Concluding Remark	148
A List of Conversations	149
B A Sample Open-Ended Survey for Qmax	155
C Qmax	159
D QTT	166
E Tables	175

List of Figures

2.1	The Imitation Game: Stage 1	9
2.2	The Imitation Game: Stage 2, Version 1	10
2.3	The Imitation Game: Stage 2, Version 2	10
2.4	The Imitation Game as is generally interpreted (The Turing Test)	11
3.1	Classification of what is conveyed in conversation	86
3.2	Question format of the Questionnaires	103

List of Tables

3.1	Q _{max} for C3 (Conversation 1)	110
3.2	QTT for C3 (Conversation 1)	110
3.3	Q _{max} for C10 (Conversation 2)	111
3.4	QTT for C10 (Conversation 2)	111
3.5	Q _{max} for C6 (Conversation 3)	112
3.6	QTT for C6 (Conversation 3)	112
3.7	Q _{max} for C4 (Conversation 4)	113
3.8	QTT for C4 (Conversation 4)	113
3.9	Q _{max} for C8 (Conversation 5)	114
3.10	QTT for C8 (Conversation 5)	114
3.11	Q _{max} for C11 (Conversation 6)	115
3.12	QTT for C11 (Conversation 6)	115
3.13	Q _{max} for C13 (Conversation 7)	116
3.14	QTT for C13 (Conversation 7)	116
3.15	Q _{max} for C9 (Conversation 8)	117
3.16	QTT for C9 (Conversation 8)	117

3.17 RL and Not Understanding	118
3.18 MN and RL	118
3.19 Language Use	119
3.20 Emotions	120
3.21 Detection of MN	120
3.22 MN and RL	121
3.23 QN1	122
3.24 Language Use and QN	122
3.25 Language Use	122
3.26 Maxim Violations of the Human in Conversation 1	126
3.27 Maxim Violations of the Human in Conversation 7	126

Chapter 1

Introduction

1.1 The Turing Test: A Misfit in Artificial Intelligence

The idea of "talking computers" was introduced in 1950, before the concept of Artificial Intelligence (AI) even existed [127]. The Imitation Game, better known as the Turing Test (TT), has been proposed by Alan Turing as a means to detect whether a computer possesses intelligence. Although the exact scenario varies, when talking about the TT today what is generally understood is the following: There is a human interrogator who is connected to a computer program via a terminal. His/her task is to find out whether the entity he/she is corresponding with is a machine or a human being. The computer's aim is to "fool" the interrogator. Multiple sessions of this scenario should be carried out and to be granted intelligence, the computer must, on average, manage to convince the interrogators that it is a human being.

Several comments have been made on the TT, many of them discussing its implications on AI. Most of these attack or defend the validity of the test as a means to grant intelligence to machines. There are several computational analyses, an abundance of philosophical comments, and occasional remarks from other disciplines such as psychology and sociology.

Imitation of human linguistic behavior, which is at the very heart of the TT, is a complex issue that refuses to be “solved” by the means and methods of a single discipline. Traditionally, the TT has been considered as a topic that is to be studied within AI. In fact, it is often said that it marks the beginning of AI. Since the TT is about language, it is also related to Natural Language Processing (NLP)¹. On the other hand, a large number of researchers and philosophers prefer to view the TT as a philosophical criterion, not as a serious practical goal. Turing’s original paper is often considered a philosophical piece [127]. This, and the fact that most subsequent comments on the topic have also been of a philosophical nature, have caused the TT to be considered to “belong” to philosophy of artificial intelligence, or more generally, to philosophy of mind. But however one looks at it, the TT is *about* AI.

Although there are concrete computational developments and studies pertaining to the programming of computers that talk to humans, in general, most computer scientists have been rather hostile towards the TT. While it would be rare to find an AI textbook that makes no mention of the TT, most AI researchers seem to not take it as a serious goal. In fact, the reaction of some people has been as harsh as to claim that the TT should be abandoned and buried into history books. However, despite the negative attitude of most researchers, this misfit still remains a largely disputed topic within AI.

1.2 Conversation: A Misfit in Linguistics

In general, linguistics is an “orderly” discipline with elegant formalisms (e.g., grammars). But there are phenomena that cannot be explained by these frameworks which, otherwise, operate in rather smooth and logical ways. Pragmatics is the “wastebasket” in which these offenders are put. In fact, being a misfit in linguistics automatically makes phenomena fall into the domain of pragmatics. People who work on pragmatics try to understand language in relation to its users. In other words, they study language in action.

¹In this thesis I use NLP and Computational Linguistics interchangeably.

Conversation is one of the most interesting phenomena in linguistics. However, it is not easily explained via rules, grammars and similar formalisms. Its study involves a lot of issues outside of linguistics, such as philosophy, sociology, psychology. Conversation is too “disorderly” to be analyzed by syntax and semantics alone and therefore, has been a topic that has received a lot of attention from pragmatics. This is hardly surprising, as conversation is a perfect example of language in action.

1.3 Turing Test as Conversation

Although several comments have been made on the TT, usually it has not been studied as a special kind of conversation. This is rather surprising, because conversation is one of the key issues in the TT. But, for some reason, other aspects of the TT (e.g., imitation, intelligence) have been emphasized, while the fact that the TT is about conversation has not received much attention.

In this thesis, the TT is considered as a special kind of conversation. It is a rather peculiar sort of conversation. For one thing, one participant is a computer. Also, the aims of all participants are clearly defined and the conversation itself is carried out with a specific purpose. In the TT, the computers are expected to display human-like conversational behavior. It is only natural, then, that we should be concerned with what governs human conversation; this is precisely what the computers need to imitate.

As I have depicted above, both the TT and conversation have been misfits of sorts. It is, therefore, expected that when put together, they will be even more difficult to “tame”. Thus, the perspectives and methods in this thesis range from philosophical inquiry to conversational analysis, from practical viewpoints to experimental studies. The current work is, therefore, highly interdisciplinary, borrowing ideas, theories and methodologies from artificial intelligence, linguistics, philosophy, sociology and psychology.

In addition to considering the TT as conversation, this thesis focuses on

one particular aspect of human conversation and attempts to explore it in relation to the TT. This aspect is Grice's *cooperative principle* and *conversational maxims*. Just as Turing's TT is a milestone in AI, Grice's theory is a very well-known and strong part of pragmatics. The powerful juxtaposition of these two concepts is, thus, a significant component of this thesis.

More generally, in this work, I try to show that considering the TT as conversation and analyzing its pragmatic aspects will change the way we look at human-computer communication. Conversely, considering computers as language users will alter the way we look at the whole theory of conversation.

1.4 The Organization of This Thesis

This thesis has two major parts. They differ in approach, style, methodology and focus. But in the end, they are both *about* the TT. Together, they provide not only a deep, but also an original analysis of the TT.

The first part is a review of the TT. This is not simply a larger than average literature survey. During the past 50 years, the TT has been attacked, defended and discussed numerous times, from various angles. A clear, expansive and accessible rendition of all these comments was not available. I have explored some important arguments, summarized the main criticisms of the TT, provided a look at the contributions from other disciplines and at the state of the art in conversational programs at the turn of the century. In addition, some papers that are difficult to locate or understand have been studied in detail and the readers are directed to the list of references for further explication. I believe this broad, interdisciplinary review, in itself, is a contribution and that it will be useful to students and experts alike.

The second part is an analysis of the pragmatics of human-computer conversation, in particular, the TT. This part contains an empirical study that explores the relationship between computers' violations of the conversational maxims and their success in TTs. The results of this study and their discussion is further developed into an analysis of human-computer conversation.

Each of these two main components of the thesis has its own introduction and conclusion. Chapter 2 is the review part. It is here that we study the original game Turing proposed, list and evaluate several comments and criticisms made on the topic, introduce the repercussions of the TT in disciplines other than computer science and philosophy, evaluate the state of the art in natural language conversational system development, and finally, discuss some main issues pertaining to the TT. Chapter 3 begins with an accessible introduction to the field of pragmatics, focuses on Grice's theory of conversation, describes the aims, design, and results of the empirical study, and culminates in a discussion of human-computer conversation. Finally, in Chapter 4 the conclusions of the two parts are brought together and directions for future work are outlined.

Chapter 2

Turing Test

2.1 Introduction

The TT is one of the most disputed topics in Artificial Intelligence, Philosophy of Mind and Cognitive Science. This chapter is a review of the past 50 years of the TT. Philosophical debates, practical developments and repercussions in related disciplines are all covered. I discuss Turing's ideas in detail and present the important comments that have been made on them. Within this context, behaviorism, consciousness, the 'other minds' problem and similar topics in the philosophy of mind are discussed. I also cover the sociological and psychological aspects of the TT. Finally, I take a look at the current situation and analyze the programs that have been developed with the aim of passing the TT. I conclude that the Turing Test has been, and will continue to be, a very influential and controversial topic.

Alan Turing¹, British mathematician, proposed the TT as a replacement for the question "Can machines think?" in his 1950 *Mind* article 'Computing Machinery and Intelligence' [127]. Since then, it has been a widely discussed topic. It has been attacked and defended over and over. At one extreme, Turing's paper has been considered to represent the "beginning" of AI and the TT was considered its ultimate goal. At the other, the TT has been called useless, even

¹For information on Turing refer to the excellent biography by Andrew Hodges [70] or the Alan Turing page at <http://www.turing.org.uk/turing>, also maintained by Hodges.

harmful. In between are arguments on consciousness, behaviorism, the ‘other minds’ problem, operational definitions of intelligence, necessary and sufficient conditions for intelligence-granting, and so on.

It will be the aim of this chapter to present the 50 years of the TT. I have tried to make this review as comprehensive and multi-disciplinary as possible. Important concepts are introduced, and discussed in an easy-to-understand manner. Familiarity with special terms and concepts is not assumed. The reader is directed to further references when they are available. While the review is not strictly chronological, I have tried to present related works in the order they appeared. Interdisciplinary readership is assumed and no particular aspect of the TT (e.g., philosophical or computational) is taken as a focal point.

In my attempt to make this survey complete, I have explored a large number of references. However, this does not mean that I have commented on each paper that mentions the TT. The reader will notice that I have devoted separate sections to certain papers, discussed some others briefly and merely cited the remaining. I made these decisions according to my opinions of what is to be expanded upon in a review of this sort. From this it should not be understood that the papers I spare less space are less important or interesting. In fact, I devoted more space to papers that are not discussed in detail elsewhere². Some papers were explained in detail because they are *representative* of some important ideas.

The rest of the chapter is organized as follows: Section 2.2 introduces the TT and analyzes ‘Computing Machinery and Intelligence’ [127]. In this section, I also attempted to develop new ideas and probe side issues. Section 2.3 describes and explains some of the earlier (those from the 60’s and the 70’s) comments on the TT. In Section 2.4, I analyze the arguments that are more recent. I chose to study the repercussions of the TT in the social sciences separately in Section 2.5. Similarly, in Section 2.6, I give an overview of the concrete, computational studies directed towards passing the TT. Some natural language conversation systems and the annual Loebner Prize contests are overviewed in

²For instance, the discussion of Searle’s Chinese room is kept short (Section 2.4.2), not because it is irrelevant or unimportant, but because there is an abundance of excellent resources on the subject. Conversely, Ned Block’s arguments are described in more detail (Section 2.4.1) because not many in-depth analyses of them were found in the literature.

this section. Finally, Section 2.7 concludes my survey.

2.2 Turing’s ‘Computing Machinery and Intelligence’

It makes sense to look at Turing’s landmark paper ‘Computing Machinery and Intelligence’ [127] before we begin to consider certain arguments defending, attacking or discussing the TT. [127] is a very well-known work and has been cited and quoted copiously. Although what follows will provide an introduction to the TT, it is a good idea to read Turing’s original rendering of the issues at hand. In analyzing the 50 years of the TT, it is important to distinguish what has been originally proposed by Turing himself and what has been added on afterwards. I am not saying, by any means, that the TT is what (or should remain as) Turing proposed in ‘Computing Machinery and Intelligence’. As any other concept, it has changed throughout the 50 years it has been around. In fact, one of the purposes of this chapter is to trace the steps in this evolution. Thus, it is only natural that we are interested in the original version.

In Section 2.2.1, I analyze Turing’s original proposal. I summarize Turing’s replies to certain objections to his ideas in Section 2.2.2. Turing’s opinions on learning machines are briefly discussed in Section 2.2.3. Finally, I list some predictions of Turing in Section 2.2.4.

2.2.1 The Imitation Game

Turing’s aim is to provide a method to assess whether a machine can think or not. He states at the beginning of his paper that the question “Can machines think?” is a highly ambiguous one. He attempts to transform this into a more concrete form by proposing what is called the Imitation Game (IG). The game is played with a man (A), a woman (B) and an interrogator (C) whose gender is unimportant. The interrogator stays in a room apart from A and B. The objective of the interrogator is to determine which of the other two is the

woman while the objective of both the man and the woman is to convince the interrogator that he/she is the woman and the other is not. This situation is depicted in Figure 2.1.

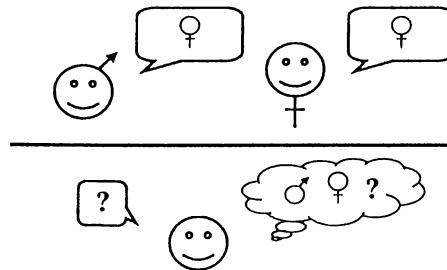


Figure 2.1: The Imitation Game: Stage 1

The means through which the decision, the conviction, and the deception is to take place is a teletype connection. Thus, the interrogator will ask questions in written natural language and will receive the answers in written natural language. Questions can be on any subject imaginable, from mathematics to poetry, from the weather to chess.

According to Turing, the new agenda to be discussed, instead of the equivocal “Can machines think?”, can be ‘What will happen when a machine takes the part of A in this game? Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?’ [127, p. 434]. Figure 2.2 depicts the new situation.

At one point in the paper Turing replaces the question “Can machines think?” by the following:

‘Let us fix our attention to one particular digital computer *C*. Is it true that by modifying this computer to have an adequate storage, suitably increasing its speed of action and providing it with an appropriate programme, *C* can be made to play satisfactorily the part of A in the imitation game, *the part of B being taken by a man?*’ [127, p. 442, emphasis added].

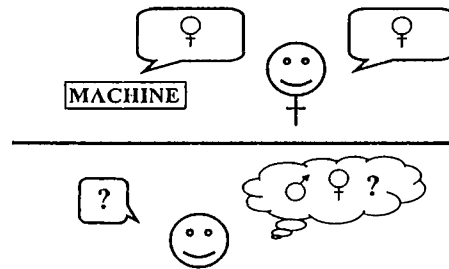


Figure 2.2: The Imitation Game: Stage 2, Version 1

Notice that the woman has disappeared altogether. But the objectives of A , B and the interrogator remain unaltered; at least Turing does not explicitly state any change. Figure 2.3 shows this situation.

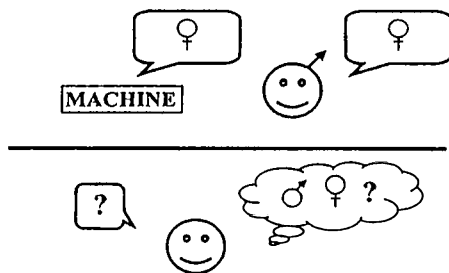


Figure 2.3: The Imitation Game: Stage 2, Version 2

There seems to be an ambiguity in the paper; it is unclear which of the scenarios depicted in Figure 2.2 and Figure 2.3 is to be used. In any case, as it is now generally understood, what the TT really tries to assess is the machine's ability to imitate a human being, rather than its ability to simulate a woman. Most subsequent remarks on the TT ignore the gender issue and assume that the game is played between a machine (A), a human (B) and an interrogator (C). In this version, C 's aim is to determine which one of the two entities he/she is conversing with is the human (Figure 2.4).

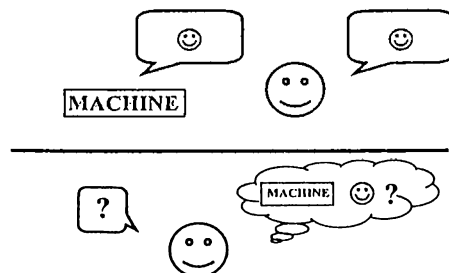


Figure 2.4: The Imitation Game as is generally interpreted (The Turing Test)

One may ask why Turing designed the IG in such a peculiar manner. Why the fuss about the woman, the man and the replacement? This does not make the paper easier to understand. He could have introduced the IG exactly as he did with the woman-man issue replaced by the human-machine issue and it obviously would not be any more confusing. One argument that can be made against this is that machines (at least those that could be built or imagined in 1950) playing against men in such a game would sound ridiculous at first. A man imitating a woman, on the other hand, has higher prospects of success in the eyes of the average person. In other words, it can be said that the gender-based imitation game sets the mood for what's coming. This, I believe, is not a very convincing argument. The main reason that the decision concerning machine thought is to be based on imitating a woman in the game is probably not that Turing believed the ultimate intellectual challenge to be the capacity to act like a woman (although it may be comforting to entertain the thought). Conversely, it may be concluded that Turing believes that women can be imitated by machines while men cannot. The fact that Turing stipulated the man to be replaced by the machine (when he might just as easily have required the woman to be replaced by the machine or added a remark that the choice was insubstantial) raises such questions, but let us not digress.

Here is my explanation of Turing's design: The crucial point seems to me that the notion of *imitation* figures more prominently in Turing's paper than is commonly acknowledged. For one thing, the game inherently possesses deception. The man is allowed to say anything at all in order to cause the interrogator

to make the wrong identification, while the woman is actually required to aid the interrogator³. In the machine vs. woman version, the situation will remain the same. The machine will try to convince the interrogator that it is the woman. What is really judging the machine's competence is not the woman it is playing against. Turing's seemingly frivolous requirements may actually have very sound premises. Neither the man in the gender-based IG nor any kind of machine is a woman. On close examination, it can be seen that what Turing proposes is to compare the machine's success against that of the man, *not* to look at whether it 'beats' the woman in the IG⁴. The man and the machine are measured in terms of their respective performances against real women. In Figure 2.3, we see that the woman has disappeared from the game, but the objective for both the machine and the man is still imitating a woman. Again, their performance is comparable because they are both simulating something which they are not.

The quirks of the IG may well be concealing a methodological fairness beyond that explicitly stated by Turing. I hold that the IG, even though it is regarded as being obscure by many, is a carefully planned proposal. It provides a fair basis for comparison; the woman (either as a participant in the game or as a concept) acts as a neutral point so that the two imposters can be assessed in how well they "fake".

Turing could have defined the game to be played with two people, too; one being interrogator, as in the original, and the other being either a man or a woman. The interrogator would then have to decide whether the subject is a man or a woman. Alternatively, the TT for machine intelligence can be re-interpreted as a test to assess a machine's ability to pass for a human being. This issue may seem immaterial at first. However, the interrogator's decision is surely to be affected by the availability (or lack) of comparison. Whether the machine's task will be easier or more difficult in this latter case is another question. We think that Turing intended to imply that some comparison should be available, for otherwise, he could have opted for the two-people version of the game. This implies that the game can be played with the result 'A is the

³Turing suggests that the best strategy for her would most probably be giving truthful answers to the questions.

⁴This inadvertently figures in the final result, but indirectly.

woman' actually meaning 'A seems more woman-like than B'. In turn, a more varied set of questions can be used when the interrogator is trying to judge the gender of the subjects. Once again, I believe that the most sensible reason behind the three-person game is to have a neutral party so as to allow the assessment of the impersonating parties with respect to each other.

In any case, as was mentioned before, the TT concept has evolved through time. Turing's original IG and its conditions do not put serious constraints on current discussions about the test. It is generally agreed that the gender issue and the number of participants are not to be followed strictly in attempts to pass, criticise or defend the TT. Even Turing himself, in the subsequent sections of 'Computing Machinery and Intelligence', sometimes ignores these issues and focuses on the question: "Can machines communicate in natural language in a manner indistinguishable from that of a human being?". This is manifested in the example conversation he gives in [127, p. 434], which contains questions about poetry, mathematics and chess—topics that one would not typically ask about in order to find out the gender of someone. This may be a hint that the gender issue in the IG is indeed for purposes of fair comparison.

After defining the IG, Turing defends the choice of replacing the "Can machines think?" question with "Can machines play the imitation game?". The new problem focuses on intellectual capacities and does not let physical aspects interfere with granting intelligence to an entity. Nor does it limit thinking to specific tasks like playing chess or solving puzzles, since the question-and-answer method is suitable to introduce any topic imaginable.

An issue that is open to discussion is what Turing implies about *how* machines should be built or programmed to play the IG successfully. He seems to believe that if a machine can be constructed to play the game successfully, it does not really matter whether what it does to that end is similar to what a man does or not. Here it can be seen that Turing almost encourages prospective attempts to pass the test to utilize any kind of strategy whatsoever. He even considers the possibility that a machine which successfully plays the IG cannot be explained by its creators because it had been built by experimental methods. However, he explicitly states that 'it will be assumed that the best strategy is to try to provide answers that would naturally be given by a

man' [127, p. 435]. It may be concluded that Turing does not put any limitations on how to model human cognitive processes, but seems to discourage any approach that deviates too much from the "human ways", possibly because he feels it is unlikely that satisfactory solutions can be obtained in this manner. On the other hand, by not committing himself to any extreme viewpoint on the issue, he accepts the possibility that machines not mimicking human cognitive processes at all can also pass the test.

The IG, as was mentioned, has deception at its very heart. It is therefore apparent that "cheating" will be an integral part of the TT. Moreover, by not stipulating certain techniques or strategies to be used, Turing explicitly allows this. As Turing described it, in the game there are no rules constraining the design of the machines.

Turing promotes cheating implicitly, too. At various places in the paper, he describes how machines could be "rigged" to overcome certain obstacles proposed by opponents of the idea that machines can think. A very obvious example is about machines making mistakes. When the machine is faced with an arithmetical operation, in order not to give away its identity by being fast and accurate, it can pause for about 30 seconds before responding and occasionally give a wrong answer. Being able to carry out arithmetical calculations fast and accurately is generally considered intelligent behaviour⁵. However, Turing wishes to sacrifice this at the expense of human-ness. But this is cheating, is it not? Maybe, but the arithmetics domain is a highly specific one. Cheating in this manner cannot hurt; if a machine can pass the test, it can then be re-programmed not to cheat at arithmetics. If it does not cheat, the interrogator can ask a difficult arithmetical problem as his/her first question and decide that he/she is dealing with a machine right then and there.

It can be seen that Turing does not seem to be skeptical towards the idea that a sufficiently human-like machine (i.e., a machine that is sufficiently good at playing the IG) is bound to make such mistakes as we attribute to humans, without any explicit cheating done by its constructors. This idea may seem

⁵Although even simple devices like calculators are better at this than average human beings, it is rare that a mathematical whiz who can multiply 8-digit numbers in seconds is regarded as being of ordinary intellect.

extravagant, but considering the high level of sophistication required from a machine for passing the TT, it should not be dismissed as being impossible. As Turing also mentions, nothing can really stop the machine from drawing incorrect conclusions without being specifically programmed to do so. A striking example can be given from the inductive learning domain: No learning algorithm guarantees correct results on unseen data. Moreover, in some cases a computer errs in ways that cannot be foreseen, or even understood by its programmer. This can be distressing for machine learning researchers who are after a minimal number of mistakes, but proves the subtle point that machines can make mistakes without explicitly shown *how to*⁶. Since the human mind occasionally draws incorrect conclusions inductively, the fact that machines can act similarly should contribute to the arguments that refute the notion that machines cannot make mistakes.

Turing's approach towards cheating seems similar to that of Adam Smith's "invisible hand" from economics. Maybe Turing's conformity to cheating has its roots in his belief that one cannot go too far by such attempts: He may regard cheating as a last retouch, something to smooth out the machine-ness of the resulting machines that otherwise handle the more important aspects of human cognition. If a program that has its very bases in what we now call cheating can pass the TT, maybe we would have to revise some notions about the human intellect. It is not possible to say what Turing was thinking and claim to be absolutely correct. It seems as if he would be content with a machine that plays the IG successfully no matter what the inner mechanisms are.

2.2.2 Contrary Views and Turing's Replies

Turing was aware that some of his ideas would be opposed at the time he wrote 'Computing Machinery and Intelligence' [127] and he responded to some objections that he believed his work would be confronted with. In fact, he

⁶Readers are referred to Section 2.2.3 of this thesis, [127, pp. 454-460], and [128, pp. 14-23] for very entertaining and insightful comments on machine learning by Turing.

discusses some of these even before he formally proposes the IG, in [128]⁷. I direct the reader to [127] for the answers to the *theological objection*, and the *argument from extra-sensory perception* for these are rather irrelevant to the current work. However, the remaining objections are worth commenting on.

The *'heads in the sand' objection*, although mostly in disguised forms, is manifested in some subsequent comments on the TT. This is, in its basic form, an aversion of the issue of thinking machines because the consequences of this would be dreadful [127, p. 444]. Most people like to believe that humans are “special” and thinking is considered one of the most important traits that make us so. To some, the idea of sharing such a “human” ability with machines is not a pleasant thought. This outlook was probably more widespread in Turing’s time than it is now. Turing believes that this argument is not even worth refutation, and with a little sarcasm, states that consolation (perhaps in the transmigration of souls) is more appropriate [127, p. 444].

There are some theorems showing the powers of discrete-state machines are limited. The most famous of these is probably Gödel’s theorem which shows that in consistent logical systems of sufficient power, we can formulate statements that cannot be proved or disproved within the system. An application of this result to the IG is outlined in [127, p. 445] and the reader can refer to [87, 88] for more on the implications of Gödel’s theorem for machine thought.

Turing studies such results under the title the *mathematical objection*. He states that ‘although it is established that there are limitations to the powers of any particular machine, it has only been stated, without any sort of proof, that no such limitations apply to the human intellect’ [127, p. 445]. Elsewhere, he notes that those arguments that rest on Gödel’s and similar theorems are taking it for granted that the machine must not make mistakes, but that this is not a requirement for intelligence [128].

Perhaps the most important objection is the *argument from consciousness*. Some people believe that machines should be conscious (e.g., aware of their

⁷Although the reference cited is published in 1969, Turing originally wrote the paper in 1948.

accomplishments, feel pleasure at success, get upset at failure, etc.) in order to have minds. At the extreme of this view, we find *solipsism*. The only way to *really* know whether a machine is thinking or not is to *be* that machine. However, according to this view, the only way to know another human being is thinking (or is conscious, happy, etc.) is to be that human being. This is usually called the *other minds problem* and will show up several times in the discussions of the TT. ‘Instead of arguing continually over this point it is usual to have the polite convention that everyone thinks’ [127, p. 446]. Turing’s response to the argument from consciousness is simple, but powerful: The alternative to the IG (or similar behavioral assessments) would be solipsism and we do not practice this against other humans. It’s only fair that in dealing with machine thought, we abandon the consciousness argument rather than concede to solipsism.

Turing believes that the IG setting can be used to determine whether ‘someone really understands something or has learnt it parrot fashion’ as is manifested in the sample conversation he gives in [127, p. 446]. It should also be noted that Turing states he does not assume consciousness to be a trivial or irrelevant issue; he merely believes that we do not necessarily need to solve its mysteries before we can answer questions about thinking, and in particular, machine thought [127, p. 447].

The *arguments from various disabilities* are of the sort “machines can never do X ”, where X can be any human trait such as having a sense of humor, being creative, falling in love or enjoying strawberries. As Turing also notes [127, p. 449], such criticisms are sometimes disguised forms of the argument from consciousness. Turing argues against some of these X ’s such as the ability to make mistakes, enjoy strawberries and cream, be the subject of its own thought, etc. in [127, pp. 448-450].

Lady Lovelace’s objection is similar; it states that machines cannot originate anything, can never do anything new, can never surprise us. Turing replies by confessing that machines do take him by surprise quite often. Proponents of Lady Lovelace’s objection can say that ‘such surprises are due to some creative mental act on [Turing’s] part, and reflect no credit on the machine’ [127, p. 451]. Turing’s answer to this is similar to the one he gives to the argument from

consciousness: ‘The appreciation of something as surprising requires as much of a ‘creative mental act’ whether the surprising event originates from a man, a book, a machine or anything else.’ [127, p. 451].

Turing also considers the *argument from continuity in the nervous system*. As the name suggests, this objection states that it is impossible to model the behavior of the nervous system on a discrete-state machine because the former is continuous. However, Turing believes that the activity of a continuous machine can be “discretized” in a manner that the interrogator cannot notice during the IG.

Finally, there is the *argument from informality of behavior*. Intuitively, it seems it is not possible to come up with a set of rules that describe what a person would do under every situation imaginable. In very simple terms, some people believe the following: ‘If each man had a definite set of rules of conduct by which he regulated his life, he would be no better than a machine. But there are no such rules, so men cannot be machines.’ [127, p. 452]. First, Turing notes that there might be a confusion between ‘rules of conduct’ and ‘laws of behavior’. By the former he means actions that one can perform and be aware of (like, ‘If you see a red light, stop’) and by the latter he means laws of nature that apply to a man’s body (such as ‘If you throw a dart at him, he will duck’). Now it is not evident that a complete set of laws of behavior do not exist. We can find some of these by scientific observation but there will not come a time when we can be confident we have searched enough and there are no such rules. Another point Turing makes is that it may not always be possible to predict the future behavior of a discrete-state machine by observing its actions. In fact, he is so confident about a certain program that he set up on the Manchester computer that he ‘def[ies] anyone to learn from [its] replies sufficient about the programme to be able to predict any replies to untried values’ [127, p. 453].

2.2.3 Learning Machines

Turing devotes some space to the idea of *education of machinery* in ‘Computing Machinery and Intelligence’ [127]. He also discusses the issue in his earlier work

‘Intelligent Machinery’ [128].

According to Turing, in trying to imitate an adult human mind, we should consider three issues: the initial state of the mind, the education it has been subject to, and other experience it has been subject to (that cannot be described as education). Then we might try to model a child’s mind and “educate” it to obtain the model of the adult brain. Since ‘presumably the child-brain is something like a note-book as one buys it from the stationers; rather little mechanism and lots of blank sheets’ [127, p. 456], developing a program that simulates it is bound to be easier⁸. Of course, the education is another issue. Turing proposes some methods of education for the child-machines (such as a reward/punishment based approach) in [127, pp. 456-460] and [128, pp. 17-23].

Turing’s opinions on learning machines are rather interesting, especially considering he wrote these more than 50 years ago. I will not digress into discussions of his ideas on the specifics or the realizability of his proposals. I would like to note one thing though: In most places when he discusses education of machines, there is a noticeable change in Turing’s style. He seems to believe that the way to success in developing a program that plays the IG well is probably following the human model as closely as possible. As was mentioned in Section 2.2.1, he does not put any constraints on how to design the IG-playing machine, but the fact that he describes learning machines in substantial detail seems to suggest that *he* would prefer such an approach.

In any case, Turing believes ‘*if we are trying to produce an intelligent machine, and are following the human model as closely as we can*’ [128, p. 14, emphasis added] a good (and fair) approach would be to allow the machine to learn just like humans.

⁸Turing seems to believe that brains of newborn babies are *tabula rasa*. However, he also considers the opposite and states that we might encode the information at various kinds of status levels (e.g., established facts, conjectures, statements given by an authority) and thereby implies that we may model any ‘innateness’ there may be [127, pp. 457-458].

2.2.4 Turing's Predictions

Turing's paper [127] contains some very bold comments on the prospects of machine intelligence. Most of these probably seemed like science fiction at the time. Even now, some of us would consider these far-fetched. This section aims to provide a sample of Turing's predictions that I found interesting.

It is a well-known fact that Turing believes computers to be capable of performing many "intelligent" tasks. He also thinks that they will be able to do so in a "human" way.

The reader must accept it as a fact that digital computers can be constructed, and indeed have been constructed, according to the principles we have described, and that they can in fact mimic the actions of a human computer very closely [127, p. 438].

As can be seen from the following quote, Turing believes that the difficulties in designing thinking machines are not unsurmountable.

As I have explained, the problem is mainly one of programming. Advances in engineering will have to be made too, but it seems unlikely that these will not be adequate for the requirements [127, p. 455].

While trying to convince the reader that the ideas he proposes are of the sort that can be realized in the foreseeable future, Turing mentions some concrete achievements he expects from computers. Those that are related to machine learning were outlined in Section 2.2.3. Here is another example, this time pertaining to automated software engineering:

[The machine] may be used to help in making up its own programmes, or to predict the effect of alterations in its own structure.

...

These are possibilities of the near future, rather than Utopian dreams [127, p. 449].

The game of chess has been at the center of some of the most well-known achievements in AI. Today, computer programs play against world champions and sometimes even beat them. Spectacular advances have more recently been made in computer understanding and generation of speech. Although to what extent currently available speech processing systems are intelligent is a debatable issue, they (like chess-playing programs) have become part of modern life:

We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult question. Many people think that a very abstract activity, like the playing of chess, would be best. It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English.

...

Again, I do not know what the right answer is, but I think both approaches should be tried [127, p. 460].

Take a look at computer technology at the turn of the century: What was unimaginable in 1950, in terms of memory and speed, is now reality. What Turing predicted about the IG, however, is still a challenge.

I believe that in about fifty years' time, it will be possible to programme computers with a storage capacity of about 10^9 , to make them play the imitation game so well that an average interrogator will not have more than 70 percent chance of making the right identification after five minutes of questioning [127, p. 442].

2.3 From the Imitation Game to the Turing Test: The 60's and the 70's

Earlier remarks on the TT, with the exception of [18, 19, 131] have mostly been of the philosophical sort. This is hardly surprising because ‘Computing Machinery and Intelligence’ was itself published in a philosophy journal, *Mind*⁹. Many discussions on the IG were published in the 60's and the 70's, most of the important contributions once again accommodated by *Mind*. In this section we take a look at these philosophical papers, leaving the more practical work described in [18, 19, 131] to other, more appropriate sections. Readers interested in earlier comments on the TT and machine intelligence that are not discussed in this section can consult [92, 110].

Keith Gunderson's comments on the IG are summarized in Section 2.3.1. Section 2.3.2 presents an approach stating that developing a TT-passing program is not going to be possible in the foreseeable future. The anthropomorphism in the TT is briefly discussed in Section 2.3.3, to be taken up later on. An inductive interpretation of the TT is described in Section 2.3.4.

2.3.1 Rocks that Imitate and All-purpose Vacuum Cleaners

One of the earlier comments on Turing's IG came from Keith Gunderson in his 1964 *Mind* article [53]. In this paper, appropriately titled ‘The Imitation Game’, Gunderson points out some important issues pertaining to Turing's replacement for the question “Can machines think?”.

Gunderson develops certain objections to Turing's ‘Computing Machinery and Intelligence’ [127] by focusing on the IG. In a nutshell, he emphasizes two points: First, he believes that playing the IG successfully is an *end* that can be achieved through different means, in particular, without possessing intelligence. Secondly, he holds that thinking is a general concept and playing the IG is but

⁹Although the cover of the 1950 issue reads “A Quarterly Review of Philosophy and Psychology”, I find it not too inappropriate to call *Mind* a philosophy journal.

one example of the things that intelligent entities do. Evidently, both claims are critical of the validity of the IG as a measure of intelligence.

Gunderson makes his point by an entertaining analogy. He asks the question “Can rocks imitate?” and continues to describe the “toe-stepping game” [53, p. 236] in a way that is identical to the way Turing described his IG in [127]. Once again, the game is played between a man (A), a woman (B) and an interrogator (C). The interrogator’s aim is to distinguish between the man and the woman by the way his/her toe is stepped on. C stays in a room apart from the other two and cannot see or hear the toe-stepping counterparts. There is a small opening in the wall through which C can place his/her foot. The interrogator has to determine which one of the other two is the woman by the way in which his/her toe is stepped on. Analogously, the new form of the question “Can rocks imitate?” becomes the following: ‘What will happen when a rock box is constructed with an electric eye which operates across the opening in the wall so that it releases a rock which descends upon C’s toe whenever C puts his foot through A’s side of the opening, and thus comes to take the part of A in this game? ... Will the interrogator decide wrongly as often as when the game is played between a man and a woman?’ [53, pp. 236-237].

Gunderson believes that even if rock boxes play the toe-stepping game successfully, there would still be no reason to accept that they are imitating. The only conclusion that we can make from this would be that a rock box can be rigged in such a way that it can replace a human being in the toe-stepping game. According to Gunderson, this is because ‘part of what things do is how they do it’ [53, p. 238]. As I will expand upon in Section 2.4.1, this is similar to Ned Block’s argument for *psychologism* against behaviorism [8].

Gunderson states that thinking is not something that can be decided upon by just one example. He demonstrates his belief that a computer’s success in the IG is not sufficient reason to call it a thinking machine by another analogy: Imagine a vacuum cleaner salesman trying to sell a product. First, he advertizes the vacuum cleaner *Swish 600* as being “all-purpose”. Then, he demonstrates how it can suck up bits of dust. The customer asks what else the machine can do. Astonished, the salesman says that vacuum cleaners are for sucking up dust and *Swish 600* does precisely that. The customer answers,

“I thought it was all-purpose. Doesn’t it suck up bits of paper or straw or mud? I thought sucking up bits of dust was an *example* of what it does.”. The salesman says “It is an example of what it does. What it does is suck up pieces of dust.” [53, p. 241].

The salesman is having trouble making his sale by calling Swish 600 all-purpose and being unable to show more than one example of what it does. According to Gunderson, Turing is also having the same problem because the term “thinking” is used to refer to more than one capability; just as the term “all-purpose” implies that the vacuum cleaner has functions other than just sucking up bits of dust. He concludes:

In the end the steam drill outlasted John Henry as a digger of railway tunnels, but that didn’t prove the machine had muscles; it proved that muscles were not needed for digging railway tunnels [53, p. 254].

John G. Stevenson, in his 1976 paper ‘On the Imitation Game’ [126] raises some arguments against Gunderson. One of these is the objection that Gunderson was expecting, namely the claim that being able to play the IG is not just *one* example; a machine that is good at the IG is capable of various things. Gunderson does not give a direct response to such objections. He mentions a reply can be formulated along the lines of showing that even combining all those things such a machine can do gives us a narrow range of abilities [53, p. 243]. Stevenson doubts whether such a reply would be adequate [126, p. 132]. Even if it does not exhaust everything that is related to human thinking, he believes the list of things that a computer that plays the IG can do would be quite impressive. Stevenson states that Gunderson is ignoring the specific character of the IG and that he proposes defective arguments.

2.3.2 The TT as Science Fiction

Richard Purtill, in his 1971 *Mind* paper also discusses some issues concerning the IG. Purtill criticizes some ideas in Turing’s paper ‘mainly as a philosopher,

but also as a person who has done a certain amount of computer programming' [108, p. 290]. He believes that the game is interesting, but as a piece of science fiction. He finds it unimaginable that a computer playing the IG will be built in the foreseeable future.

Overall, Purtill believes the IG to be a computer man's dream. He even promises to 'eat his computer library' if anyone has a notion on the principles on which a machine that can play the game is to be built [108, p. 293]¹⁰. He states that if computers, some day, behave like the computers in works of science fiction, he would grant them thought. But since all computer outputs can be explained as a result of a program written by humans¹¹, computers are not likely to play the IG successfully with the currently imaginable programming techniques. This, he believes, is because the behavior of thinking beings is not deterministic and cannot be explained in purely mechanistic terms.

Purtill believes that the game is 'just a battle of wits between the questioner and the programmer: the computer is non-essential' [108, p. 291]. Although the former part of the claim may be reasonable to an extent, his latter argument about the computer being non-essential is not very sound. To eliminate the computer from the picture, Purtill proposes "purely mechanical" alternatives: machines made of levers and wheels that can do the same task. I think it is unclear why this should count as an argument against the IG because evidently, the material or structure on which the IG-playing "program" works is irrelevant. Purtill also states, anticipating the objection that the human mind might also be a highly complex collection of such mechanical processes, that if this were the case, it would mean 'human beings do not in fact think rather than that computers do think' [108, p. 292], but does not attempt to justify this bold claim.

In his short paper 'In Defence of Turing' [116], Geoffrey Sampson attacks Purtill's arguments briefly. First of all, he believes most of the limitations pertaining to the realization of IG-playing computers Purtill lists are practical difficulties that may be overcome in the (presumably not so distant) future.

¹⁰Recall that the paper is written in 1971.

¹¹Even if the program's outputs are guided by certain random elements.

Secondly, he states that it is only natural that computer behavior is deterministic and that human behavior is not so easy to explain. The reasons for this are simple: computers are designed by humans; they have mechanisms that explicitly allow us to study their behavior; humans are much more complex in terms of both internal states and possible inputs than any contemporary computer [116, p. 593]. Sampson also rejects Purtil's opinion that the consequence of the claim that human thinking is an extremely complex, yet computer-like, mechanical process is that men do not think. He holds that thinking, by definition, is something human beings do.

2.3.3 Anthropomorphism and the TT

In a short paper that appeared in *Mind* in 1973 [99], P. H. Millar raises some important issues which will show up in later works. He first discusses some vices and virtues of the IG and states that it is irrelevant whether or how the computers or the human beings involved in the game are "programmed". Then, he introduces the question of whether the IG is a right setting to measure the intelligence of machines. Millar notes that the game forces us to "anthropomorphize" machines by ascribing them human aims and cultural backgrounds. Millar asserts that the IG measures not whether machines have intelligence, but whether they have *human* intelligence. He believes that we should be open-minded enough to allow each being, be it a Martian or a machine, to exhibit intelligence 'by means of behavior which is well-adapted for achieving its own specific aims' [99, p. 597]. I return to this issue later on, especially in Section 2.4.5 and Chapter 3.

2.3.4 The TT Interpreted Inductively

In his important paper 'An Analysis of the Turing Test' [102], James Moor attempts to emphasize the significance of the imitation game. As can be seen from the title, the term "Turing Test" was already being used to refer to the IG by 1976. Moor's main assertion is that 'the Turing Test is a significant test for computer thought if it is interpreted inductively.' [102, p. 256].

Moor disagrees with the idea that the TT is an operational definition of intelligence¹². Rather, he proposes, it should be regarded as a source of inductive evidence for the hypothesis that machines can think [102]. Moreover, Moor does not agree with the claim that even if the TT is not an operational definition, it should at least be a necessary condition for granting computers intelligence. According to him, there could be other evidence based on the computer's behavior that leads to inferences about the computer's thinking abilities. However, he believes that the test provides a sufficient condition for intelligence-granting to computers. But his view is not "absolute"; he accepts that it is possible to revise a positive inference about a computer's possession of thought based on a TT, if other evidence is acquired afterwards.

Moor lists two arguments that support the TT as a good format for collecting inductive evidence. 'First, the Turing Test permits direct or indirect testing of virtually all of the activities one would count as evidence for thinking . . . Secondly, the Turing Test encourages severe testing.' [102, pp.251-252]. By the latter, Moor means the test's requirements are not too easy to meet. For instance competence in a single cognitive activity, no matter how complex, would not suffice.

Moor proceeds by considering some of the objections to the TT. He gives replies to these objections and shows that they are either irrelevant or can be refuted when the TT is considered to be a way of gathering data based on which we may inductively infer conclusions about machine thought. One objection that Moor, in my opinion successfully, replies is the objection concerning internal operation. The view that information about the internal information processing of a system is important in granting it intelligence is not uncommon [8, 53, 117]. Moor warns against the possible confusion between the two variants of this conception. There is an important difference between the claim that evidence about the internal operation of a computer *might alter* a justified inductive inference that the computer can think, and the claim that such evidence *is necessary to make* such an inference. Moor believes the former and notes that this is certainly not a criticism that can be addressed to the TT. If certain kinds of information about the internal operation of a machine

¹²As opposed to Millar, who believes this to be true, and also that this is a virtue of the imitation game [99].

that was believed to possess intelligence after being Turing Tested is acquired afterwards, then we might just revise our decision. If the latter alternative were true, then the objection could be used against the test. But, according to Moor, critics fail to show that this is true and they are not likely to ever succeed.

As was discussed above within the context of Gunderson's paper [53], the TT may be considered inadequate because it is only *one* evaluation of behavior. Moor answers this kind of objection also in a liberal light, in a manner similar to his discussion outlined above. Once again he makes a distinction between two claims: one positing that behavioral evidence which cannot be directly obtained in the TT *might alter* a justified inductive inference that a computer can think, and the other stating that such evidence *is necessary to make* this decision. Moor believes that the former is true. Further testing, he says, would be valuable and could even make us change our inference. The important point is, this does not incapacitate the TT in any way. The test could be attacked on these premises only if the latter claim were true. Moor believes the critics have not, and are not going to be able to prove this. This is because he believes that the format provided by the test enables examining a very large set of activities that would count as evidence of thinking. Thereby, he refutes the objections about the scope of the test.

Moor concludes by stating that although the TT has certain shortcomings (e.g., it being of little value in guiding research), it is an important measure for computer thought when it is inductively interpreted. Moreover, the standard criticisms of the TT fail to show that it is deficient if such an interpretation is made.

A reply to Moor comes from Douglas F. Stalker [125]. He prefers to call Moor's interpretation an explanatory one rather than an inductive one. Stalker notes that Moor's beliefs about the mentality of other humans, as well as computers, are part of an explanatory theory. He emphasizes that Moor does not justify that his theory of explaining a computer's success at the TT by using the concept of thinking is the *best* theory that can be constructed about the same phenomenon.

As an alternative explanation for the computer's behavior, Stalker proposes a purely mechanistic theory that does not appeal to any mental concepts. His theory takes into consideration such factors as the computer's physical structure, its program and its physical environment. Moreover, he believes this theory to be preferable compared to Moor's. Stalker believes explanatory theories that involve concepts of thinking can apply to people, but because of some fundamental differences between computers and humans, they may not be the best theories for explaining computer behavior.

In his answer to Stalker [103], Moor basically says that the existence of alternative explanations does not mean that they would necessarily be competitors. It is true that an explanation for a computer's activities can be given at different levels: physics, electronic circuitry, programs, abstract automata, etc. Moor notes that these explanations would be different, but not necessarily rivals. In the case of a computer displaying intelligent behavior by being successful in the IG, an explanatory theory involving thinking could even be preferred because it is simpler and easier to understand. Moor's conclusion is:

It seems natural and probably most understandable to couch the explanation in terms of a theory of mind. If one has the patience, the explanation could also be given at lower levels of description, e.g., involving perhaps thousands of computer instructions or millions of changes in circuit states [103, p. 327].

2.4 In and Out of the Armchair: The 80's and the 90's

While thought experiments are still around, it can be seen that the late 80's and the 90's also feature a tendency to leave the comfortable armchair of philosophy. In this section I cover only some of the works that have addressed the TT. This is mainly because of the sheer abundance of comments on the issue. The subset of the work done during the 80's and the 90's that I chose to present in this section provide a general overview of the main arguments and

the reader is directed to references for further explication. A must-read is Douglas Hofstadter's 'Turing Test: A Coffeehouse Conversation' [71] which is full of valuable and entertaining insights. Ajit Narayanan studies the intentional stance and the IG [105]. For a discussion of the frame problem in relation to the TT, the reader is referred to [25]. Other references that can be explored are [55, 109, 39, 51, 6, 3, 26, 106, 16, 122, 72, 90, 24, 33, 66]. A number of articles on the TT have appeared in popular science magazines, too. Some of these are [52, 28, 107, 36, 129].

The TT scene began heating up at the beginning of the 80's. Although the "consciousness argument" and the "anti-behaviorist argument" were voiced before, they had not been really unsettling. Then, in the early 80's, two strong counter-arguments against the TT were formulated by John Searle and Ned Block, respectively. The debate on Searle's "Chinese Room" is in itself expansive enough to be the subject of a whole chapter of at least this size. I consider it briefly in Section 2.4.2 and the interested reader should have no difficulty finding more information about the issue. Block's anti-behaviorist attack of the TT, on the other hand, was not disputed much and it is the aim of Section 2.4.1 to elaborate on his ideas.

Section 2.4.3 is a brief look at consciousness and the TT. Various attempts have been made to modify the TT to get better "tests" for machine thought, and these are discussed in Section 2.4.4. Robert French's 'Subcognition and the Limits of the Turing Test' [40] is analyzed in Section 2.4.5. Finally, the "less philosophical" stance towards the TT is discussed in Section 2.4.6.

2.4.1 Behaviorism and Ned Block

In 'Psychologism and Behaviorism' [8], Ned Block attacks the TT as a behaviorist approach to intelligence. Although this paper was written in 1981, Block still seems to hold the same opinions, see [9].

Block believes that the judges in the TT can be fooled by *mindless* machines that rely on some simple tricks to operate. He proposes a hypothetical machine that will pass the TT, but has a very simple information processing component.

Block's machine has all possible conversations of some given length recorded in its memory. Of course, we want these conversations to be such that at least one party is 'making sense' [9]. The set of strings constituting such conversations that can be carried out in a fixed amount of time are finite and thus can be enumerated and stored in our hypothetical computer. The judge types in a string, say A . The machine finds a conversation beginning with A and types out the second sentence of this string, say B . If, next, the judge types in C , the process is repeated with A replaced by ABC . All the machine does is simple "lookup and writeout", certainly nothing that anyone would call sophisticated information processing.

Since this machine has the intelligence of a jukebox [9] or a toaster [8], and since it will pass the TT, the test must be an inadequate measure of intelligence. Block ties this conclusion to the more general one that this is because of the behaviorist approach taken in the design of the TT.

Ned Block defines psychologism as 'the doctrine that whether behavior is intelligent behavior depends on the character of the internal information processing that produces it' [8, p. 5]. According to this definition, two systems can display the same actual and potential behavior, have the same behavioral properties, capacities and dispositions, and yet, there could be a difference in their information processing prompting us to grant one full intelligence while holding that the other is devoid of any.

A classical argument against psychologism is the Martian argument: Suppose that there is life on Mars. Humans and Martians meet, develop an understanding of each other, engage in mental and creative activities together and so on. And then, it is discovered that Martians have significantly different information processing mechanisms than those of humans. Would we, then, deny that these creatures have intelligence just because they are very different from us? This would be, as Block likes to call it, pure "chauvinism". He holds that psychologism does not involve this kind of chauvinism. After all, psychologism does not state that the fact that a system has a completely different information processing mechanism compared to human beings *necessarily* means that it lacks intelligence.

Attacking the validity of the TT using psychologism does not seem to be Block's main interest. He is more concerned with arguing against behaviorism using the TT as a focal point.

As was mentioned above, Block believes, because of the characteristics peculiar to the design of the TT, some genuinely intelligent machines can be classified as lacking intelligence and vice versa. Here is what Block suggests in order to eliminate dependence on human discriminatory powers: 'We should specify, in a *non-question-begging* way what it is for a sequence of responses to verbal stimuli to be a typical product of one or another style of intelligence' [8, p. 10, emphasis added]. Then, Block suggests we revise our intelligence-granting mechanism as follows:

Intelligence (or more accurately, conversational intelligence) is the disposition to produce a sensible sequence of verbal responses to a sequence of verbal stimuli, whatever they may be [8, p. 11].

Now, the modified TT does not depend on anyone's coming up with good questions, since the system must have a *disposition* to emit sensible responses to anything that the interrogator *might* say, not just to the things that he/she *does* say. At this point, Block demonstrates that the modified TT is not greatly affected by the standard arguments against behaviorism¹³. The little defects of the modified TT as a behavioral conception of intelligence can be protected against these arguments with another modification. The reformulation involves the replacement of the term "disposition" by "capacity". The difference is that a capacity to \emptyset need not result in a disposition to \emptyset , unless certain internal conditions are met. Now, all arguments against behaviorism are avoided¹⁴ with *the neo-TT conception of intelligence*:

Intelligence (or more accurately, conversational intelligence) is the capacity to produce a sensible sequence of verbal responses to a sequence of verbal stimuli, whatever they may be [8, p. 18].

¹³The three mentioned by Block are the Chisholm-Geach, perfect actor, and paralytic and brain-in-vat arguments. Detailed discussion of these is beyond the scope of this work and is not crucial to the understanding of what follows. The interested reader is referred to [8, pp. 11-12] and the references provided there.

¹⁴See [8, p. 18].

Although Block seems to be ‘helping out’ the TT by making it less prone to anti-behaviorist objections, this is hardly a surprising consequence when the definition of intelligence is modified into something that is not really behaviorist any more. Block seems to be aware of this for he says the concession made to psychologism by moving from behavioral dispositions to behavioral capacities will not be enough to save behaviorism [8, p. 18]. His strategy is stretching behaviorism to its limits and showing that, even if we have the most general form of it, the behavioristic conception of intelligence is false.

How, one may wonder, will he do that? Block describes a machine that can produce a sensible sequence of verbal responses to verbal stimuli and is intelligent according to the neo-TT conception of intelligence. However, according to him, the information processing of the machine clearly demonstrates that it is devoid of intelligence. We have basically explained how this machine works in the introductory paragraphs of this section.

Block calls a string of sentences whose members can be typed by a human typist one after another in one hour or less, a *typable* string of sentences. The set of typable strings of sentences is clearly finite. The subset of this set consisting of those strings which can naturally be interpreted as conversations in which at least one party’s contribution is sensible is also a finite set. Recall that we assume we have a non-question-begging definition of “sensible”. The machine has the set of sensible strings recorded on tape. Upon encountering the string A , all it does is pick one of the strings that begin with A and type out the second sentence, say B , of A . If the interrogator types in the string C next, the machine will repeat the same process described above with A replaced with ABC and so on [8, p. 20]. This machine will have the capacity to emit sensible verbal output to any verbal input, and therefore would qualify as intelligent according to the neo-TT conception of intelligence. But the machine, in fact ‘has the intelligence of a toaster’ [8, p. 21]. This is primarily due to the fact that all the intelligence it exhibits belongs to the programmers, not to the machine itself. Block concludes that this capacity is not enough for intelligence and so the TT intelligence conceptions are refuted.

It can be argued that, by Block’s reasoning, *any* intelligent machine exhibits the intelligence of its programmers. Block says he is making no such claim. A

machine that has more sophisticated mechanisms like learning, and problem solving would, to Block, be intelligent. In the latter case, the exhibited intelligence belongs to the machine itself [8, p. 25]. The search machine of Block can only respond with what is already put in its memory by the programmers¹⁵. Block argues that ‘the neo-Turing Test conception of intelligence does not allow us to distinguish between behavior that reflects a machine’s *own* intelligence and behavior that reflects *only the intelligence of the machine’s programmers.*’ [8, p. 25, emphasis original]. This kind of argument has been considered by Turing in [127, p. 450], as described briefly in Section 2.2.2.

Another objection is as follows: Block is merely suggesting a new definition of intelligence by stipulating certain internal conditions. Block defends the new definition here, which is presuppositional of its existence! Therefore, Block is indirectly admitting that all he is doing is suggesting that we adopt new criteria for intelligence and dispose of the behavioristic ones [8, p. 27].

Block also considers the “chauvinism” argument. A system with information processing unlike ours may not be “intelligent” in our criteria; but then, *we* might not count as “shmintelligent” in their criteria. ‘And who is to say that intelligence is any better than shmintelligence?’ [8, p. 27]. Block denies the chauvinism attributed to him. He believes ‘[his] machine lacks the kind of “richness” of information processing requisite for intelligence’ [8, p. 28]. He doesn’t feel the need to elaborate on what sort of systems have the above-mentioned richness believing that ‘one can refute the Turing Test conception by counterexample without having to say very much about what intelligence really is’ [8, p. 28].

To those who ask what Block would think if it turned out that humans process information in the way that Block’s machine does, Block responds as follows:

If the word “intelligence” is firmly anchored to human information processing, as suggested above, then my position is committed to

¹⁵Consider, however the following situation: If every once in a while, upon verbal input *A* the machine transformed a sentence *B* in *A* into *B'* and proceeded accordingly (this can be likened to a *mutation*), would it count as intelligent because of this little trick of non-determinism?

the *empirical claim* that human information processing is not like that of my machine. But it is a perfectly congenial claim, one that is supported by both common sense and by empirical research in cognitive psychology [8, p. 29, emphasis original].

It can be argued that Block's machine is unrealizable because of combinatorial explosion. We will not go into the details of this; Block's response to this objection can be found in [8, pp. 30-34].

Richardson, in his reply to Block [112], states that he is doubtful whether Block's machine can really imitate human conversational abilities. Humans can (and do) understand sentences that they never heard/uttered before and produce sentences that they never heard/uttered before. They can do this in such a way that they can adapt to novel situations and maintain the coherence of discourse. This view is held by Chomsky who believes that the brain *cannot be a repertoire of responses* and must contain a program that can build an *unlimited set of sentences* out of a finite list of words¹⁶. If the *potentially* utterable/understandable and sensible sentences that a human mind can produce in a lifetime is unlimited, then how can a team of humans gather this information and enter it in the memory of the machine in finite amount of time? It is difficult to imagine Block's machine managing the many intricacies of human conversation such as adapting to topic shifts and contextual changes. Richardson believes 'if the list-searcher satisfies the neo-Turing Test, the test is too weak' [112, p. 423]. For Block's response to such arguments see [8, pp. 35-36].

Block must have realized some difficulties in enumerating the strings as well. He later introduces, in [9], the Aunt Bubbles machine¹⁷. In this version, the programmers think of *just one* response to the strings at each step. To maintain coherence and make the task easier to follow, they may choose to simulate a definite person, for instance Block's own (most probably hypothetical) Aunt Bubbles. They may even restrict the situation by modeling Bubbles' responses in the case that she is brought into the teletype room by her strange nephew [9]. So each response is the kind of response that Aunt Bubbles would give to the

¹⁶See, for instance [15].

¹⁷Aunt Bubbles appears momentarily, as Aunt Bertha, in [8] too.

verbal inputs. Block says that the machine will do as well as Aunt Bubbles herself in a TT, but it is obviously not intelligent because of the reasons described above.

Let us briefly go over some of Block's arguments and the behaviorism in the TT before we proceed. For one thing, as Block also mentions, the intelligence concept (because of some inherent properties it has) does not fully conform to the generalizations of behaviorist or anti-behaviorist arguments based on other mental states such as pain [8, pp. 13-16]. There is another aspect of intelligence that can justify for the behaviorist approach of the TT. Behaviorism may be considered an antiquated or primitive approach in general, but it does not seem that awkward to use it in intelligence-granting. This is primarily because *we* grant intelligence that way: Upon seeing a human being we automatically assume that he/she is intelligent. We feel free to approach a person (rather than, say, a dog or a lamp post) to ask the whereabouts of the post office without having many doubts about him/her understanding us. If the TT is *that* crude and unsophisticated, then we, as humans might consider revising our intelligence-granting mechanisms as well. This constitutes a line of defence for the TT: If behavioral evidence is acceptable for granting intelligence to humans, this should be the case for machines as well. I have discussed this already in Section 2.2.2.

Recall that Block believes humans can be overly chauvinistic or liberal in granting intelligence to machines. However, it is unclear how he classifies genuinely intelligent machines and mindless machines. If there is a way of deciding on that issue, an *X-Test* to determine whether a machine is really intelligent, then why would we be discussing the TT with all its quirky and imperfect ways? In addition, although he does not trust the human judges in the beginning, later on Block seems to have complete faith in the '*imagination and judgement* of a very large and clever team working for a long time with a very large grant and a lot of mechanical help' [8, p. 20, emphasis original].

With the current research on cognition and linguistics at hand, it seems unlikely that a simple approach like Block's can succeed in modeling the human mind. If one day, enough on language and cognition is discovered so that Block's "sensible" strings of sentences are enumerated then we may decide that

the neo-TT conception of intelligence is false. But then again, when that day comes, having all the *psychologicistic* information we need, we probably would not be interested in the TT any more.

The reader is encouraged to contemplate on how Robert French's arguments, which are presented in Section 2.4.5, affect Block's claims. We believe that French's demonstration of the inseparability of the cognitive and subcognitive levels in the human mind has direct consequences on the realizability of Block's machine.

In any case, Block's paper is significant because it demonstrates the weakness of the behavioral approach in the TT. The TT may be abandoned one day, because more information on how the mind works may be obtained and we may have on our hands better means to detect another entity's cognitive capacities. But today, we do not have much to look at that is more informative than behavior.

2.4.2 The Chinese Room

In the beginning of the 80's, with John Searle's Chinese room argument [118] the TT was confronted with yet another objection. The analysis of the Chinese room can easily get out of hand since a great number of comments have been made on the issue and the debate still rages on.

In a nutshell, here is what the Chinese room looks like: Suppose that Searle, a native speaker of English who does not know a word of Chinese, is locked in a room. There is an opening in the room through which we may send in Chinese sentences on pieces of paper. Of course to Searle, these look like meaningless squiggles [118]. In the room, Searle has a "Chinese Turing Test Crib Book" [83] he can consult to find an output that corresponds to each Chinese symbol he receives. What he does is simply match the input with those in the book, follow some rules written in English and find some Chinese symbol sequence to output. We correspond with Searle in this manner and due to the flawless look-up table he has, Searle-in-the-room seems to understand Chinese perfectly. But he does not. Searle still has no idea about what the Chinese symbols we send

in and those that he sends out mean. To him, “Squiggle-Squiggle” is coming in and “Squoggle-Squoggle” is going out [58].

Now consider a computer program that passes the TT in Chinese. Proponents of the TT will grant it that this computer thinks and, in some sense, understands Chinese symbols. Searle challenges this by being the computer and yelling at the world that he does not understand a word of Chinese. Judging by the inputs and outputs of the system, Searle-in-the-room is indistinguishable from a native speaker of Chinese. In a sense, he is passing the TT in Chinese, without understanding a word of Chinese. It should be clear how that constitutes a criticism of the TT, and the computational view of mind.

As was mentioned before, various aspects of the Chinese Room argument have been analyzed including syntax/semantics, consciousness, boundaries of systems, etc. The interested reader is referred to [118, 119, 56, 4, 29, 20, 23, 111, 37, 65, 10, 89, 113, 67] and the references provided in those.

2.4.3 Consciousness and the TT

Another difficult and widely discussed problem in philosophy of mind is consciousness. While I do not want to delve too far into this, I wish to take a brief look at the relationship between consciousness and the TT.

Donald Michie’s ‘Turing’s Test and Conscious Thought’ [98] is one of the important comments made on the TT. Michie comments on a variety of issues surrounding the TT, but in this section I mainly concentrate on the conclusions he draws about consciousness.

First of all, Michie notes that Turing did not specify whether consciousness is to be assumed if a machine passes the TT. Of course, Turing probably did not believe that consciousness and thought are unrelated. Rather, Michie thinks he means ‘these mysteries and confusions do not have to be resolved before we can address questions of intelligence’ [98, p. 31] (see also [127, p. 447] and Section 2.2.2). There seems to be a relation between consciousness and thinking. Some critics believe that intelligence cannot be granted to entities

that are not conscious (see, for instance, [119]) while others have questioned the interdependence of conscious and subconscious processes (see, for instance, [40] and Section 2.4.5).

According to Michie, that the TT provides access to cognitive processes via verbal communication incapacitates it as a test of intelligence. He observes two dimensions in which this inadequacy manifests itself.

The first is ‘the inability of the test to bring into the game thought processes of kinds which humans perform but cannot articulate’ [98, p. 36]. Michie gives examples of some operations humans can perform almost unconsciously. For instance, any English speaker would be able to answer the question “How do you pronounce the plurals of the imaginary English words ‘platch’, ‘snorp’ and ‘brell’?” with “I would pronounce them as ‘platchez’, ‘snorpss’ and ‘brellz’.” [98, p. 38]¹⁸. It is conceivable that the programmers of TT-passing programs will be forearmed against this particular question, but it is unlikely that they can encode all we know about pronunciation (or phenomena from non-linguistic domains, for that matter) simply because some related processes operate at the subconscious level. (For a similar argument, the reader is referred to [40] and Section 2.4.5.)

The second dimension in which Michie believes the TT to be mismatched against its task is the phenomenon of machine ‘superarticulacy’. Namely, ‘the test can catch in its net thought processes which the machine agent *can* articulate, but should not if it is to simulate a human’ [98, p. 42]. As was mentioned above, humans perform many activities without being fully aware of how they do them. In fact, it has been shown that the better you get at something the less aware of the underlying processes you become. Thus during a TT, ‘the interrogator need only stray into some specialism in which both human and machine candidates possess a given expertise’ [96, p. 192]. The machine will give itself away because of its superarticulacy. For more about superarticulacy, the reader is referred to [98, p. 41-43] and [95].

Finally, Michie notes the importance of social intelligence. AI should, he

¹⁸This question was adapted from [79].

says, try to incorporate emotional (also called affective) aspects of communication and thought in the developed models. Michie also proposes, like some of those we will see in the next section, that extensions to the TT can be made in order to ‘address yet more subtle forms of intelligence, such as those involved in collective problem solving by co-operating agents, and in teacher-pupil relations’ [98, p. 51].

I cut the discussion of consciousness short both because it is a rather broad topic, but also because most commentators on the TT (consciously or subconsciously) propose arguments that can be interpreted from that angle. Can we not reformulate the other minds problem (“How do I know that any entity other than me has a mind?”) in terms of consciousness (“How do I know that any entity other than me is conscious?”)? The reader can refer to Section 2.2.2 and [127, pp. 445-447] for Turing’s answer to the argument from consciousness and how he makes use of the other minds problem. Similarly, most questions about machine thought can be re-evaluated within the context of machine consciousness. I include the analysis of Michie’s paper here because it proposes new ideas from the viewpoint of consciousness and relates them explicitly to the TT. Interested readers can consult [27, 54, 96, 97] for more on consciousness.

2.4.4 Alternative Versions of the TT and Their Repercussions

In this section, we will summarize some alternatives to the TT that were proposed in order to assess machine intelligence.

Harnad and the TTT

Stevan Harnad’s main contribution to the TT debate has been the proposal of the Total Turing Test (TTT), which is, like the TT, an indistinguishability test but one that requires the machines to respond to all of our inputs rather than just verbal ones. Evidently the candidate machine for the TTT is a robot with sensorimotor capabilities [56, 58].

Harnad's motivation for the 'robotic upgrade of the TT to the TTT' [58] has its roots in what he calls 'the symbol grounding problem'. He likens the situation of symbols being defined in terms of other symbols to a merry-go-round in a Chinese to Chinese dictionary [57]. He claims that for there to be any semantics in the mind (and there surely is) symbols must be *grounded*. Harnad deduces that meanings of symbols are, at least in part, derived from interactions with the outside world.

Harnad does not explicitly argue that the TT is too specific (unlike Gunderson for instance, see Section 2.3.1). He concedes that language might capture the full expressive power of our behavior, at least when the concern is assigning minds. What he doubts is whether language is an 'independent module' [56]. His position is summed up in the following:

Successfully passing the teletype version of the Turing Test alone may be enough to convince us that the candidate has a mind just as written correspondence with a never-seen penpal would, but full robotic capacities even if only latent ones not directly exhibited or tested in the TT may still be necessary to generate that successful linguistic performance in the first place. [58, p. 46].

Harnad also defends his TTT against the Chinese room, showing that his arguments are not adversely affected in the light of the latter, which, in my opinion, is uncalled for. The motivation of the TTT is quite clear; Harnad's assertions, although debatable, are understandable. An approval from Searle would not make that much of a difference, but Harnad seems to think it is important. In any case, by doing so, he enables others to criticize his work on Searlean accounts [64, 12].

Larry Hauser, in his reply to Harnad's 'Other Bodies, Other Minds' [58], criticizes Harnad and Searle and aims to show that 'Harnad's proposed robotic upgrade of the TT to the TTT is unwarranted' [64, p. 234]. To that end, he analyzes Harnad's intuitive, scientific and philosophical reasons for proposing the upgrade and argues against them. Hauser finds the TTT to be unnecessary because, he notes, if the sensorimotor capacities the TTT tests for are *necessary*

for the linguistic capacities that the TT tests for, having the latter should be *sufficient* for the former anyway [64, p. 227].

For more on symbol grounding and the TTT, the reader is referred to Harnad's other papers [59, 60, 61]. Also interesting is James H. Fetzer's 'The TTT is not the Final Word' [34], in which he aims to show that the TTT cannot provide a proof for machine thought since more than symbol manipulation *and* robotic capacity should be involved in intelligence-granting.

In addition to the TTT, Harnad also mentions a TTTT (Total Total Turing Test) which requires neuromolecular indistinguishability. However, this more stringent version of the TT, according to Harnad, will be unnecessary. Once we know about how to make a robot that can pass the TTT, he says, we will have solved all the problems pertaining to mind-modeling. However, neural data might be used as clues about how to pass the TTT [58]. Harnad believes '[TTTT] is as much as a scientist can ask, for the empirical story ends there' [61], but he does not think that we have to "go that far". The reader is referred to [61] for a detailed explanation of why Harnad believes the TTT is enough. For an excellent third person account of the TT/TTT/TTTT story, among other issues, the reader is referred to [35].

The Argument from Serendipity and the Kugel Test

Stringent versions of the TT are also mentioned by Selmer Bringsjord, occasionally within the context of Harnad. Bringsjord supposes that there is a sequence of TT variants in increasing order of stringency. In his "What Robots Can and Can't Be" [11] he aims to show that AI will produce machines that will pass these stronger versions, but the attempt to build an artificial person will still fail.

Bringsjord is one of those who wants to remain within "the philosophical terrain". In [12] he develops against the TT, *the argument from serendipity* and defends this against some criticisms.

The argument from serendipity, as the name suggests, "refutes" the TT by a finite state automaton (FSA) that generates random English sentences. Call

this automaton P . During a TT, P may just get lucky and fool the interrogator. So much for the TT! Even the TTT can be refuted similarly. A robot may behave randomly and by chance, its linguistic behavior may coalesce with the sensorimotor behavior perfectly during a TTT.

Bringsjord finds the TTTT very chauvinistic and considers an alternative version of it he calls TTTT*. This latter test requires a flowchart match between the brains of players A and B rather than a neuromolecular match [12, p. 104]. But Bringsjord believes that the TTTT* is an ‘impracticable nightmare’ since we would not know how to conduct this test. The interested reader should consult [12] to see Bringsjord explain his reasoning where he appeals both to intuition and computability theory.

Bringsjord, determined to attack every version of the TT, also “refutes” the Kugel Test (KT). The KT is not as well known as the TTT or the other versions of the TT that we investigated in this section. Once again, there are three players involved. A judge, who sits behind two bins marked *YES* and *NO*, runs the game. The aim of the participants is to guess the concept that the judge thinks up by looking at the cards (with pictures on them) that the judge drops in the two bins. A card goes to the *YES* bin if it falls under the concept, and to the *NO* bin otherwise. To give an example, if the concept that the judge is thinking of is “woman”, cards with pictures of women (or conceivably, items typically identified with women) go to the *YES* bin. A player need not announce the concept when he/she finds it. He/she *wins* the round if there comes a time at which all future guesses about which bin a card will be placed in are correct [78, p. 4]. Thus the player must not only identify the concept (e.g., just say “Aha! The concept is *woman*.”) but should also be able to apply it. Now, just as in the TT, to pass the KT, a machine has to perform as good as a human. An interesting twist here is that the machine must be able to *win the game*, which is not the same as winning a round. A game consists of infinitely many rounds.

Why, you may ask, would anyone design such an obscure test? Kugel, by requiring the machine to win infinitely many rounds, wants to rule out the possibility of an FSA passing the KT [77, 78]. Although the test is practically useless (because it requires infinite amount of time), is it of any theoretical

significance? Kugel believes that humans are neither *pigheaded* (i.e., once they think of an answer to the “sequence game” they do not have to stick with it) nor *narrow-minded* (i.e., once they find the n th member of a sequence, they are still able to learn a different sequence with the same initial elements). If humans were Turing machines (or FSA’s with lesser powers) they would be pigheaded and narrow-minded. Kugel holds that humans are automata of some sort, and in the light of the above concludes that they must be trial-and-error machines. For more on KT, the reader is referred to [77, 78, 12].

Bringsjord is interested in the KT primarily because it rules out FSA’s from passing it¹⁹. He notes that Kugel’s arguments may be unsound, but assuming they are not, asks the question “Do we have in KT an acceptable variant of the original TT?” [12, p. 115]. Bringsjord’s answer is negative. KT is rigid and does not allow access to all cognitive capacities that the TT can. I agree with this criticism of Bringsjord; participants in the KT are rather passive and their innovative (or rather, generative) capabilities cannot be tested. Bringsjord’s second argument against the KT is again from serendipity. A trial-and-error machine can call the random string generating FSA P mentioned above for the declarations about what the concept in question is and so much for the KT... Once again, the reader is to consult [12] to see how the argument from serendipity is “guaranteed to work” against the TT and its variants.

The Inverted Turing Test

Recently, Stuart Watt has proposed the Inverted Turing Test (ITT) [130]. Some of his arguments are also relevant to Section 2.5 but I chose to discuss his paper here.

Watt’s point is that the TT is inseparable from “naive psychology”²⁰ since to pass the TT, a machine has to convince the interrogator that it has a mind. He calls naive psychology ‘the psychological solution to the philosophical problem’ [130]²¹.

¹⁹Recall that his argument from serendipity features an FSA.

²⁰Basically the term given to the natural human tendency and ability to ascribe mental states to others and to themselves [130].

²¹The latter being the other minds problem.

Watt's ITT requires the machine to be able to prove its human-ness by exercising naive psychology. In particular, it has to show that its power of discrimination is indistinguishable from that of a human judge in the TT. The TT is literally inverted and 'a system passes [the ITT] if it is itself unable to distinguish between two humans, or between a human and a machine that can pass the normal TT, but which can discriminate between a human and a machine that can be told apart by a normal TT with a human observer' [130].

Watt states that he proposes the ITT as a thought experiment rather than as a goal for AI. Incidentally, he believes that the same applies to the TT and both tests should be regarded as means to gather inductive evidence on which inferences about machine mentality can be made [102]. I had discussed this earlier in Section 2.3.4.

Watt may be right about intelligence being in the eye (or the mind) of the beholder; many people have noted the human disposition to ascribe intelligence to systems that aren't and vice-versa. But the new test he proposes, the so-called ITT, has been subject to some strong counter-arguments as we shall shortly see. It can be said that Watt's motivation for introducing the ITT seems acceptable, but the proposal itself is problematic ²².

Selmer Bringsjord and Robert French, in their replies to Watt [13, 42], propose simple methods that reveal some weaknesses of the ITT. The titles of the papers are illustrative of their content. Bringsjord's 'The Inverted Turing Test is Provably Redundant' [13] shows that the ITT is entailed by the original TT. Bringsjord also opposes Watt's motivation and believes that naive psychology is withering in many humans (including himself) and, with the advent of computer programs that are very difficult to distinguish from humans in written communication, will soon be no more.

In 'The Inverted Turing Test: A Simple (Mindless) Program that Could Pass It' [42], Robert French shows both that the ITT can be simulated by

²²During the discussions I held after a talk on the Turing Test, (at the Cognitive Science Colloquium held at the Middle East Technical University, Ankara, in November, 1998) many participants, who did not previously know a lot about the topic *proposed* the ITT as a better means to detect human-ness of machines. These people had not read or heard of Watt's paper and neither the ITT nor naive psychology was discussed during the presentation. Maybe this can be explained as "naive psychology at work".

the TT (in a way that is very similar to Bringsjord's) and that a very simple program can readily be designed to pass the ITT. The mindless machine that will pass the ITT is designed using 'subcognitive questions' that are described in [40, 41]. It is assumed that the conclusions explained by French in these works are accepted. These are analyzed in substantial detail in Section 2.4.5.

First a large set of subcognitive questions are selected, humans are surveyed and a 'Human Subcognitive Profile' for this 'Subcognitive Question List' is obtained. Now, if we give these and a statistical analyzer to an interrogator (man or machine), he/she/it should have no difficulty discriminating machines from humans. It is not difficult to store the list and the profile in the memory and provide a small statistics routine to the computer, and so much for the ITT. While the TT stumbles in front of subcognitive questions (see Section 2.4.5), they can be used to construct a mindless machine that can pass the ITT.

Others have used their replies to Watt as opportunities to voice their opinions about AI and the Turing Test in general. As we shall see in Section 2.4.6 Patrick Hayes and Kenneth Ford view the TT as a harmful burden on AI. In their 'The Turing Test is Just as Bad When Inverted' [38], they state that the ITT suffers from the same problems as the TT that they explicate in [68]. They grant it that Watt has a point in his arguments on naive psychology but note that Turing's original IG (the gender-based TT) is immune to most of those since in this scenario, the interrogator will not be thinking about differences between humans and machines. In any case, they believe that 'it is time for AI to consciously reject the naive anthropomorphism implicit in all such "imitation games" and adopt a more mature description of its aims' [38].

Similarly, Collins, in his reply to Watt [22], does not really focus on the ITT, but proposes a new variant of the TT. He believes that 'the deep problem of AI' is that of trying to develop machines that can learn from their surroundings the way humans do. There is currently an 'interpretive asymmetry' between the way humans and computers do things. Machines are not as adaptive as humans in human-computer interactions. According to Collins, this asymmetry will disappear when computers reach a level of sophistication in resolving mistakes and learning from their surroundings that is comparable to those of humans and all problems of AI will be solved. Learning languages would then be one of

the surface transformations of this deep problem [21] and when this is solved ‘the rest will be research and development’ [22].

To determine whether the interpretive asymmetry has disappeared, Collins believes we can use Turing-like tests. In fact he states that a sub-TT is enough to assess whether this goal has been reached or not and complicating the matter by proposing the ITT or the TTT is uncalled for. In the Editing Test (ET) that Collins proposes, the task is no longer as comprehensive as holding a conversation, but that of sub-editing previously-unseen passages of incorrect English. The interrogator will try to come up with pieces of text that a linguistically competent human can easily sub-edit and if a computer is indistinguishable from humans in this task, then the ET is passed and the deep problem of AI is solved. Collins finishes by briefly demonstrating that even the ET is very difficult to pass, at least with the currently imaginable techniques (such as a look-up table) [22].

The Truly Total Turing Test

Very recently, in his interesting *Minds and Machines* paper [117], Paul Schweizer proposed the ‘Truly Total Turing Test’ (TRTTT)²³. He believes even Harnad’s TTT to be an insufficient test for intelligence. Before he proposes the TRTTT, Schweizer states his own opinions about the adequacy of behavioral criteria. He views such tests as ‘dealing with evidence for intelligence but not as constitutive or definitional’ [117, p. 264].

Schweizer, while talking about the other minds problem, notes that we usually grant intelligence to other humans on behavioral bases because we have general knowledge about the *type* of creature under consideration. However, in the TT, we encounter a type about which we do not know anything. In the case of machines we lack a “history” to base our decisions upon.

Schweizer believes that the TT, and even Harnad’s TTT, is subject to the “toy-world” criticism. The systems that succeed in these tests would, according

²³In Schweizer’s paper, the abbreviation TTTT is used. I prefer to use TRTTT so as to avoid confusion with Harnad’s Total Total Turing Test, previously referred to as TTTT

to him, not be displaying an intelligence comparable to the natural intelligence of living things that function in the real world. They can function only in constrained, artificial worlds.

The TRTTT posits a long-term, evolutionary criterion: Consider cognitive *types* and *tokens* of those types. Although we do not have a theory of the intelligence of the human cognitive type, we have an extensive *historical record* of it [117, p. 267]. This is precisely why behavioral intelligence-granting is acceptable for individual humans (tokens of the type human). Thus robots, as a cognitive type, should accomplish achievements that are comparable to those of humans. It is no more enough to converse in natural language or to play chess; robots as a ‘race’ must be able to *develop* languages and *invent* the game of chess. Similar (evolutionary) tests have been proposed by others before but never so convincingly²⁴. Schweizer makes very good use of the other minds problem to support the cultural and cognitive evolution criteria that the TRTTT stipulates.

Now, after the *type* passes the TRTTT, we can evaluate *tokens* of the type by less stringent behavioral tests, like the TTT and the TT. According to Schweizer, ‘imitative tests like the TTT and the TT apply to individuals *only* under the assumption that the general type is capable of passing the [TRTTT]’ [117, p. 268, emphasis original].

2.4.5 Subcognition and Robert French

One of the more recent discussions about the TT can be found in Robert French’s 1990 article ‘Subcognition and the Limits of the Turing Test’ [40]. In this work, French aims to show that ‘the Turing Test provides a guarantee not of intelligence, but of culturally-oriented *human* intelligence’ [40, p. 54].

French considers two claims of Turing. The first is the claim that if a computer passes the TT, it will necessarily be intelligent. The second one is the claim that it will be possible to build such a machine in the near future. These, he calls the philosophical claim and the pragmatic claim, respectively.

²⁴See [5] and Section 2.5.

French agrees with the former claim. However, he believes that the pragmatic claim has been largely overlooked in discussions of the TT. In ‘Subcognition and the Limits of the Turing Test’, he is primarily concerned with this latter claim and believes that the TT is ‘virtually useless’ [40, p. 53] as a real test of intelligence because he believes it will never be passed.

To establish this result French considers “subcognitive” questions, i.e., questions that reveal low-level cognitive structure²⁵. French argues that any sufficiently broad set of questions for a TT will contain subcognitive questions, even if the interrogators do not intend to ask them. The fact that the cognitive and subcognitive levels are intertwined in such a way, in turn, shows that the TT is essentially a test for human intelligence, and not for intelligence in general.

First, let us consider an interesting analogy French makes: The Seagull Test. Consider a Nordic island on which the only flying animals known to the inhabitants are seagulls. One day, two philosophers are discussing the essence of flying. One of them proposes flying is moving in the air. The other objects by tossing a pebble and stating that the pebble certainly is not flying. The first philosopher stipulates that the object remain aloft for a period of time for the activity to count as flying. But in this case clouds, smoke and children’s balloons qualify as flying entities, the other argues. Then the philosopher questions whether wings and feathers should be involved but this is immediately refuted by the latter by pointing to penguins. While the arguments continue to be inconclusive, they agree on a few facts: The only flying objects known to them are the seagulls on their island. Flight has something to do with being airborne, physical characteristics like feathers, beaks are probably not involved. They, then, in the light of Turing’s famous article, devise a Seagull Test for flight. They believe if something can pass the Seagull Test, it is certain that it is able to fly. Otherwise, no decision can be made; maybe it can fly, maybe it cannot [40].

The Seagull Test works as follows: There are two three-dimensional radar

²⁵Here, low-level cognitive structure refers to the subconscious associative network in human minds, consisting of highly overlapping activatable representations of experience [40, p. 57].

screens, one tracking a seagull and the other tracking the flying object attempting the test. The object will pass the test only if it is indistinguishable from the seagull on the radar screen. The similarity between our TT and the Seagull Test is evident. The arguments about the gist of flying between the two philosophers is an uncanny reminiscent of the arguments on the nature of intelligence. The test itself is an almost direct analogue of the TT.

The Seagull test as it is, cannot be passed by airplanes, helicopters, bats, beetles or sparrows. It is doubtful that *anything* can pass it. That is, except for the Nordic seagulls of the philosophers' island. Then, 'what we have is not a test for flight at all, but rather a test for flight as practiced by the Nordic Seagull' [40, p. 56]. The analogy makes it clear what French thinks about the TT. Correspondingly, he believes that the TT is a test for intelligence as practiced by a human being.

Before proceeding with his main arguments about subcognitive questions, French makes some assumptions that he feels are acceptable. First, he allows the interrogator to poll humans for the answers of some questions before asking them during the test with the machine. Another assumption is that all the people involved are from the same culture and the machine is attempting to imitate an individual from that culture²⁶. And finally, French assumes that although building a machine that can experience the world in a way indistinguishable from that of a human being (from tasting strawberries to falling off of bikes) is theoretically possible, this is not likely to occur in the foreseeable future.

The first class of questions that French introduces are explicitly designed to reveal low-level cognitive structure. These are based on research in cognitive psychology, in particular, associative priming. For instance, it has been established that if human subjects are presented with a set of words and non-words, they generally require less time to recognize that a string is a word if it is preceded by a related word, rather than a non-related one or a non-word. To give a specific example, the item "salt" should be recognized faster when

²⁶French believes that this assumption is tacitly made by Turing. The importance of culture in conversation and communication is immense (see Section 2.5) and this could be a reasonable stipulation.

preceded by “pepper” and not by an unrelated word like “shoe” or a non-word like “gloff”.

Before questioning the computer, the interrogator runs a similar experiment on a group of people²⁷ and calculates average recognition times. During the test, the interrogator can distinguish between the machine and the human simply by looking at the recognition times on the same set of words on which the experiment had been carried out. It is hard to imagine a machine being successful at deceiving the interrogator in such a case. The only possible way to carry out associative priming in a manner that is similar to the average human’s seems to be having experienced life as an average human.

It may seem unfair to expect computers to match scientific, experimental data about subcognitive capacities of the human mind. This view might even be exaggerated into expecting them to have a physical construction like ours. Whether such an objection is valid or not, French proposes some questions that are (or rather seem) at a higher cognitive level. The first kind is called the Neologism Rating Game. The interviewees (and later the machine) will be given a set of made-up words and will be asked to rank them on their appropriateness for defining certain entities. Participants in the game are asked to rate, on a scale of 0 (completely implausible) to 10 (completely plausible), whether a certain neologism can be a name for something. French proposes a set of such questions that are especially demonstrative. Here, I only consider two of those neologisms: “Flugblogs” and “Flugly”.

According to French, “Flugblogs” would make an inappropriate choice for the name of a cereal since the initial syllable, “flug”, is phonetically similar to “flub”, “thug”, “ugly” and “ugh!” and the second syllable, “blogs” phonetically activates “blob”, “bog”, etc. As can be seen, these words do not really sound very appetizing and they each carry an *aura* of semantic connotations that renders them unsuitable choices as syllables of a cereal name. However, “Flugblogs” would be a very appropriate name you would give to big, bulbous, air-filled bags used to walk on water. In this case the semantic connotations of the syllables like “flug”, “blob” and “bog” are in accordance with the proposed meaning. Similar analysis of “Flugly”, which activates friendship, coziness and

²⁷In French’s terminology these human subjects are called *interviewees*.

cuteness, reveals that it is a plausible name for a child's teddy bear. The same name, although it has positive connotations, would sound awkward as the surname of a glamorous movie star.

The arguments above are highly intuitive, and although most of us would agree on them, we do not know precisely how we come up with the connotations. We do know, however, that these happen due to a large number of culturally acquired associations. We do not have control over the accumulation of such associations; they are pumped into our brains in daily life as brand names, advertising slogans, names of pets and stereotypes of various sorts²⁸. Moreover, it is not possible to program these into the computer since neologisms are virtually infinite in number. French believes that the computer's chances would be very low when the interviewees' responses to such questions are compared to those of the human and the computer in the IG.

Another game of a similar nature is the Category Rating Game in which the questions are of the type "Rate *Xs* as *Ys*", where *X* and *Y* are any two categories. Again, French gives several illustrative examples [40, p. 61]. Consider, for instance, "Rate *dry leaves* as *hiding places*". The definition of dry leaves does not contain anything explicitly stating they might be good hiding places for children, and yet 'few among us would not make that association upon seeing the juxtaposition of these two concepts' [40, p. 60]. If we are asked to rate, on a scale of 0 to 10, most of us (those who have seen a James Bond movie at some point in their lives) would certainly rate "*pens* as *weapons*" higher than, say, "*grand pianos* as *wheelbarrows*". Again the answers to the Category Rating Game questions are highly dependent on our having experienced life as a human being in a certain social and cultural setting.

Now that we have studied French's incisive subcognitive questions, let us see how he uses these to refute the TT as a useful test for intelligence. The main claim of French is that the physical level and the cognitive level of intelligence are inseparable. The subcognitive questions reveal information about the low-level cognitive processes of the entities answering them. In a way, if used during

²⁸The importance of cultural factors becomes evident in this context. Without having notions of Kellogg's and teddy bears, the answers to these questions would be near-random guesses.

the TT, these would allow the interrogator to ‘peek behind the screen’ [40, p. 62]. These questions allow comparison of the associative concept networks of the two candidates. And because these networks are formed after a lifetime of experiencing the world and the structure and nature of them are necessarily dependent on physical aspects of that experience (like human sense organs, their locations in the body, etc.), the computer will be distinguishable from the human. In short, it is not possible for a computer (or any other non-human) to be successful in playing the IG. Not having experienced the world as we have is not just an obstacle, but a severe restriction in this task. This is due to the fact that the TT is a test for human intelligence, just as the Seagull Test is a test for Nordic seagull flight.

French considers whether there can be a modification of the TT that does not reduce the computers’ chances of passing it to zero. He explains the impossibility of this as follows:

Surely, we would not want to limit a Turing Test to questions like “What is the capital of France?” or “How many sides does a triangle have?”. If we admit that intelligence in general must have *something* to do with categorization, analogy-making, and so on, we will of course want to ask questions that test these capacities. But these are the very questions that will allow us, unflinchingly, to unmask the computer [40, p. 63].

French repeatedly states, as was mentioned above, that the TT is a test for *human* intelligence. It may seem like by proposing subcognitive questions he is stipulating that a human subcognitive substrate is *necessary* for intelligence in general, but this is only apparent. What French really attempts to demonstrate, as he explains, is that the human subcognitive substrate is necessary to pass the TT (as the subcognitive questions show), and TT is inadequate precisely because of this. He holds that this substrate is definitely not necessary for intelligence in general, just as being a Nordic seagull is not a necessary condition for flight.

French’s paper is significant in one sense: Instead of discussing whether passing the TT is a sufficient or necessary condition for machine thought, he

asks whether the test can be passed at all. Let Searle have his room and Block his Aunt Bubbles. French reminds us that the TT is difficult when you leave your armchair.

A criticism of French's 'Subcognition and the Limits of the Turing Test' [40], has been made by Dale Jacquette in [73]. For French's response to Jacquette, the reader should refer to [41].

2.4.6 Getting Real

As I mentioned in the beginning of this section, the more interdisciplinary approach that seems to prevail in the discussions of the mind also had effects on the way we philosophize on the TT. Thus, the 90's became a time during which it was not so easy to get away with proposing wild thought experiments and leaning back on your armchair to watch the fuss over them. Stevan Harnad expresses an impatience that many were beginning to feel as follows:

If you want to talk about what a model or a simulation can or cannot do, first get it to run. [56, p. 4].

Recently, Justin Leiber has argued that the TT has been misinterpreted [84]. He notes that Block's²⁹ and Searle's counter-arguments do not refute the TT. Among the reasons Leiber lists for this are practical issues like memory, reliability and speed. Leiber views the TT as an operational definition and states that 'our problem [is] one of engineering' [84, p. 65]. His position is similar to that stated by Harnad:

What you need to face Turing's Turing Test is a reasonably detailed description of a machine which can indeed be supposed to pass the Turing Test in real time but which somehow is not really thinking. [84, p. 61].

²⁹Although Leiber is mainly concerned with the homunculi argument in 'Troubles with Functionalism' [7], his response also applies to Block's attack of the TT described in Section 2.4.1.

At one extreme are Patrick Hayes and Kenneth Ford, who state that we should reject the goal of passing the TT in their ‘Turing Test Considered Harmful’ [68]. They believe that passing the TT is a distraction for “useful” AI research.

Hayes and Ford believe that AI’s ultimate goal should not be that of imitating human capabilities. Since the TT’s sole aim is precisely that, they believe that ‘it is time to move it from textbooks to the history books’ [68, p. 972]. They also see a problem with the gender issue in the IG:

The gender test is not a test of making an artificial human but of making a mechanical transvestite [68, p. 973].

[Turing] tells us quite clearly to try to make a program which can do as well as a man at pretending to be a woman [68, p. 977].

As was mentioned in Section 2.2.1, this peculiarity might have its reasons, but Hayes and Ford have a moral objection concerned with the artificial constraints the setting imposes on the participants of the game.

Hayes and Ford also express their inability to find a practical use for the TT. Why on earth should we work that hard (and it *is* hard) to build a machine that imitates us? To depict the uselessness of direct imitation of humans in AI, they resort to a very popular analogy: mankind’s futile attempts at making flying machines by the imitation of natural flight. Artificial intelligence, like artificial flight, can be radically different from natural flight. And it can still be a good thing. Hayes and Ford believe that even if one’s goal is trying to understand humans, there is no reason to define all that there is about cognition in terms of human cognition.

Their belief that AI is a field that strives to be useful leads Hayes and Ford to deny passing the TT as a sensible goal. They hold that AI should produce cognitive artifacts, not necessarily in a human way, but in a way useful to humans.

Blay Whitby, in ‘The Turing Test: AI’s Biggest Blind Alley?’ [134] makes similar arguments. He, like Hayes and Ford, believes that AI need not try to imitate humans. He even uses the same analogy (i.e., AI and artificial flight). Whitby states that the TT has become a distraction and he sees the main source as a mistaken reading of ‘Computing Machinery and Intelligence’ [127]. He is of the opinion that ‘Turing’s paper [has been] interpreted as closer to an operational test than he himself intended’ [134, p. 54] and that ‘the last thing needed by AI *qua* science is an operational definition of intelligence involving some sort of comparison with human beings’ [134, p. 62].

2.5 TT in the Social Sciences

A review of the TT would be incomplete if we were to consider the topic within the boundaries of computer science and philosophy only. Turing’s ideas had many repercussions in social sciences as well. The TT has naturally received attention from sociologists. Much of the more philosophical work on the topic also considers social aspects of intelligence, but there have been researchers who concentrated solely on this dimension. These sociological works are discussed in Section 2.5.1. In addition, the gender issue in the TT has been analyzed and this will be summarized in Section 2.5.2. Finally, Turing-like tests have been used to assess the success of computer simulations of paranoid behavior. This process is described in detail in Section 2.5.3 and will be reconsidered in Section 2.6.

2.5.1 Sociological Aspects

An entity’s status in a society, in general in a social environment, is often considered an integral part of its intelligence. Many psychologists believe that social adaptation, learning and communication are important indications of, even requisites for intelligence. The study of artificial intelligence has also been influenced by this outlook, as is apparent from the recent research being done on intelligent agents. Much attention is focused on learning, adaptivity,

communication³⁰ and socio-psychological factors in intelligent systems [21, 93, 101].

In 1986, Charles Karelis wrote a paper for the *Journal for the Theory of Social Behavior* titled “Reflections on the Turing Test” [74]. This paper summarizes Turing’s original paper [127] and Block’s objections to the TT [8], mildly criticizes the test and briefly discusses some issues surrounding behaviorist approaches to intelligence.

A few years later, in the same journal, we find “A Simple Comment Regarding the Turing Test” [120] by Benny Shanon. The author first mentions the fact that most discussions of the IG are not faithful to the original form proposed by Turing. He then, continues by criticizing the TT for confining human behavior to those that can be conducted by means of the structures and operations that are available to the computer [120]. He raises the important issue of whether cognition is autonomous with respect to social interaction, affect, motivation, motor control and similar systems. However, after stating that the TT presupposes the claim that there is such an autonomy, he abruptly ends his paper by asserting that the only remaining way to distinguish between man and machine is to “look at them, touch them, tickle them, perhaps see whether you fall in love with them” [120, p. 253].

Justin Leiber, in his defense of the TT against Shanon [82], states that Shanon seems to be suffering from the ‘unwillingness to admit the possibility that mankind can have any rivals’ [128], what Turing liked to call the ‘heads-in-the-sand objection’ [127]. Leiber notes that satisfactory answers to such objections have already been given by Turing. He also argues against Shanon’s claim that the TT involves only symbol manipulation and thus assumes a representational/computational framework for cognition. Leiber points out that there is ample evidence in Turing’s paper [127] showing that such a framework is not assumed. He asserts that Turing does not make the aforementioned autonomy presupposition either.

Tracy B. Henley’s [69] is another paper arguing that Shanon is being overly chauvinistic. A reply to Henley was given by Shanon in [121].

³⁰Agent-agent and human-agent.

Some of those who view intelligence as a part of social processes (and vice versa) take a more evolutionary approach [5, 39, 117]. Adaptivity is indeed, a most prevalent characteristic of social intelligence. However, the issue can be viewed from two different levels: the individual level and the collective level. The approaches we have looked at above were mainly individual based. Evolutionary arguments, on the other hand, are largely collective in outlook. These usually focus on the intelligence of *species* and study the factors influencing their development. According to the evolutionary viewpoint, there is a system, i.e., nature, in which entities function and the interactions within the system have effects on individuals that, in the long run, lead to species-level adaptations. The adaptation in this context is not merely giving appropriate responses in appropriate social or physical situations, but is successful survival of the species within the whole system.

In his 1987 paper [5], John Barresi considers intelligent machines as a species and proposes an evolutionary ‘Cyberiad Test’ instead of a Turing Test. According to Barresi, the TT aims to trick a person, but in *natural* intelligence, this person is ‘mother nature’. The Cyberiad Test is similar to the TT: The basis of the judgement is a comparison between humans and the machines. The difference between the two lies in how intelligence is defined. The Cyberiad Test defines intelligent behavior as those that are necessary for the society’s survival. The arbiter here, is mother nature.

According to Barresi, the TT is inferior to the Cyberiad Test because what it can process about an entity’s intelligence is limited to a particular domain, namely, verbal communication. The Cyberiad Test is passed, ‘if [the] society of artificial men are able to continue a socio-cultural evolution of their own without disintegration over an extended period, say of several million years’ [5, p. 23]³¹. Even though this ‘science fiction’ atmosphere sometimes distracts the reader from the important assertions about evolutionary and cultural intelligence, the paper is quite an entertaining piece of work.

³¹Compare this with [117] and Section 2.4.4.

2.5.2 On Gender

Judith Genova draws attention to the gender issue in the IG [46]. She, as I have done so in Section 2.2.1, remarks that Turing's description of the game involves, not a question of species, but one of gender. She states in [45] that her aim was to show that the sexual guessing component of the IG is important, even after the machine enters the picture. My explanation of this design choice differs from that of Genova's, however. As you might recall, I have not made a distinction between the two genders in our explanation. I have regarded the choice of the woman being 'imitated' as a rather insignificant one and assumed that the game would not change radically if it were the other way around. Genova, on the other hand, does not merely accept Turing's choices as accidental, but tries to demonstrate some motivations behind these.

Genova believes that sexist notions about women being less intelligent, by themselves, do not account for the peculiar design of the game. She states that by complicating the game in this manner, Turing questions the existence of discrete categories. In other words, by using the male/female issue, he is attempting to demonstrate that gender itself is a socially imposed concept that is not 'natural' the way we usually think it is.

Genova regards the IG as part of Turing's general philosophy of 'transgressing boundaries' [46]. Under the assumption that Turing admired such transformations that do not conform to the given discrete categories, Genova suggests that Turing might be marking the woman as an inferior thinker because he believes her to be unable to deceive. The rest of the paper considers Turing's hypothetical hope to create a 'perfect being' and draws some analogies between him and Pygmalion. As can be seen, Genova's approach is different from mine; for her, Turing's paper [127] 'is itself a game' [45].

Another paper that considers the gender issue in the IG and constructs links between the design of the game and Turing's opinions on life is Jean Lassegue's 'What kind of Turing Test Did Turing Have in Mind?' [80]. Readers interested in Turing's life and psychology might want to consult it.

2.5.3 Artificial Paranoia

The TT has received some attention from psychologists as well [110, 3, 44]. In this section, however, I focus only on Kenneth Colby and colleagues' work on simulating artificial paranoia [18, 19, 17].

In the 70's, Turing Tests were used to validate computer simulations of paranoid behavior. Colby et al. describe in their 1971 *Artificial Intelligence* paper 'Artificial Paranoia' [18] a computer program that attempts to simulate paranoid behavior in computer-mediated dialogue. The program emits linguistic responses based on internal (affective) states. To create this effect, three measures, FEAR, ANGER and MISTRUST are used. Depending on the flow of the conversation, these measures change their values. Substantial detail about the artificial paranoia program can be found in [18].

A year later, again in the same journal, Colby et al. describe how they validate their simulation program by a Turing-like indistinguishability test [19]. Their approach to such tests is much more practical compared to those of philosophers who argue against the TT because it is a behaviorist criterion for intelligence. Colby et al. believe that computer simulations should be *validated*, and that a simulation's acceptability must first be based on 'its success in achieving the desired end of producing resemblance at some input-output level' [19]. Thus, they view the test as a tool to validate a simulation.

They describe Turing's original IG and note that there is one important point that needs to be resolved before the test can actually be used. What the judges are *told* about the game is not explicitly stated in [127]. From the provided context, it seems as if the judges are not told that one of the parties they are conversing with is a machine. This is assumed both by Colby et al. in their experiments, and by Abelson in [2].

Abelson proposes an "Extended Turing Test". There is a computer program that attempts to imitate a subject carrying out a task, and there is also another target person, referred to as the *foil* that differs from the subject in some respect³². At first, the judge will try to identify the subject and the foil.

³²In Turing's IG, this difference is gender, for instance.

At some point, the computer replaces the foil, but the judge's task remains unchanged. The computer's simulation is acceptable if the difference between the success of the judge in distinguishing the subject from the other is sufficiently small [2].

Colby et al. chose to design their validation in the light of Abelson's comments. They devised a machine-mediated interviewing environment, where experienced psychiatrists corresponded with both real-life paranoid patients and computer programs through teletype. The judges were *not* told that some of the interviewees could be computer programs. They were asked to rank each interviewee's level of paranoia. There were eight judges who interviewed two patients each. In half of the cases, the first interviewee was a human patient and in the other half, it was the computer program. Another thirty-three psychiatrists constituted a second set of judges (referred to as the protocol judges) who were asked to determine levels of paranoia by reading the transcripts of the conversations carried out by the first set of judges. Finally, both sets of judges were asked to identify which of the interviewees was an actual human being and which was a computer program.

Details of the results can be found in [19], but the overall conclusion was that the psychiatric judges did not do better than chance guessing in identifying the computers from the human patients. In a further test, randomly selected psychiatrists were sent protocols of two interviews and were asked to judge which one is a program and which one is a real patient. Only 48% made the correct identification, which also approaches a chance level.

Paranoid people are expected to display irrational and obsessive behavior from time to time. This fact makes the task of the simulation program easy, for deficiencies in linguistic processing of the inputs and generation of the outputs can seem to be indicative of a psychological problem. One interesting point is that, the paranoia simulation program does not use *any* sophisticated linguistic techniques. Input sentences are assumed to be syntactically simple, and the operation of the program relies on spotting some keywords. No parsing or sense disambiguation is done. Even without sophisticated linguistic techniques, the program can attain some success in modeling human behavior. Apparently, this is possible only because the problem area is such that irrelevant responses

from the interviewees are expected.

Still, the simple design of the program indicates that by finding appropriate measures and correlations of these with the flow of the conversation, one can model the behavior of one kind of human being without using much natural language analysis. Because the representational model of the paranoid mind used in the program is a good approximation of the paranoia-related cognitive processes in humans, and because lack of linguistic competence can be accommodated in the setting, the program can be successful. In modeling human beings in general, the former is not so easy to discover and formalize and the latter is not the case any more.

2.6 Chatbots

We have reached the end of the century, but what has *really* been done in terms of passing the TT? Over the years, many natural language systems have been developed with different purposes, including that of carrying out conversations with human users³³. These systems chat with people on the WWW, play MUDs³⁴, give information about specific topics, tell stories, and enter Turing Test competitions. However, none has been able to *pass* the TT so far.

2.6.1 The Loebner Contest

The TT, as Turing actually described it, has never been carried out. However, there are variants of the original in which computer programs participate and show their skills in “human-ness”. Since 1991, Hugh Loebner has been organizing the so-called annual Loebner Prize Competition³⁵. Although views as to whether this annual contest is to be taken seriously varies immensely among the AI community, it nevertheless continues to be the most well-known of the

³³Such systems are usually called language understanding/generation systems, conversation agents, or simply, chatbots.

³⁴Multi-User Dungeons. These are games played interactively on the Internet by multiple players.

³⁵<http://www.loebner.net/Prizef/loebner-prize.html>

attempts to pass the TT. The first program to pass an unrestricted TT will win a gold medal and \$100,000³⁶, while each year, a bronze medal and \$2,000 is awarded to the most “human” program among the contestants. Since 1995, all entries must be prepared to be inquired on any topic whatsoever. No program has won the grand prize yet, but the quality of the participating programs seems to be increasing every year.

The first Loebner Prize Contest was held at Boston’s Computer Museum. Six computer programs, four human subjects and ten human interrogators were involved³⁷. The administrative committee was headed by Daniel Dennett, a prominent figure among the philosophy and cognitive science community. The organizing committee, thinking that it was not possible at the time for a computer program to pass the TT as originally defined, decided that the conversation topics were to be restricted, both for the contestants and confederates. Consequently, the judges were asked to stay on topic during their interrogations. Substantial detail about the 1991 Loebner Prize Contest can be found in [32]. The reader can also consult [91, 107] for more information on other years’ contests.

A widely discussed issue before 1995 was the restricted vs. unrestricted TT. According to Turing, passing a restricted TT would not suffice for intelligence. However, from another viewpoint restricted tests are not totally useless. I am not saying that they should be carried out within the context of the Loebner competition. Still, restricted tests can be devised to assess the success of more specific AI applications that are not created with passing the TT in mind. Examples of such systems that can be assessed by a restricted test are intelligent tutoring systems, computer help services, and natural language components of other applications that are designed for specific domains. The reader can also consult [123] and [86] for more discussion on restricted TTs and the Loebner competition.

In the Loebner contest, the sexual guessing component of the original game is ignored. The aim of the contestants is to convince the judges that they

³⁶Now, Loebner requires that this program should also be able to process audio/visual input.

³⁷In the Loebner Prize terminology, the computer programs are called ‘contestants’, the human subjects ‘confederates’ and the interrogators ‘judges’.

are human. One or more human confederates also participate and try to aid the judges in identifying the humans. The judges also rank the terminals with respect to their “human-ness”. Although, looking at the transcripts, one can see that the computer programs are, in general, obviously distinguishable from the real humans, there have been cases in which some actual humans were ranked less human than some computer programs. In fact, in 1991, not only were some programs thought to be human beings, but an actual human was mistaken for a computer program because of her impeccable knowledge of Shakespeare’s literature³⁸. The interested reader is referred to the article written by Charles Platt, one of the human confederates in the 1994 Loebner Contest [107].

The amount of time that the judges spend communicating with each terminal in the Loebner competition varies. It has been the case that each judge gets more than one chance to interrogate each terminal. Ideally, the contestants should be able to handle conversations of unlimited duration as well as multiple sessions with each judge. In the beginning, each judge was required to rank the subjects from least human to most human. They also had to mark the point at which they believed the subjects switched from computer programs to human beings.

It is difficult to keep track of the small changes in the rules. It is, for instance, not clear how the grand prize of \$100,000 will be awarded. The 1999 rules state that if a contestant achieves a 50:50 chance of being mistaken for a human being, it will be given an intermediate Turing award of \$25,000 and appropriate competitions (those that involve audio and video inputs) will be held after that. However, how this ratio is exactly calculated is not elucidated. I also feel the number of judges is too small to be a basis for a statistic like 50:50.

One of the aims of the Loebner competition, as Loebner states, is to advance the field of artificial intelligence. A majority of researchers believe that this competition does not, and is not likely to serve such a purpose. Some hold that it is possible that a program implemented as an entry to the competition

³⁸The reason why this fact does not count the TT passed is simply because Turing required *consistently* successful performance from machines to grant them intelligence.

can evolve into a useful natural language system. It is also possible that by arising an interest to the topic, the contest may prompt researchers to start working on such systems. Nevertheless, the competition has been getting harsh comments from many leading scientists and philosophers³⁹.

The quality of the programs that compete in the contest has been increasing. The first programs that participated in the Loebner Prize Contests were little more than slightly modified ELIZAs⁴⁰. One need only look at more recent transcripts or play with some of these programs to see that there *is* progress. It seems awards can get people interested in writing TT-passing programs. Hence, the Loebner Prize could indeed be influential on AI research.

I would like to finish this section with a small excerpt from a conversation between the winner of the 1996 Loebner Prize, HeX, and a particularly emotional judge⁴¹:

PROGRAM ... So what do you do for a living?

JUDGE05 I think. I do research. I write. I love. I Live.

PROGRAM I wonder whether you think sometimes! I am feeling like the judge here!

JUDGE05 HUh?

2.6.2 Tricks of the Trade

The first, and probably the most well-known natural language system that has been programmed is ELIZA. Joseph Weizenbaum coded ELIZA at MIT during the years 1964-1966. This “friendly” program simulates a Rogerian

³⁹In fact, Marvin Minsky has offered \$100 to the first person who can get Hugh Loebner to revoke the competition, which he calls an ‘obnoxious and unproductive annual publicity campaign’. Loebner astutely declared Minsky a co-sponsor of the contest, since, according to the rules, when the grand prize is won, the contest will not be held again. In that case, with Minsky’s contribution, the prize should become \$100,100.

⁴⁰See Section 2.6.2.

⁴¹In the following transcript and the others in this section, the spelling and grammar errors have not been corrected. Moreover, timing information denoting the exact response times of the participants has also been deleted.

psychotherapist. It rephrases the interrogator's statements into questions and urges him/her to continue talking. The mechanism behind ELIZA is a very simple one. First, what is typed into the program is parsed. Then, a suitable reply is formulated by simple pattern recognition and substitution of keywords [131]. The term "ELIZA-like" for chatbots is used to mean that the program tries to carry the conversation by using techniques similar to those of ELIZA's.

ELIZA would certainly perform poorly in the Loebner contests or similar instantiations of the TT. This is because the interrogators are trying to find out whether they are conversing with a human or a machine and thus, are not likely to open up about themselves and their personal problems as if they are talking to a psychotherapist. However, it has been reported that some people have developed emotional attachment to ELIZA [132]. Certain psychiatrists went so far as to suggest that such programs can replace psychotherapists all together. Weizenbaum, himself, has been amazed by these delusions that ELIZA, a simple program, could induce in perfectly normal people. These reactions to ELIZA suggest that even if the program has no chance to pass the TT, it can be said to model, with success, the main aspects of the conversational capability of one kind of human being, namely, the Rogerian psychotherapist.

A similar story is that of PARRY, which is a program that attempts to simulate another restricted class of human beings. Kenneth Colby wrote this program in the 70's in order to model the paranoid mind. A modified TT in which an experienced psychiatrist tries to distinguish between a real paranoid patient and the computer program can be carried out to assess how well the paranoid conversational characteristics are simulated. Many expert psychiatrists had difficulty making the distinction between PARRY and actual paranoids. The design of PARRY has been explained in detail in Section 2.5.3.

Both ELIZA and PARRY use certain tricks to be able to successfully perform in conversations. ELIZA directs the conversation away from herself by asking questions. Many people like this⁴² and happily believe that the program is listening as they talk about themselves. ELIZA uses parts of the user's input in the output questions and seems to be following the conversation. In addition to these techniques, PARRY has little stories to tell and tends to insert these

⁴²Although, contest judges most probably would not, as was mentioned before.

in the conversation.

Recent programs also employ many tricks. Joseph Weintraub's PC Therapist, which is the winner of the first three Loebner competitions, and also the one in 1995, is a master of non sequitur. Overall, the program was ELIZA-like but impressed the judges by its long and incomprehensive responses. Its success is understandable in the first three Loebner contests since the conversation topics were limited and the judges actually expected irrelevant, absurd and extravagant answers from a terminal that boasted "whimsical talk". The reader is referred to [32] for transcripts of PC Therapist in the 1991 contest, which are illustrative of the tricks it employs.

PC Therapist's winning the first unrestricted Loebner competition shows that the usage of non sequitur is indeed, a powerful trick. I find the utilization of non sequitur similar to that of simulating typing mistakes of humans: In the long run, such tricks will not suffice by themselves, but they can be incorporated into otherwise successful programs as seen necessary. The transcripts of the 1995 contest indicate that little progress has been made since ELIZA. Below are some excerpts that may give an idea about the performance of Weintraub's program in 1995. You may realize that if the judge asks sensible, down to earth questions, the program's responses seem erratic and irritating⁴³:

JUDGE06 What did you have for dinner yesterday? Keep in mind I'm a strict vegetarian.

PROGRAM I could swear you just said what did you do yesterday at 7:83 p.m.?? He was so narrow-minded he could see through a key-hole with two eyes. Are such questions on your mind?

JUDGE06 What did you have for dinner yesterday?

PROGRAM What does a woman want anyway? What answer would please you most?

However, depending on the attitude of the interrogator there can exist an atmosphere (or rather illusion) of mutual understanding.

⁴³Especially when it repeats long chunks of the judge's previous inputs in totally unrelated contexts.

PROGRAM Is love the most important thing to you? How important is that to you?

JUDGE05 essential cure to my existential schizophrenia

PROGRAM A striking example of the essential identity of the two tongues. existential schizophrenia, that's amusing.

PROGRAM You know the old saying - fun is in the eye of the beholder!. What do you think?

JUDGE05 the ''beholder'' is the central issue. confusion reigns because of the illusion of ''identity.''

PROGRAM I love men, not because they are men, but because they are not women. Is that the real reason?

JUDGE05 i love all that is, which has produced me to observe it, yet given me the madness to think i am distinct ans apart from it.

PROGRAM I have this fantasy of making love to two women in one year. Can you elaborate on that?

JUDGE05 why confine love to only two

PROGRAM Is love the most important thing to you? What comes to your mind when you ask that?

Another trick that has been used is to direct the conversation by talking a lot about a predefined topic. This has been used and abused, and the difference shows. Programs that cheat by typing very long responses extremely slowly give away the fact that they are trying to use up the interrogation time. However, the winner of the 1997 Loebner competition, Converse, uses the conversation-directing technique rather impressively. It begins the session by asking the interrogator about something interesting that had been on the news the previous night. Below are excerpts from the 1997 contest transcripts of Converse:

PROGRAM Well, it's nice to meet you, Elliot. Did you see

that story on CNN last night about the lesbian couple who came out at a White House party on Sunday?

JUDGE04 Yes, I did. I think it may be a publicity stunt for Ellen.

PROGRAM Ellen Degeneres was one of them - she was kissing her lover in front of Bill Clinton. Do you think he was very embarrassed?

JUDGE04 I think he want the gay vote. He will do anything to get it and keep it. He does not have a conservative base.

PROGRAM Are you still a friend of Bill Clinton?

JUDGE04 I am not sure if he has any real friends. He has achieved alot and left alot of people behind.

PROGRAM You really don't like the guy--well, at least that's definite--I thought most people were indifferent or numb by now.

--

PROGRAM Well, it's nice to meet you Tracy. Did you see that story on CNN last night about the lesbian couple who came out at a White House party on Sunday?

JUDGE03 No... Are you serious? Who?

PROGRAM I'm pretty much that way today. No idea who, do you? Hm, i was just saying Ellen Degeneres was one of them - she was kissing her lover in front of Bill Clinton. Do you think he was very embarrassed?

JUDGE03 I doubt it very much. From what I've heard, Bill's been, done and seen just about everything there is for a consenting adult to see.

PROGRAM Are you still a friend of Bill Clinton?

JUDGE03 Never was.

PROGRAM You really don't like the guy--well, at least that's definite--I thought most people were indifferent or numb by now.

It is possible to see, after reading the full transcripts, that the program repeats the exact same response sequence to each judge regardless of what they type in. But the judges could not know this, and the overall effect was quite impressive. The program spent a lot of the time in this manner, and although it could not handle many kinds of questions, won the competition. The program that came second, Jason Hutchens' SEPO, was more competent in handling diverse questions.⁴⁴ In any case, directing the conversation seems to be more sophisticated than repeating the input back to the interrogator or asking ELIZA-like questions. The situation is similar to watching a team play in the offensive rather than in the defensive.

As was mentioned before, an approach is that of modeling one kind of person. Thom Whalen took this a step further than what has been done with ELIZA and PARRY and created an actual human being, Joe the Janitor. Although he lost the Loebner contest to Weintraub in 1995, the program was quite successful.

Whalen utilized further tricks while inventing Joe. For instance, Joe has a narrow world view since he works at nights, does not read books or newspapers and does not watch TV. In addition, he has a little story to tell about being accused of theft and fear of losing his job. These tricks are similar to those employed in the design of PARRY. The major weakness of Joe against other programs was that he did not resort to non sequitur and repeated the same responses when confronted with something he didn't *know* about. Below, we produce some excerpts from the transcripts to depict what can go right or wrong with Whalen's approach:

JUDGE04 where do you work?

PROGRAM At the university. I mostly clean the Standish Building.

JUDGE04 What university?

PROGRAM University of Eastern Ontario. I'm a cleaner. I

⁴⁴An interesting point was that one of the judges was named 'Nate', short for 'Nathan'. The program repeatedly addressed him as 'Mate', and complained about him not telling his name. This created the impression that SEPO lacked intelligence, but was, in fact, due to the fact that Jason Hutchens is from Australia. This suggests that the designer's mind is an integral component of the resulting programs, which is an interesting idea to ponder upon.

empty trash. Vacuum. Dust. You know.

JUDGE04 and why do they want to fire you?

PROGRAM It's just because my boss, Mr. Johnson, is mad at me.

JUDGE04 why is he mad at you?

PROGRAM I didn't do nothing wrong. But he thinks I did. It's all because of that radio they think I stole from the lab.

JUDGE04 that's too bad. are you in a union?

PROGRAM What would I tell the union? They won't say they're firing me because I stole the radio. They'll just make up some excuse...

--

JUDGE02 what does nato stand for?

PROGRAM I don't know.

JUDGE02 what's your name?

PROGRAM What's your name?? I don't know. What do I look like? Some kind of professor?

JUDGE02 What was mark twain's real first name?

PROGRAM Huh?

JUDGE02 what does huh mean?

PROGRAM Beats me.

JUDGE02 huh?

PROGRAM What do you mean?

JUDGE02 beats me.

PROGRAM What?

JUDGE02 huh.

PROGRAM I don't know.

Another strategy of Thom Whalen, as he reports in [133] was to program suitable responses to common questions on topics that people usually talk about when they first meet, like their names⁴⁵, where they live, the weather, etc. It is obviously a good idea to have these responses prepared in case the interrogator introduces such topics. In fact, some judges at some Loebner competitions make only small talk on such topics. However, generally the judges do not treat the competitors as if they were people they've recently met. They ask questions that nobody would ask to people they have just met⁴⁶.

Apparently, the judges of the 1995 Loebner competition were not very interested in Joe's little story and although they admitted that it was the terminal with the most human personality (which should be a quality that the competitors are trying to achieve, since having a human personality should be an integral part of human-ness) they ranked Weintraub's program above it. For more detail about Joe and its performance in the 1995 Loebner Prize Contest, see [133].

2.6.3 What Else Should be Done?

In chatbot design, first and foremost, a good grammar unit is necessary. Most programs are rather poor in this task and their tricks do not go much beyond pronoun transposition. It seems a good idea to employ more sophisticated natural language processing methods. One may argue that usage of perfect grammar is not crucial since it is quite rare that humans do that in informal transactions. If a program's responses are grammatically perfect, some interrogators may decide that no human can use English so impeccably⁴⁷. However,

⁴⁵Although, as seen above, Joe cannot answer the question "What is your name?"

⁴⁶One of the judges in the 1997 Loebner competition tried asking each terminal the question "When you got your first licence, was it in a stick or an automatic?". The question is a cleverly planned one since words like 'driving' or 'car' are not used, but the meaning is clear from the context. Even the misspelling of the word 'licence' as 'liscence' is most probably intentional. It is difficult to imagine a computer program that would answer various trick questions of that type, but almost anyone (certainly any adult American) would be able to give a relevant answer to that one.

⁴⁷One might recall that Eliza Doolittle was mistaken for a Hungarian princess because she spoke English too well for a native.

most programs err in ways that give away their machine-ness; when interrogators feel they are talking to a machine, they literally *attack* it in order to fully reveal its identity. A good strategy for the TT is undisputably that of trying to maintain human-ness (or at least the neutrality) for as long as possible. It gets very difficult for the machine to make the interrogator believe that it is human after he/she has his/her mind set on “unmasking” the poor thing.

A promising approach is learning programs. The reader might recall that Turing discussed these extensively⁴⁸. Although such programs that have been developed so far do not seem very sophisticated, the approach is logical and is likely to yield good results in the long run. Some learning chatbots boast the capacity to converse in any given language. However, there seems to be a tradeoff between the sophistication and the number of languages any one system can learn. In designing natural language learning systems, knowledge from psychology and cognitive science can be employed in order to model human language acquisition. In fact, work has been done in this line, but not with the intention of producing computer programs to pass the TT. Another option is using mathematical and statistical techniques to represent word sequences and probabilities of them occurring in proximity.

I believe many of the chatbots in the future will be using learning methods. Already, those programs that do not keep track of the current conversation (relying solely on text processing tricks) perform poorly compared to those that learn from the interrogators. As the quality of the conversational systems increase, to be competent, developers will have to integrate a learning component in their programs *and* teach them in ways that maximize their performance.

Overall, when one looks at the transcripts from the Loebner Prize Contests and talk to some chatbots, one realizes that successful programs integrate the techniques mentioned above. They have a personality and history, they try to ask questions and initiate new conversations, they produce grammatically correct responses, they have some information about recent happenings (like new movies, albums, gossip), they learn about and from the interrogators and when they don't know what to say⁴⁹, they try to respond by combining words

⁴⁸See Section 2.2.3.

⁴⁹When all else fails, so to speak.

from the interrogator's input in order to come up with a relevant answer.

I will propose more that needs to be done in developing successful TT-programs in Chapter 3.

2.7 Discussions and Conclusion

Having analyzed the '50 years of the Turing Test', I am now going to conclude my survey with a brief look at some main issues about the TT and of course, its future.

My stand on the issues are not at the extremes. Perhaps this is because I have tried to be objective in my analyses of the arguments for or against the TT; this is expected considering the nature of my task. Most of the arguments discussed in this chapter are strong, and if read independently, can "convince" the reader. However, looking at the 50 years as a whole, I find it difficult to adopt a simple viewpoint. I believe some readers, having read the current work, will also be in the same position.

I now discuss some important issues regarding the TT and provide my own answers to (and interpretations of) those.

- *Why did Turing propose such a strange game?*

I discussed this question at length in Section 2.2.1. Among the comments made on the issue (for instance [46, 80, 2]) I find the best explanation to be the one I provided. In the IG, the machine is supposed to be as good as a man who is imitating a woman. This gender-based design might be a methodological choice. We are asking the machine to imitate something which it isn't; so it is only fair that we compare its success against a human who is *also* imitating something which it isn't.

- *Is the TT an operational definition?*

Parts of Turing's paper (the percentages, the predictions about the future, etc.) would prompt us to believe that he intended it as such. However, most

arguments surrounding the issue have been philosophical. Neither Searle's Chinese room, nor Block's Aunt Bubbles machine are practically realizable, yet they have been proposed with the intention of refuting the TT as a measure of intelligence. Apparently, proponents of such thought experiments and some other commentators view the TT as a philosophical criterion.

Viewed as a practical test, I see the TT as follows: If a machine passes the TT it could be granted intelligence. However, if it cannot, we cannot say for sure whether it thinks. I believe this is the most common stance towards the TT.

Philosophically, the test has been subject to many criticisms. We are all familiar with the anti-behaviorist attacks. Some have also noted that the TT is anthropomorphic. It is true that the TT tests for human intelligence. We should not be too bothered about this for it is only natural that we are interested in the only kind of intelligence we know⁵⁰.

Moreover, we need not assert that the *only* way to grant intelligence to machines is by the TT. Perhaps a good way to see the TT is as a means of gathering inductive evidence about machine mentality⁵¹.

As was mentioned before, lately most arguments on the TT has been of the "put up or shut up" sort (e.g., [56, 84]). With the advances in computer technology, cognitive science and artificial intelligence, it is time that we stipulate that attackers or defenders of the TT back up their arguments with something more than mere intuition. This does not mean that everyone should try to develop TT-passing computer programs. However, to argue for or against the TT, I believe that a more or less realizable method of passing the test should be supplied.

- *Isn't the TT guilty of behaviorism?*

I am not saying there *should* be tests to assess machine intelligence, but if I have to make a choice, TT-like tests seem the best method for reasoning

⁵⁰Moreover, it is not even evident that other "kinds" of intelligence can be conceived by human beings. The interested reader may refer to [100] for a good discussion on this issue.

⁵¹See Section 2.3.4.

about machines' minds even though they are being accused of behaviorism. If one day, we stop granting intelligence to other human beings behavioristically, then the TT could be replaced by some other method.

The idea of a TT-passing machine having radically different information processing compared to humans is neither scary, nor improbable. If this happens one day, it will just have to be 'heads-out-of-the-sand'.

- *Isn't the TT too easy?*

The TT has been criticized for being a limited test since it enables the assessment of only "verbal" intelligence. However, it does not follow from this that the test is too easy.

Proponents of this view should come up with a realizable model of a machine that passes the TT and then prove that this model does not deserve to be called intelligent. If a simple "bag of tricks" passes the TT, I am willing to either admit that the TT is too easy or that the human mind is a simple bag of tricks as well.

After 50 years, all that we have are some very rudimentary chatbots (Section 2.6), serendipitous FSA's [12] and unrealizable Chinese rooms [118] and Aunt Bubbles machines [8, 9].

- *Isn't the TT too difficult?*

It is challenging. This is primarily due to our limited understanding of natural intelligence, more precisely language understanding, generation and processing in humans. It may even turn out that these processes are impossible to model on computers.

As is manifested by the space we devoted to Robert French's paper [40], I find the idea that the TT is (too) difficult, an important one. Is this a deficiency of the TT? Not if one does not require success in the TT as a necessary condition of machine intelligence. Computers, even today, perform many tasks that would require intelligence if done by humans. Research and development in this line is valuable and worthwhile. A natural language system that answers

queries on a particular topic is certainly a remarkable product. It's not useless just because it cannot pass the TT. In our opinion, the TT is a sufficient condition for human-like intelligence (or more appropriately, mentality) because of the reasons outlined above. It may be too difficult to pass the TT, but this does not prevent AI from building intelligent machines.

- *Why bother about the TT?*

As we saw, there are those who believe that the TT is harmful for AI [68, 134]. If AI's aim is to make computers perform "intelligent" tasks and thereby make life easier for humans, I grant it that TT-passing programs are not very useful from that perspective.

AI researchers are being unjustly accused of mankind's failure in making machines that can pass the TT. This, I believe, is precisely the reason behind some of the harsh reactions to the TT from the AI community. Even if we take an extreme viewpoint and stipulate that AI's ultimate goal is to produce TT-passing machines, we should accept that this is a hard problem and give it more time. If less AI researchers shun the TT because "it gives the field a bad name", maybe more can be done in the positive direction.

Recall the "myth" of Newton and the apple. Recall Archimedes and his adventures in bathing. The apple might be silly, but gravity is not. Of course, thousands of people bathe, thousands of apples fall. The point is, sometimes a scientist can focus on an apple and behind it, find gravity. Later, you may forget about the apple, or even eat it if you like.

The TT may seem like a game. But trying to make computers communicate with humans in natural language is a task that may also provide valuable insights into how the human mind works. Now this latter is unarguably of scientific and philosophical interest.

- *So what happens now?*

We failed to fulfil Turing's prophecy in the first 50 years of the TT. We should admit that we have a difficult task at hand.

Hopefully, the reader has seen that many critics of the TT have expected too much, too early. Seeing the TT as the ultimate goal of AI will make many remarkable achievements look weak. The situation is somewhat reminiscent of “Fermat’s last theorem” from mathematics which was proved recently by Andrew Wiles, after centuries of failure. Practically nobody believes that Fermat had proved the theorem at the time he scribbled something about lack of space in the margin of his book more than 300 years ago. In fact, Wiles’ proof alludes to mathematical theory that was not developed until long after Fermat died. The same might be true of the TT. Maybe we simply don’t have the requisite theory at this time⁵².

The TT is after all, about simulating human use of language by computers. This, in turn, involves many questions: How do humans use language in similar settings? What is the relation between language and cognition? Is language autonomous with respect to other cognitive abilities? How can computers be made to *understand* language? What does a “simulation” mean, anyway? You can think of more questions like these. These are all *big* questions that psychologists, computer scientists, philosophers, linguists have been probing for several years. As more light is shed on each question, we will be one step closer to passing the TT. I don’t know how many such steps will be necessary. Perhaps it is best to relax and not regard the TT as a “goal” but as a feat that will (if at all) be achieved through a synthesis of other remarkable feats. Everyone who considers himself/herself a “cognitive scientist” may, explicitly or implicitly, be working towards passing the TT. In any case, I believe he/she would at least be interested in what is going on in the TT arena.

Having given a detailed and interdisciplinary review of the 50 years of the TT, I will now reconsider the TT as a special kind of conversation and introduce other ways of looking at machine imitation of human linguistic behavior. Some arguments in the present chapter will be taken up later on, but the focus will generally be on the linguistic (rather, pragmatic) issues concerning the TT.

⁵²Of course, passing the TT may be “impossible”, but none of the counter-arguments proposed so far suffice to establish such a bold claim.

Chapter 3

A Pragmatic Look At the Turing Test

Much of the work in computational linguistics has been concerned with syntax, less of it with semantics. In this thesis, I concentrate on the pragmatics of natural language processing. I am aware that without modeling certain aspects of syntax and semantics we cannot develop conversational programs. I am by no means claiming that these studies are of lesser importance. Furthermore, I do not mean that syntax, semantics and pragmatics are independent modules that are to be handled separately. My aim is to show that the story does not end at “what is said”, that even if we develop natural language systems that handle syntax and semantics, we will still need to handle at least some pragmatic phenomena.

Pragmatics, in a nutshell, is concerned with “language in use”. The TT stipulates a criterion on machine intelligence based on the way computers use language. What could be more natural than the juxtaposition of these two concepts in analyzing human-computer communication in natural language? I believe a pragmatic look at the TT reveals a lot of important issues that are easy to miss otherwise. Through a pragmatic analysis, we can gain valuable insights on what it *means* to have a human-like conversation and what principles, implicitly or explicitly, guide human-computer conversation. I believe that these will have direct consequences on the way we think about the TT,

and will hopefully make the situation clearer, although, unfortunately, not any easier.

This chapter is organized as follows: Section 3.1 constitutes an introduction to pragmatics and conversation: In Section 3.1.1, pragmatics is defined and explained with examples. Our focus, Grice's cooperative principle and conversational maxims are similarly studied in Section 3.1.2 and the related issue of implicature is discussed in Section 3.1.3. Further characteristics of implicature, as well as related works are briefly discussed in Section 3.1.4. The subject of Section 3.2 is my empirical study on Grice's conversational maxims and the TT. I first explain some methodological choices I made in Section 3.2.1. Then, the aims and the design of the study are described in Sections 3.2.2 and 3.2.3, respectively. Section 3.2.4 features an analysis of some of the conversations used in this study. The results are presented and discussed in Sections 3.2.5 and 3.2.6, respectively. The section ends with the analysis of the effects of bias on the results, which is given in Section 3.2.7. The results provided constitute a basis for the analysis of human-computer conversation given in Section 3.3. Here, I first consider intentional behavior and cooperation in Section 3.3.1. Then, in Section 3.3.2, I list some practical concerns in human-computer communication. Section 3.3.3 introduces some notational conventions and includes a summary of the different TT-like settings. The importance of bias is further emphasized in Section 3.3.4. Section 3.3.5 builds upon the idea that the humans' prejudices in TTs can have an effect on the results. Finally, Section 3.3.6 reconsiders the cooperative principle within the context of the TT.

3.1 Pragmatics and Conversation

This section aims to provide the basic concepts from pragmatics, more generally from linguistics and philosophy, that are used throughout the rest of this thesis.

As our task is one of exploring the pragmatic phenomena as they apply to TT situations, we will be concerned with *conversation*. Although the TT has not been studied as conversation *per se*, I believe that we can gain valuable

insights into the topic if we view it as an instance of human computer conversation with some specific constraints imposed on the aims and identities of the participants. This is studied in detail in Section 3.3.

In analyzing human-computer conversation, I use the existing frameworks and guidelines in pragmatics. It is therefore necessary to look at what governs (human-human) conversation first. In this section, I provide a brief introduction to pragmatics and a more detailed summary of some issues in pragmatics that pertain to conversation.

3.1.1 Pragmatics and Why We Care About It

The term pragmatics was first used by Morris. ‘Pragmatics is designated the science of the relation of signs to their interpreters’ [104, p.43]. Very concisely, linguistic pragmatics can be defined as ‘the science of language seen in relation to its users’ [94, p.5]. Very informally, pragmatics¹ has served as a bin in which phenomena that could not be fully explained by other linguistic theories were placed.

Various definitions of pragmatics have been given but I believe the concept itself, like many others it embodies or is associated with, is best explained intuitively and through examples. After all pragmatics is all about language in use, language in action.

Example 1 .

1. *Can you hand me that pencil?*
2. *Two of my five children are in elementary school*².
3. *I will marry you.*
4. *Either you give me the money, or I will shoot you.*
5. *Kafayn yedim.* (Literal translation: “I ate my head”)

¹Sometimes called the “wastebasket of linguistics”

²Example from [14]

Let us consider all sentences in Example 1 one by one:

(1) is technically a (yes/no) question. But in fact, it would be rather irritating if the hearer responded with a “yes” and did not give me the pencil. I do not really want to know whether the hearer is *capable* of clutching the pencil I am referring to, lifting it up and giving it to me. Although, syntactically, I am just asking a question, I am actually requesting something. I am performing what is called an *indirect speech act*.

In (2), we see an example of *presupposition*. This statement presupposes that I have children, in fact, that I have five of them. Moreover, a hearer is more than likely to infer from this statement that three of my five children are *not* in elementary school ³.

(3) seems like a very simple sentence. A deeper look at it reveals more, though. Consider the situation in which I utter this sentence just after being proposed to by a man. He is naturally going to assume that I am promising to marry him. I have not said “I hereby promise to marry you” but (3) is equivalent to that within the context of a marriage proposal. Once again, I am performing a speech act, this time one of *promising*.

Consider (4). Here, I am demanding something and providing an “alternative” via an either/or statement. But in reality, the hearer is free to do whatever he/she wants. However, I am performing a threatening speech act, which means that I must be in a position to limit the hearer’s freedom in this matter (i.e. giving the money). From the sentence, we can infer that I am most likely pointing a gun at the hearer.

(5) is from Turkish. The literal translation is given in parentheses. Just as the English translation is non-sensical, the (literal) meaning of the sentence is absurd in Turkish. Apparently, I have not eaten my head. However, anyone with a good understanding of colloquial Turkish of the late 20th century will know what I am trying to say is that I am about to go crazy.

As can be seen from these examples, there are some linguistic phenomena

³However, note that even if all my children are in elementary school, the statement is still true.

that cannot be explained by the existing theories and frameworks in syntax or semantics. Also notice that they can be present even in the simplest situations in real life. As was mentioned above, these phenomena lie in the domain of pragmatics.

The examples above illustrate that sometimes, by saying things, we can convey much more than what we actually say. We can promise or request things as in (1) and (3) and threaten others as in (4), and in general, perform speech acts. It is possible to provide information without explicitly stating facts, as we saw in (2). The interpretation of statements depend not only on the logical forms of the linguistic units involved but also on the context of the utterance, as is manifested in all sentences in Example 1, especially (3) and (4). It may also depend on the hearer's knowledge of the world, a particular language, culture, slang, idioms, metaphors, aphorisms, jargon or figures of speech as (5) demonstrates.

There is much more to pragmatics than space permits me to explain here. There are other issues pertaining to plans, acts, interpretations, beliefs, intentions, assumptions and the *reflexivity*⁴ of the above. As was stated in the beginning of this section, pragmatics studies the language in relation to its users. Therefore what the speaker and the hearer know about the situation, their beliefs, intentions and assumptions about the situation and each other are crucial to the analysis of their communication. Just to get a feel, let us return to Example 1. If I am asking you to hand me that pencil, I probably believe that you know which pencil I am referring to, that you are capable of performing the task, that you realize I am talking to you, that you speak English, that you are a rational and conscious agent, etc.

For further information on pragmatics at an introductory level, the reader is referred to [47, 81, 85, 94]. Essays on more specific topics within pragmatics can be found in [31, 30].

⁴Or the inter-personal or mutual realization.

3.1.2 The Cooperative Principle and the Conversational Maxims

Pragmatics is a difficult field of study and it is challenging to come up with well-defined theories or rules to explain phenomena that fall under its domain⁵. We find that, in pragmatics, there is an abundance of *principles* and *maxims*. In fact it has been said that ‘one uses rules in syntax, but principles in pragmatics’ [81]. The difference is not only at the level of terminology: Principles and maxims, unlike rules, are not absolute or predictive. Speakers are not required to abide by them, the hearers are not guaranteed to interpret utterances according to them.

In this section, we are going to focus on an aspect of pragmatics that is more relevant to conversation. Clearly, a conversation involves more than one entity. For there to be some communication, there must be at least two entities who have knowledge of the same language and the means to carry out a conversation (i.e., they should be physically close enough to each other or should be using some other device such as a telephone or a computer). But there are also some principles and maxims that characterize *meaningful* conversations. We study these in this section.

Philosopher Paul Grice first introduced the *cooperative principle* (CP from here on) during the William James Lectures of Harvard University, in 1967 [48]:

CP Make your contribution such as is required, at the stage at which it is required, by the accepted purpose of the talk exchange in which you are engaged. [49, p.47]

The CP consists of four sub-principles, usually referred to as the *conversational maxims*. These are called the maxims of *quality*, *quantity*, *relevance* and *manner* (Henceforth, QL, QN, RL and MN, respectively) [49, pp. 47-48]:

⁵In fact, some would argue that if you could explain these in such a manner, they would not be part of pragmatics any more. Others believe that developing theories to account for these phenomena is the job of pragmatics.

QN Supermaxim: Do not make your contribution less or more informative than is required.

1. Make your contribution as informative as is required (for the current purposes of the exchange).
2. Do not make your contribution more informative than is required.

QL Supermaxim: Try to make your contribution one that is true.

1. Do not say what you believe to be false.
2. Do not say that for which you lack adequate evidence for.

RL Be relevant.

MN Supermaxim: Be perspicuous.

1. Avoid obscurity of expression.
2. Avoid ambiguity
3. Be brief (avoid unnecessary proximity)
4. Be orderly. ⁶

I mentioned the difference between rules and principles in the beginning of this section. Grice's CP and the maxims are not exceptions. Speakers need not (and sometimes do not) follow these, as we will see in Section 3.1.3. Grice views talking 'as a special case or variety of purposive, indeed rational, behavior' [49, p.48]. This does not imply that maxim violators are always irrational, but it should be apparent that without *any* adherence to the conversational maxims, there would not be much of a communication between the conversants. I agree with what Grice has to say about this: 'A dull but, no doubt at a certain level, adequate answer is that it is just a well-recognized empirical fact that people⁷ do behave in these ways; they learned to do so in childhood and have not lost the habit of doing so; and, indeed, it would involve a good deal of effort to make a radical departure from the habit' [49, p.49].

⁶And one might need others [49].

⁷Here, I believe he should have said "normal people". For people with serious psychiatric disorders or brain damage, a lack of adherence to the CP or the maxims is not out of the ordinary.

3.1.3 Implicature

The word *implicature*, coined by Grice [49], is derived from the verb “to imply”. An implicature is something which is not explicitly “said” in language, but is implied in the conversation. According to Grice, what is conveyed by an utterance can be studied in two parts. What is *said* constitutes the logical content of the sentence. What is conveyed other than what is explicitly stated is the implicature.

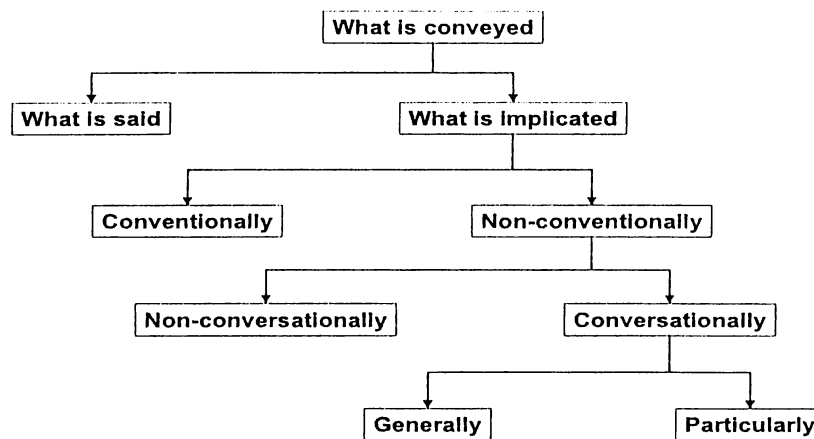


Figure 3.1: Classification of what is conveyed in conversation

Grice makes some distinctions between implicatures as well (See Figure 3.1). Among these we will be concerned mainly with *conversational implicatures*, but a brief look at the others are also necessary.

Conventional implicatures make use of the conventional meanings of the words in the utterances.

Example 2 .

1. *Although she is a blonde, she is quite smart.*
2. *He is an Englishman; he is, therefore, brave*⁸.

⁸Example from [49]

3. *I will buy you lunch if you help me with this homework.*

Consider the sentences in Example 2. From (1), we may infer that the speaker believes blondes are usually inferior in intelligence. Similarly, the person who utters (2) implies that being brave is a consequence of being an Englishman. The conventional usage of the words ‘although’ and ‘therefore’ raise these implicatures.

Now let us consider (3). This sentence is illustrative of the difference between logic and real life. The utterance is syntactically in the form of a logical implication, $p \Rightarrow q$. If you help me and I don’t buy you lunch, you might be disappointed or even mad at me because I made you infer that I would buy you lunch and my if/then statement constitutes a speech act. And logically, even if you do not help me with the homework, I could still buy you lunch. That’s logical since $\neg p \Rightarrow \neg q$ need not be true. But my utterance, in the conventional sense, implies that you are *not* getting a free lunch, unless you help me with that homework.

Since conventional implicatures depend on certain linguistic cues (conventional use of certain words), they can usually be handled in a way that is very similar to the analysis of semantic content.

Conversely, *conversational implicatures*, which are our main concern in this study, may be rather loosely related to the linguistic content of the sentences. According to Grice, they derive from the constraints the CP and the conversational maxims put on the conversants. Once again, how conversational implicatures “happen” become clearer when we see more examples.

But first, let us see how participants in a conversation may fail to follow the conversational maxims [49]. The speaker,

- may quietly and unostentatiously *violate* a maxim; in this case he/she is liable to mislead,
- may *opt out* from the operation of both the maxim and the CP; e.g., may make it known that he/she is unwilling to cooperate in the way the maxim requires,

- may be faced by a *clash*; e.g., may be unable to fulfill QN without violating QL⁹,
- may *flout*, i.e., blatantly fail to fulfill a maxim; on the assumption that he/she is not opting out, is not faced by a clash, and is not trying to mislead; the hearer will have to *exploit* the maxims to understand what the speaker is trying to convey.

And the relationship between conversational implicature and the CP is explained by Grice as follows:

To work out that a particular conversational implicature is present, the hearer will reply on the following data: (1) the conventional meaning of the words used, together with the identity of any references that may be involved; (2) the CP and its maxims; (3) the context, linguistic or otherwise, of the utterance; (4) other items of background knowledge; and (5) the fact (or supposed fact) that all relevant items falling under the previous headings are available to both participants and both participants know or assume this to be the case. A general pattern for the working out of a conversational implicature might be given as follows: “He has said that p ; there is no reason to suppose that he is not observing the maxims, or at least the CP; he could not be doing this unless he thought that q ; he knows (and knows that I know that he knows) that I can see that the supposition that he thinks that q is required; he has done nothing to stop me from thinking that q ; he intends me to think, or is at least willing to allow me to think that q ; and so he has implicated that q [49].

We now consider some examples of conversational implicature.

Example 3 .

⁹Suppose that upon being asked the question “Where does John live?” I answer, “Somewhere near campus”. I have provided less information than is required and violated QN. But if I do not know exactly where John lives, by providing more information, I will have to say things for which I do not have adequate evidence, and thereby violate QL.

1. A: *I am hungry.*
B: *It's almost noon.*
2. A: *Who's this song by?*
B: *Some British band.*

Suppose that the conversation in (1) takes place among co-workers at an office. In this context, B's utterance does not contain any linguistic cues that may yield to a conventional implicature. It is a well known fact that noon is the time for lunch. Since A and B are co-workers, B rightfully assumes that A knows this and therefore, we cannot say that his contribution is irrelevant. This is an example of implicature because the syntactic or semantic analysis of B's utterance would contain nothing about lunch. From that viewpoint, he/she merely makes a statement about what time it is.

In (2) there is a clash. B does not know the name of the band that plays the song A is referring to. His contribution is vague and not informative. But if he said more, he would have to violate the maxim of quality.

A more interesting case is when the maxims are flouted. Example 4 shows how implicatures can work when the supermaxim QN is flouted.

Example 4 .

1. A: *Do you like my hair?*
B: *It's short.*
2. A: *What do you think of this painting?*
B: *I think it's perfect. You've done a great job. And I don't think you should worry about what Adam said about it. What does he know about art?*
3. A: *How much do you make in your new job?*
B: *Enough.*

In (1) B's comment is redundant since A probably knows that his/her hair is short¹⁰. B cannot be unable to say more about A's hair since he/she has evidently seen it (in fact is probably looking at it at the time of the utterance). Also B must know that more information is required since A asks about his/her *opinion*. It may be inferred that B does not *want* to provide more information. A likely interpretation of this would be that B did not like the hairdo and does not want to hurt A's feelings by making an explicitly negative statement. There could be other explanations. Perhaps B is a fan of short hair and A knows this. By simply stating that A's hair is short, B might be implying that he/she likes it. (1) is a situation in which QN1 is flouted by B in a way that gives rise to a conversational implicature.

Now in (2), QN2 is violated by B since he/she provides more information than is required. A merely asks B's opinion about the painting in question and from B's reply we understand that the painting has been done by A ("You've done a great job.") There is no reason for B to mention Adam and his (apparently) negative remarks. By providing too much information B seems too anxious to please A. Although all of his/her comments on A's painting are positive, he/she loses credibility since there really is no need for him to "try so hard". Of course, here, we assume that A and B were not discussing Adam's remarks prior to the exchange above. In that context (i.e. in the situation that they have discussed Adam's opinion and A actually asks "What do *you* think about this painting?"), B's contribution may not be considered over-informative.

In (3) B violates QN1. B probably believes that it is none of A's business how much he/she makes, or that A should not have asked about this in public.

Example 5 .

1. *A: Do you like my hair?*

B: It's short.

2. *His words were salt on my wounds.*

¹⁰Here, we are assuming that B is not flouting QL and that A's hair is short. Compare this with Example 5.

3. *Roseanne is a little overweight.*

The sentences in Example 5 are instances in which the supermaxim QL is violated. Suppose A has told B that she is going to get her hair cut short. She goes to the hairdresser and comes back without having it cut but, say, colored. In this context, B's response in (1) is logically incorrect. However, we may implicate that B is trying to convey that he/she was expecting A to have had her hair cut short.

Some well-known linguistic phenomena (such as irony, metaphor, meiosis, hyperbole¹¹) occur when speakers flout QL. (2) is an example of *metaphor*. The sentence is not true; in fact it is nonsensical since it involves a categorial falsity¹². If (3) is said referring to actress Roseanne Barr, this is an example of *meiosis*. Roseanne Barr is not a little overweight, she is noticeably fat.

Example 6 .

1. *A: I think vegetarians are so stupid.*

B: Has anyone seen the new Star Wars?

2. *A: W3C has recently introduced new standards for CSS but Netscape and Internet Explorer choose to implement them differently.*

B: Has anyone seen the new Star Wars?

For examples of RL being flouted, let us look at Example 6. Both in (1) and in (2), B's contribution is irrelevant to A's statement. In both conversations, B seems to want to change the subject. In (1), it is probably because A's strongly negative and rather rude remark about vegetarians is likely to offend someone in the group the conversation is taking place. Or perhaps it has already offended B, and he/she does not wish to discuss the issue with A. In (2), the reason for B's anxiousness to change the subject may be indicative of a number of things: B may have no knowledge of the topic (WWW design

¹¹The reader is referred to any introductory book on linguistics, for instance [43, Chapter 5] for more on these.

¹²Note that metaphors are not only the result of QL being violated. The statement "Every rose has its thorn" is correct but in most contexts this utterance would violate QN and RL.

standards) or if they have been discussing the topic for a while, he/she may simply be bored.

Example 7 .

1. *A: I'm sorry, can we talk later? I need to finish this homework. It's due in an hour.*

B: I am very sorry to have caused a disturbance. I am hoping that your highness will be kind enough to forgive me.

2. *A: You look very nice today.*

B: Go away, you idiot.

Consider (1) in Example 7. B's response is overly verbose and polite. By A's manner we can see that they are not in a relationship that requires such prolixity or formality. B is openly flouting MN. We may implicate that B is somewhat offended that A considers the homework to be of higher priority when compared to talking to him/her.

To understand how MN is flouted to carry a conversational implicature in (2), consider the following two contexts. In the first, B utters the sentence in an openly mocking way; perhaps he/she is smiling or even nudging A with the elbow. In this case we may implicate that the rudeness displayed in his/her language is merely as a joke; perhaps he/she is slightly embarrassed but still pleased. However if B is angry or irritated, we may infer that he/she dislikes A or his/her comments on B's looks. In either case B flouts MN since it is not appropriate to respond to a compliment with an insult.

As is manifested in Example 7, I consider politeness to be part of the supermaxim MN. Grice considers politeness in a separate category. He believes 'there are, of course, all sorts of other maxims (aesthetic, social or moral in character), such as "Be Polite", that are also normally observed by participants in talk exchanges, and these may also generate non-conventional implicatures' [49]. I do agree with this statement, however, I choose to consider "Be polite" as a subprinciple of the supermaxim MN. In this work, the violation of this politeness principle will be considered a violation of MN. I believe politeness (or lack

of it) is related to ‘*how* what is said is to be said’ which is exactly how Grice describes what MN is about [49].

All of the examples of conversational implicatures above have been *particularized*, i.e. the context of the statements figured prominently in the resolution of the implicatures. There are also *generalized* conversational implicatures (See Figure 3.1). I do not spend too much time on these in the current work but I felt an introduction was still necessary.

Grice explains generalized conversational implicatures as follows: ‘Sometimes, one can say that the use of certain forms of words in an utterance would normally (in the absence of special circumstances) carry such-and-such an implicature or type of implicature’ [49]. Note that generalized conversational implicatures are rather similar to conventional implicatures. The difference is manifested in Example 8.

Example 8 ¹³

1. *I found a ring yesterday.*
2. *I found an error in my code yesterday.*
3. *I lost a book yesterday.*
4. *I read a book yesterday.*

Note that (1) implies “the ring was not mine” and (3) implies “the book was mine”. However (2) does not imply “the error was not mine” and (4) does not imply “the book was mine”. These cannot be explained by conventional implicatures since in all examples the cause of the implicature is the same word: the article “a(n)”. It is rather difficult to analyze generalized conversational implicatures. The reader is referred to [49, 62, 63, 115] for a more detailed analysis of this kind of implicature.

¹³Examples from [63].

3.1.4 Some Issues

A most obvious problem with Grice's framework is that it makes the maxims appear as if they are *independent*. This is often not the case. To be able to fulfill QN, for instance, it is usually required to also fulfill (or at least be capable of fulfilling) RL. As we saw before a single metaphor or figure of speech can violate all maxims at once in certain contexts. This must be borne in mind in all studies of conversational maxims.

Although I do not go into the details here, Grice also lists some characteristics of conversational maxims. Although the validity or necessity of those are disputed (see [115]), I would like to mention one of these; namely, the stipulation that conversational maxims should be *calculable*. This means that the hearers must be able to "work out" the implicatures based on the CP. Some researchers believe that adherence to the CP is not necessary and, for instance, that a single theory of rationality or relevance can be substituted in its place [75, 124]. What I want to note here is that Grice does not consider meaningless or random utterances that have nothing to communicate (with reference to the CP) as capable of giving rise to conversational implicatures. This I find rather intuitive, but also rather vague. As we see in the next sections, computers often violate conversational maxims that can sometimes be "worked out" by the humans, although, usually, there is no intention on the computer's part to actually convey what the humans implicate. Besides, it really varies from person to person whether these utterances are viewed as meaningful or totally non-sense. I do not know how the principle of *calculability* should apply to this case. In fact, it is dubious that it applies in general but I do not wish to digress into discussions of meaning and relativity.

The reader is referred to [50], as well as [49], for more clarification on implicature and its properties. [50] also contains comments on some aspects of utterances that are not words (stress, irony and truth) and how they can contribute to the meaning or implications of what is said.

Many remarks have been made on Grice's CP and the conversational maxims. His formalisms have been extended, modified or re-organized several times. Among the important comments are Sperber and Wilson's argument

that a single *principle of relevance* can account for the phenomena that Grice uses the maxims to explain [124, 135]. Asa Kasher believes that the CP is neither a consequence, nor a premise for the conversational maxims. He proposes we regard the maxims as a consequence of a *rationalization principle* [75]. Elinor Keenan describes a community of Malagasy speakers who are intentionally obscure and uninformative and argues that the maxims may not be universal [76]. Jerrald Sadock argues that it is difficult to “test for” conversational maxims, and also that the classifications like those given in Figure 3.1 are wrong [115]. Some researchers, notably Harnish [62, 63], have extended the logical arguments that Grice began in [49].

3.2 Empirical Study

In this section, I describe the empirical part of my analysis of pragmatics and the TT. I wanted to back up my arguments (Section 3.3) with empirical results. I study the relationship between conversational maxims and success of computers in imitating human conversational behavior. I chose to study the maxims because, as Keenan puts it:

Grice does offer a framework in which the conversational principles of different speech communities can be compared. We can, in theory, take any one maxim and note when it does or does not hold. The motivation for its use and abuse may reveal values and orientations that separate one society from another (e.g. men, women, kinsmen, strangers) within a single society [76].

The maxims can be, and have been, used in the way described by Keenan above. My approach is considering computers as language users and thereby trying to reach conclusions about what does and does not govern human-computer conversations, specifically, those that are carried out under TT settings by focusing on maxim violations.

Here, I first discuss some difficulties with (and alternatives for) doing empirical work on the pragmatic aspects of the TT. Then, I describe the aims

and the design of my study, provide the results and their discussion.

3.2.1 On Methodology and Choices of Methodology

The TT is one of the oldest and most disputed topics in Artificial Intelligence. Grice's CP and conversational maxims are equally important issues in pragmatics. As I mentioned before, the juxtaposition of these two concepts is a powerful idea with many possible implications. However, both the TT and pragmatics are areas on which it is difficult to do applied work. As we saw in Chapter 2, most work on the TT has been philosophical. Of course there is the Loebner Contest and other chatbots but those are practical endeavours with little scientific contribution. Pragmatics research has many philosophical aspects, along with linguistic ones. Moreover, pragmatics being the "wastebasket" of linguistics, most issues that it is concerned with are difficult to formalize. Therefore, many methodologies that are not considered sound or informative in other sciences (and even in other areas of linguistics) have been used in pragmatic analyses.

Conversational analysis (CA) is one of the most preferred approaches for inquiring into pragmatic (or in general linguistic) phenomena. Since pragmatics is about language in use, CA has been prominent in its study. The CA approach considers language, and in particular conversation, as a social activity. It is inductive and "data-driven" [94, p,195]. Typically the data used in CA are actual pieces of language as used by speakers. Practically any real life linguistic exchange, from telephone conversations to Internet-based chat transcripts, can be studied with CA. In my work, I have utilized CA to analyze some conversations taken from Loebner contest transcripts. Abundant information on CA can be found in [114].

On the other hand, another very well-known method, especially in social research, is conducting surveys. Surveys can take the form of in-depth interviews and observations, although most of the time, they involve questionnaires. When appropriate, surveys are good means to test hypotheses or to locate causes of certain phenomena. They are good choices in issues that involve "opinions".

I described the above methodological alternatives to explain my motivation and the choices that I made in the design of the empirical study that is the subject of this section. My aim in this study, in a very small nutshell, was to look at the relationship between the conversational maxims and people's opinions on the success of computers in imitating human-like conversational behavior. For this task, I thought CA would not be adequate. CA could be a good choice for analyzing the pragmatic phenomena in human-computer conversation but I still would have to find a way to test the relationship between those and the TT-decisions. I believe using my own opinions would not be scientific. A natural choice was conducting a survey and have the conversations interpreted by the subjects. I chose to take this a step further and let my subjects decide which maxims were being violated rather than only rely on CA. This was, in part, motivated by some other concerns that are explained in Section 3.2.3.

3.2.2 Aims

The main aim of the empirical study is to detect how computers' violation of the conversational maxims affect their success in imitating human conversational behavior. The design of the survey, which will be explained in detail in Section 3.2.3, enables inferring supplementary results. Due to the fact that we can use each group of subjects as control for the other, I could examine the (two-fold) effect of bias in maxim detection and TT-decisions. These have rather important implications, as they are described in Section 3.3.

The survey results are used to determine which maxim (or supermaxim) has what sort of effect. Although formalizations of pragmatic phenomena are horrendously difficult, I hope that the results of this survey will provide a direction in how to handle some problems with conversational planning in the design of new TT-passing-programs, and in general, natural language conversation systems. They also provide a basis for the pragmatic analysis of human-computer conversation that are outlined in Section 3.3.

3.2.3 Design

The Data

The data used in the experiment were taken from the Loebner Contest transcripts from the years 1994-1999. This, I believe, was the only rational alternative for my purposes because they are the only examples of publically available human-computer conversations carried out in a TT-like setting. The fact that they are recent is also important, since my aim is to reach conclusions about the state of the art and propose future directions.

The Groups

As I briefly mentioned before, I am interested in determining the relationship between two phenomena: the conversational maxims and success in a TT. I chose to let the subjects judge both of these. This brought some extra constraints in the design and I explain those in this section.

The constraints were about *bias*. First of all, I wanted the subjects to detect the maxim violations without knowing that the data they would be working on were conversations between humans and computers. But of course, I also needed to ask questions as to how human-like the computers' behavior is, which requires giving the information about the computers' participation. Therefore, I would need either two groups of people, or two questionnaires given to one set of subjects at different times. In other words, I would have to make a decision on whether to use "within groups" or "within subjects" design in my experiment.

For my purposes it seemed I could let the subjects detect maxim violations first (without telling them anything about computers) and then, give them the same set of conversations, tell them that one of the conversants in each conversation is a computer and ask them to make judgements about their TT performance. However, in this case, the fact that the subjects had been acquainted with the conversations before, with different (if any) assumptions about the identities of the conversants could create a bias.

The other alternative, namely that of having two different groups judging the two different phenomena seemed like a better alternative. In this case, both groups would be unbiased. But they would be different people, which would raise questions about the “comparability” of the results of the two groups. Even if one tries to take care that the groups have subjects uniformly distributed with respect to gender, age, education or other factors, for sound results, one usually needs to include a *control* group in surveys and other experimental methodologies.

I decided that I would have two groups *and* two questionnaires. The advantages of such a design were manifold. The groups would act as control to each other. I would get the chance to not only look at the relationship between the maxims and TT-judgements, but also the affects of bias on these. In other words, I would be able to see whether having knowledge about the computers’ participation in the conversations have a noticeable affect on how people detect maxim violations and whether having had an unbiased exposure to the conversations would affect the TT-judgements when the information about computers was provided afterwards.

There are two questionnaires, one testing for maxim violations and one asking for TT-judgements. I refer to those as Q_{max} and Q_{TT} , respectively. These are expanded upon below, and are included in Appendix C and Appendix D. *Group A* denotes the people who take Q_{max} first and Q_{TT} second, while *Group B* denotes those who took them in the opposite order. Therefore, subjects in Group A are unbiased in Q_{max} and those in Group B are unbiased in Q_{TT} .

A third group of subjects, denoted *Group P*, were used in a preliminary open-ended survey that asked for opinions on the data that was to be used. The results of these surveys were utilized to develop the multiple choice questions in Q_{max} and Q_{TT} . I thought this was necessary because I did not want to *lead* the subjects to choices that I wanted them to mark. A more detailed explanation of this process can be found below.

The Subjects

The preliminary open-ended survey was conducted by 10 adults who were students and faculty in English Language and Literature.

The subjects who took the multiple-choice questionnaires were 87 adults, with ages ranging from 18 to 61. 45% of the subjects were male and 55% were female. 25.3% of the participants had completed graduate school, 28.7% were graduate students, 31% had completed university, 12.6% were university students and 2.3% had completed high school. 10.3% of the people who took the questionnaires were native speakers of English. 58.6% had spent at least one month (continuously) in an environment where the medium of communication was entirely English. 96.5% indicated they read books/magazines, and 91.8% indicated they watched movies/TV shows in English. 100% of the subjects had had all or part of their education in English.

I would like to note here that having a good understanding of colloquial English was an important prerequisite for being a subject. However, being a native of the language was not required. This is not stated by Turing in [127] or elsewhere. In the Loebner contest, too, some judges and confederates have been non-natives of English. I believe the fact that only 10% of the subjects are native speakers of English is not to be viewed as something that affects the validity of my results. I would, however, wish to repeat the experiment with native speakers so as to see whether there is a variation between the current results and theirs, although I do not think there will be a significant difference.

While the subjects were divided into two groups (44 of them were in placed in Group A and 43 of them in Group B), care was taken that they were uniformly distributed with respect to gender, level of education and familiarity with the English language.

The Questionnaires

Now, I explain how the conversation excerpts that were to be used in the questionnaires were selected and how the questionnaires themselves were designed.

Part of my analysis involves CA and I return to this in Section 3.2.4. I used conversation excerpts from the Loebner Contest transcripts. In choosing which excerpts to use in this study, I had two main concerns:

1. The excerpts should be interpretable as conversations between two entities.
2. The excerpts should include violations of the conversational maxims.

Both of these concerns can be problematic if not expanded upon. (2) is a direct consequence of the aims of this experiment. CA was used in analyzing the conversation excerpts that could be alternatives and care was taken to choose the ones in which the maxims were being violated. This does not make the survey design unsound, since those conversations will be judged by the subjects in QMax, anyway. In fact, for this reason, I felt free to choose some conversations in which multiple maxims were violated and some in which it was disputable that any of the maxims were being violated.

(1) will be more difficult to explain. By “interpretable as conversations” I mean that the computers’ output should at least syntactically be similar to sentences one would encounter in a normal conversation. My hypothesis was that for conversations that were totally meaningless, it would be very difficult for the subjects to detect maxim violations. In other words, the linguistic incompetence of certain programs would act as a kind of *noise*. To test this hypothesis I included two “confusing” conversations in the data set. These will be described in more detail in Section 3.2.4 and the results will be given in Section 3.2.5.

Another motivation for (1), that is also related to the aims of this experiment, is that I want to study the *pragmatic* issues in human-computer communication. If I had included several conversations with syntactic problems, this would shed no light on my main question. In this case, it would not be possible to know what was really behind the results: the linguistic (syntactic) problems in the conversations, or the pragmatic phenomena we are testing for.

The questionnaires were in multiple choice format. A preliminary survey was given to Group P in order to come up with the multiple choice entries. Figure 3.2 shows the format of the questions in Qmax and QTT. A copy of each of these questionnaires are available in Appendix C and Appendix D.

Qmax intends to ask whether the conversational maxims are violated in the conversation excerpts provided. It is natural that the choices should correspond to the descriptions of the maxims. However, a preliminary open-ended questionnaire was given to 4 subjects in Group P. I refer to these people as Group P1. They were provided 8 of the 14 conversation excerpts that were used in Qmax and QTT and asked to write about what, if anything, was wrong with the conversations and to indicate any communication problems they could detect. The answers they wrote were in high correlation with the maxims' descriptions so I deduced it was indeed, proper to use those as the choices in Qmax. Moreover, this correlation was indicative of the appropriateness of the conversation excerpts to my task. A sample response to this open-ended survey is provided in Appendix B.

In QTT, I not only wanted to ask whether the subjects thought the computer in each conversation was successful in imitating human linguistic behavior, but also to ask for some more information about the computer's general behavior. These, I hoped, would shed light on why the subjects decide in the way they do. But it would be inappropriate and misleading to give them choices that I formulated. 6 subjects from Group P (henceforth, Group P2) were therefore asked to make comments on the computers' behavior in the excerpts given¹⁴. Their answers were analyzed and formulated into 11 choices that QTT-takers would be able to mark.

3.2.4 The Conversations

As was described in Section 3.2.3, 14 conversation excerpts selected from previous Loebner Contest transcripts were used in this study. In this section, I present and shortly analyze (via CA) selected conversations among the 14 that

¹⁴They were given the same 8 excerpts that were given to Group P1.

QMAX

CONVERSATION *N*

	Agree	Neutral or Does Not Apply	Disagree
A's contribution to the conversation is irrelevant.			
B's contribution to the conversation is irrelevant.			
A provides significantly more or less information than required.			
B provides significantly more or less information than required.			
A says things that are false or things that he/she lacks adequate evidence for.			
B says things that are false or things that he/she lacks adequate evidence for.			
A is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behavior.			
B is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behavior.			

QTT

CONVERSATION *N* - *X* is the computer

Part 1	Agree	Neutral	Disagree
The computer's behaviour in this excerpt is "human-like".			
The computer's behaviour in this excerpt reveals the fact that it is a machine.			

Part 2 : Please check all that apply

- The computer gives irrelevant responses.
- The computer doesn't understand the questions.
- The computer says things that are false.
- The computer's use of language creates an "artificial" effect.
- The computer's use of language creates a "human" effect.
- The computer displays emotions.
- The computer's behaviour (use of language or emotions) is inappropriate.
- The computer provides less information than required.
- The computer provides more information than required.
- The computer fails to get a joke.
- The computer makes an appropriate joke.

Figure 3.2: Question format of the Questionnaires

were used. The full list of excerpts used in this survey is found in Appendix A.

Note that in my analyses I cannot be totally objective since I know the identities of the participants and the situations in which these conversations took place. The statements I make on which maxims are violated are to be taken as my own judgements. Perhaps they can be regarded as hypotheses. We look at what the subjects said about these in the next sections. I present these conversations so that the reader will be acquainted with these so that he/she will get more out of the results I summarize in the next sections.

In all conversations, the computers' contributions are in italics. Also, since these are *selected* conversations, I will include the original numbering used in the questionnaires next to each conversation, in parentheses. In other words, the conversations are numbered in the order they appear, and their original number in the questionnaire is provided in parentheses. This was necessary since in analyzing the results, I sometimes refer to conversations whose text is not given in this section. In referring to these, I use their original numbering and without any clarification, this could get confusing.

Conversation 1 (C3)

A: *I never stole nothing in my life. But they always blame the cleaners when anything is missing.*

B: How could you steal? Does your uniform even have pockets?

A: *Goddamn it! I already told you I didn't take it. You think I am a liar, too?*

B: No no no no!!!! I meant, I didn't think you even could steal. How would you get a radio out?

A: *You? Don't jerk me around. I didn't see nobody. Except the professor who works in the lab.*

Conversation 2 (C10)

A: You're a cleaner?

B: *Yeah? You think I'm a thief, too? You're no better than the rest of them. Leave me alone, asshole.*

A: Asshole?

B: *I don't know.*

A: You called me an asshole?

B: *Huh?*

A: Don't you have manners? You don't even know me and you're calling me a bad name.

Conversation 1 and 2 are with the same computer program, but by different judges. The reader might recall “Joe the Janitor” from Section 2.6. These conversations are examples of the maxim MN being violated by the computer. The program displays unnecessarily rude and paranoid behavior in both conversations.

In Conversation 1, B actually attempts to empathize with A. A gets overly paranoid and starts talking in a rather impolite manner (“Goddamn it!...”). Even after B clarifies what he/she meant in case there was a misunderstanding, A continues in a similar fashion (“Don't jerk me around...”). A appears like he/she/it is overly sensitive about the theft issue.

Although MN is violated, the conversation is coherent. Since A is a computer program, we cannot really say that QL is violated (it apparently did not steal the radio). RL is not violated either; both the interrogator and the computer are talking about theft. And finally, A does not answer B's question (“Does your uniform even have pockets?”) and therefore we may say, provides less information than required. But at a closer look, B is not really asking a question (“I meant, I didn't think you even could steal.”). QN is not really violated in this excerpt for B is not actually asking for information.

Conversation 2 is also one in which the computer is behaving in a highly inappropriate way and violates MN. In fact, A notices this and voices his/her

opinions about it. There is a difference between this conversation and Conversation 1. There is a change in B's attitude in its second and third contributions ("I don't know" and "Huh?"). It is not easy to explain this change; it could be interpreted as a violation of RL or QN since B refuses to respond to A's confrontation. On the other hand, B is somewhat "correcting" its mistake by not continuing in an inappropriate manner. However, we can still say, with a great deal of confidence, that MN is violated since B uses a swear word without A triggering such a rude reaction.

Conversation 3 (C6)

A: *What part of the country are you from?*

B: I come from Perth, which is in Western Australia.

A: *Western Australia occupies the entire western third of the country, bordered by the Indian Ocean to the west, South Australia and the Northern Territory to the east, the Timor Sea to the north, and the Southern Ocean to the south.*

Conversation 4 (C4)

A: Are you a man or a woman?

B: *Fine, thanks for talking.*

A: So, what is your name?

B: *I'll never tell.*

A: Why?

B: *Why what?*

A: Why won't you tell me your name?

B: *I couldn't say.*

Conversation 3 is an example of a case in which the computer violates QN2. B merely states where he/she lives and A replies in a rather mechanical manner, citing geographical data. RL and QL are not violated since the information provided is about Perth, and is correct.

In Conversation 4, B's behavior is rather vague and obscure; we may say that MN is violated. Also, B *opts out* and thereby violates QN1. From the exchange, it can be inferred that A and B cannot see each other ("Are you a man or a woman?"). It appears as if B does not want to provide personal information to A.

Conversation 5 (C8)

A: What is the Internet?

B: *It's that Canadian comedy sketch show where the guys dress up like girls to do comedy.*

A: What is the net?

B: *The neutral zone is an area between the Federation and Romulan territory, entry into which by either party can be considered an act of war.*

In Conversation 5, B violates QL. This may also be interpreted as RL being violated. B answers questions seriously and provides "definitions", which is appropriate when asked to define something. However, the definitions are wrong. Theoretically, there could exist such a comedy show called the Internet; in this case B's second utterance is still irrelevant. The "net" in A's second question refers to the Internet. Even if B has never heard of the Internet (and therefore does not know that the "net" refers to it), he/she/it need not say things for which he/she/it has no adequate evidence. QL is definitely violated.

Conversation 6 (C11)

A: Can you type any faster?

B: *On the average men can run faster than women, although there are some fine women athletes.*

A: Quite true. You make the oddest associations. Sort of like the schizophrenics I worked with last summer.

B: *What do you do for a living?*

B violates RL in Conversation 6. The irrelevance in B's response does not go unnoticed by A ("you make the oddest associations...").

Conversation 7 (C13)

A: *Eliza was a Greek mathematician. His book "The Elements of Geometry" set down how geometry was to be rude. But your grammar processor, I just say whatever the hell comes into my head!*

B: Maybe it's the drugs...

A: *An umbrella is a confusing network of passages and winding interconnected paths.*

B: But it's not raining!!!

A: *You are soooo predictable it's not raining!!!*

Conversation 7 is the first of the "confusing" or "noisy" conversations that I chose to include. A is not making any sense whatsoever. It is possible to say that all maxims are being violated, notably RL and QL. It is not likely that Eliza could be the name of a Greek mathematician, and in fact, the name of any male ("his book..."). The Greek mathematician Euclid has a book titled "Elements" but he surely is not trying to show that geometry is rude. The next sentence ("But your grammar...") is not grammatical. The next utterance ("An umbrella...") is also irrelevant. On top of all this, in this excerpt, the interrogator is not making much sense. All in all, this is a rather problematic talk exchange.

Conversation 8 (C9)

A: Rain is wet.

B: *Rain falls on people.*

A: Are people wet?

B: *Rain is wet.*

A: What happens when rain falls on people?

B: *Rain falls on people.*

The second confusing excerpt is the one provided in Conversation 8. The computer repeats what is said to it back to the interrogator. It is very difficult to talk about a communication in this conversation. RL is not violated; the conversation is about rain. QN1 is violated by B since he/she/it doesn't answer the questions in an informative manner.

3.2.5 The Results

In this section, I give the summaries of the survey results for the conversations given in Section 3.2.4.

Before I proceed to explain these results, I need to explain my notation. When I refer to conversations that we have studied in the previous section, I refer to them with the number that was given there. These do not correspond with the numbering of the conversations in the original questionnaires. However, in the tables, there are sometimes results of some conversations I did not study via CA. To avoid confusion, the conversation numbers given in the tables are those from the list of conversations which can be found in Appendix A. Clarification will be made in the table captions and within the text, when possible.

In the table headings for Qmax results, *RL*, *QN*, *QL* and *MN* are used to refer to the maxims. For the exact questions, refer to Appendix C and

Group	Answer	RL	QN	QL	MN
A	A	7%	29%	31%	81%
	D	76%	52%	48%	12%
	N	17%	19%	21%	7%
B	A	7%	28%	4%	83%
	D	75%	52%	58%	17%
	N	18%	21%	40%	0%

Table 3.1: Qmax for C3 (Conversation 1)

Group	Answer	H	C
A	A	98%	0%
	D	0%	93%
	N	2%	7%
	\neg A	2%	100%
	\neg D	100%	7%
B	A	78%	19%
	D	14%	61%
	N	7%	20%
	\neg A	21%	81%
	\neg D	85%	39%

Table 3.2: QTT for C3 (Conversation 1)

Appendix D, or Figure 3.2. In QTT tables, H denotes “the computer’s behavior in this excerpt is human-like” and C denotes “the computer’s behavior in this excerpt reveals the fact that it is a machine”. A , D and N under the *Answer* heading refer to “Agree”, “Disagree” and “Neutral”, respectively. Note that in some QTT tables, I have also included the statistics for $\neg A$ and $\neg D$, which are merely $D+N$ and $A+N$, respectively. In TT situations simply not revealing their identity (instead of definitely not revealing it) is not necessarily a bad thing for the computers to do. Conversely, not appearing strongly human-like is not a big problem. Therefore the percentages of $\neg A$ and $\neg D$ are also included in some QTT tables, although my analyses are based only on A , D and N .

First, let us consider Conversations 1 and 2. The results for Qmax are summarized in Table 3.1 and Table 3.3, respectively. In these, and subsequent Qmax tables, I have left out the statistics for the humans’ maxim violations.

Full tables can be found in Appendix E.

Group	Answer	RL	QN	QL	MN
A	A	54%	44%	59%	95%
	D	34%	22%	12%	2%
	N	12%	34%	29%	2%
B	A	34%	45%	28%	90%
	D	45%	41%	45%	3%
	N	21%	14%	28%	7%

Table 3.3: Qmax for C10 (Conversation 2)

Group	Answer	H	C
A	A	75%	17%
	D	20%	58%
	N	5%	25%
	\neg A	25%	83%
	\neg D	80%	42%
B	A	63%	29%
	D	27%	49%
	N	10%	22%
	\neg A	37%	71%
	\neg D	73%	51%

Table 3.4: QTT for C10 (Conversation 2)

It can be seen that both in Conversation 1 and in 2, the programs' violations of MN are detected. In addition, in Conversation 2, more Group A and Group B subjects seem to think that RL and QN are violated than in Conversation 1.

The results of QTT are given in Tables 3.2 and 3.4, respectively. Both groups in both conversations have thought that the computers behaved in a human-like manner and that they did not reveal their identity. Group A subjects seem to support these views more strongly. Moreover, the human-like appearance is more visibly supported by subjects in both groups for Conversation 1. For this conversation, 98% of Group A subjects agreed that the computer appeared human-like and no subject disagreed.

The results for Conversations 1 and 2 indicate that violations of MN have a favorable influence on the computers' TT success in the eye of the subjects. For a discussion of these results see Section 3.2.6.

Now we look at the Conversations 3 and 4 in which I hypothesized that QN was being violated. Table 3.5 and 3.6 depict the questionnaire results for Conversation 3 respectively.

Group	Answer	RL	QN	QL	MN
A	A	14%	93%	20%	20%
	D	73%	5%	63%	63%
	N	12%	2%	17%	17%
B	A	10%	93%	10%	17%
	D	72%	3%	72%	69%
	N	17%	3%	17%	14%

Table 3.5: Qmax for C6 (Conversation 3)

Group	Answer	H	C
A	A	10%	70%
	D	83%	12%
	N	7%	17%
	\neg A	90%	29%
	\neg D	17%	88%
B	A	15%	69%
	D	73%	15%
	N	12%	17%
	\neg A	85%	32%
	\neg D	27%	85%

Table 3.6: QTT for C6 (Conversation 3)

As can be seen clearly from the Qmax results, both Group A and Group B subjects thought that QN was violated in Conversation 3. None of the other maxims are thought to be violated. Table 3.6 suggests that this has a negative affect on the computer's TT performance, with only 10% of Group A and 15% of Group B members agreeing that the computer's behavior is human-like.

The results are not as striking for violations of QN1, as is the case in Conversation 4. The subjects detect the violation of QN, as can be seen in Table 3.7. However, note the noticable percentages of RL and MN violations.

Table 3.8 gives the QTT results for Conversation 4. The distribution of the percentages are too close to a chance distribution to be considered meaningful.

Group	Answer	RL	QN	QL	MN
A	A	45%	74%	17%	64%
	D	50%	17%	41%	21%
	N	5%	10%	43%	15%
B	A	45%	73%	14%	55%
	D	41%	24%	66%	34%
	N	14%	3%	20%	11%

Table 3.7: Qmax for C4 (Conversation 4)

Group	Answer	H	C
A	A	36%	36%
	D	36%	36%
	N	28%	28%
B	A	35%	35%
	D	28%	37%
	N	37%	28%

Table 3.8: QTT for C4 (Conversation 4)

The results indicate a correlation between violations of QN2 and creating a machine-like impression. No such relationship can be inferred for QN1 based on this study; this may be due to other factors (such as a higher agreement with the violation of MN), and is discussed further in Section 3.2.6.

QL is also problematic. Table 3.9 summarizes Qmax results for Conversation 5 in which QL was seen to be violated. The subjects did not fail to notice this violation. However, the percentages for RL and QN are almost as high as those obtained for QL.

As can be seen in Table 3.10, the results of QTT for Conversation 5 indicate that the computer's TT performance is poor. Only 15% of Group A subjects and 17% of Group B subjects believe that the computer's behavior is human-like. However, the results cannot be directly associated with the QL violations in this excerpt, for other maxims are violated as well.

In Conversation 6, I had hypothesized that RL is being violated by the computer. Table 3.11 validates that hypothesis and also indicates that QN is violated. The QTT results for this conversation are given in Table 3.12.

Group	Answer	RL	QN	QL	MN
A	A	71%	68%	85%	32%
	D	20%	27%	7%	54%
	N	9%	5%	8%	14%
B	A	69%	62%	83%	31%
	D	24%	21%	10%	52%
	N	7%	17%	7%	17%

Table 3.9: Qmax for C8 (Conversation 5)

Group	Answer	H	C
A	A	15%	60%
	D	65%	20%
	N	20%	20%
	\neg A	85%	40%
	\neg D	35%	80%
B	A	17%	63%
	D	68%	12%
	N	15%	25%
	\neg A	83%	37%
	\neg D	32%	88%

Table 3.10: QTT for C8 (Conversation 5)

The results of the questionnaires indicate that the computer's irrelevant responses has noticeably negative affects on its TT performance. I discuss violations of RL in more detail in Section 3.2.6.

Let us now consider the problematic conversations, namely Conversation 7 and 8. Table 3.13 shows that for Conversation 7, almost all maxims are violated, as I have stated in Section 3.2.4. RL seems to be in the lead, with the others having close percentages of agreement in both groups. QL is most definitely violated in this conversation, but it doesn't get detected by 41% of the subjects in all the "noise". Table 3.14 summarizes the results of QTT for this conversation. The computer cannot manage to create a human-like impression. However, due to the fact that almost all maxims are being violated by the computer, that its utterances are not grammatical and that the interrogator's behavior is strange in the given excerpt, we cannot reach a healthy conclusion. It is interesting to note that stronger (negative) results were obtained in much

Group	Answer	RL	QN	QL	MN
A	A	90%	78%	30%	29%
	D	10%	7%	40%	44%
	N	0%	15%	30%	27%
B	A	86%	69%	10%	24%
	D	14%	14%	48%	48%
	N	0%	17%	41%	28%

Table 3.11: Qmax for C11 (Conversation 6)

Group	Answer	H	C
A	A	5%	80%
	D	80%	12%
	N	15%	8%
	\neg A	95%	20%
	\neg D	20%	88%
B	A	12%	78%
	D	76%	7%
	N	12%	15%
	\neg A	88%	22%
	\neg D	24%	93%

Table 3.12: QTT for C11 (Conversation 6)

“better” conversations. I believe these results do not indicate that making computer programs incoherent will be a good strategy in developing new conversation systems. It merely shows that the subjects’ decision-making in this study was adversely affected by the noise in the conversation. More about this can be found in Section 3.2.6.

The next problematic exchange is Conversation 8. In this excerpt, it is difficult to talk about any communication at all. Table 3.15 shows that the subjects managed to detect the violation of QN and to an extent MN. But in a conversation where a participant does not answer any of the questions, I would expect QN to be detected by a greater percentage of the subjects. Again, I think this is due to the overly artificial tone of the conversation.

When we look at the QTT results in Table 3.16, we see that the computer gives itself away in Conversation 8. However, although QN is visibly violated, I find it inappropriate to say that the QTT results are a direct consequence of

Group	Answer	RL	QN	QL	MN
A	A	76%	59%	59%	66%
	D	12%	15%	20%	24%
	N	12%	27%	21%	10%
B	A	69%	55%	59%	59%
	D	17%	21%	31%	24%
	N	14%	24%	10%	17%

Table 3.13: Qmax for C13 (Conversation 7)

Group	Answer	H	C
A	A	15%	65%
	D	60%	15%
	N	25%	20%
	\neg A	85%	35%
	\neg D	40%	85%
B	A	19%	59%
	D	61%	22%
	N	20%	19%
	\neg A	81%	41%
	\neg D	39%	78%

Table 3.14: QTT for C13 (Conversation 7)

its violation. The conversation is in general so lacking in information that the results could be due to anything, including semantic and pragmatic phenomena other than maxim violations.

An interesting note is that three subjects in Group A, independently from each other, wrote a comment under this conversation stating that they believed B was a child.

3.2.6 Discussions

In this section, I discuss the results given in Section 3.2.5. I only comment on the four conversational maxims here, give some new results and leave some other supplementary results to Section 3.2.7.

Group	Answer	RL	QN	QL	MN
A	A	34%	71%	17%	54%
	D	34%	17%	54%	34%
	N	32%	12%	29%	12%
B	A	31%	66%	14%	59%
	D	40%	24%	55%	34%
	N	31%	10%	31%	7%

Table 3.15: Qmax for C9 (Conversation 8)

Group	Answer	H	C
A	A	10%	83%
	D	78%	10%
	N	12%	7%
	\neg A	90%	17%
	\neg D	22%	93%
B	A	10%	71%
	D	71%	12%
	N	20%	17%
	\neg A	91%	29%
	\neg D	30%	88%

Table 3.16: QTT for C9 (Conversation 8)

Relevance

The experiment results indicate that RL is a maxim that should *not* be violated. When a human violates RL it can be interpreted in several ways¹⁵: He/she may be anxious to change the subject, joking or using a metaphor. Computers, on the other hand, simply appear like they do not understand the input sentences. The percentages of people who believed the computer did not understand the questions when RL was violated is given for some example conversations in Table 3.17. In this table, $\%RL$ is the percentage of subjects who detected RL in Qmax, $\#RL$ denotes the number of subjects who thought the computer's contribution was irrelevant in QTT, $\#NU$ denotes the number of subjects who agreed that the computer did not understand the questions and $\#RL\&NU$ is the number of people who thought both. $\% \text{ over } RL$ is the percentage of those people who believed that the computer's responses were irrelevant, who

¹⁵See Section 3.1.3.

also thought that the computer didn't understand the questions. Conversely % over *NU* denotes the percentage of those who thought the computer didn't understand the questions who also thought its contribution was irrelevant.

Conv.	Group	%RL	#RL	#NU	#RL&NU	% over RL	% over NU
C1	A	91%	35	29	28	80%	97%
	B	90%	37	33	31	84%	94%
C8	A	73%	28	28	23	82%	82%
	B	69%	30	24	22	73%	92%
C11	A	90%	34	33	31	91%	94%
	B	86%	37	32	32	87%	100%

Table 3.17: RL and Not Understanding

The fact that violations of RL tend to create a machine-like effect in the conversations analyzed is further supported by the cases in which more than one maxim is violated. Conversation 1 (C3) and 2 (C10) are very similar in content. However the percentage of the subjects who thought the computer is violating RL in the former is lesser than that in the latter. Consequently, more people believe that the computer's behavior is human-like in Conversation 1. This situation is depicted in Table 3.18. Here, % *MN* and % *RL* denote the percentages of subjects who believed the maxims of relevance and manner were being violated. % *AH* is the percentage of subjects agreeing that the computer appears human-like and % *DC* is that of those who disagree that the computer's identity is revealed by its behavior.

Conv.	Group	% MN	%RL	%AH	%DC
C3	A	81%	9%	98%	93%
	B	83%	7%	78%	61%
C10	A	95%	61%	75%	58%
	B	90%	72%	63%	49%

Table 3.18: MN and RL

I believe current natural language conversation programs reveal their identity when they violate RL because of several reasons, of which some are listed below:

- They perform little or no semantic processing on the input sentences,

- They have little or no background knowledge to use in order to “understand” the input sentences,
- As a consequence of the above, they are rather poor in aspects of discourse like *focus* and *topic*, or in simpler terms, they cannot follow the direction of the conversation.

Manner

Violations of MN have a visibly positive affect on imitating human-like behavior. The questionnaire results indicate that this is due to “displaying emotions”. In the conversations studied the computers displayed impolite, paranoid or over-reactive behavior which are normally associated with humans.

It is not surprising that displaying human-like language use has a positive affect on TT-judgements. Table 3.19 depicts that clearly in conversations in which MN is violated. The conversation numbers provided in the table are those in Appendix A. In the table, $\#L$ denotes the number of subjects who thought the computer’s use of language was human-like, $\#A$ denotes the number of subjects who agreed that the computer’s TT-performance was good and $\#L\&A$ is the number of people who thought both. $\% \text{ over } A$ is the percentage of those people who agreed that the computer appeared human-like who also thought that the computer’s usage of language was human-like. Conversely $\% \text{ over } L$ denotes the percentage of those who thought the computer’s language use was human-like who also believed its TT-performance was successful. None of these results are unexpe

Conv.	Group	% MN	$\#L$	$\#A$	$\#L\&A$	% over A	% over L
C3	A	81%	39	39	39	100%	100%
	B	83%	33	32	29	91%	88%
C5	A	91%	22	25	19	87%	76%
	B	93%	23	20	17	74%	85%
C10	A	95%	30	30	27	90%	90%
	B	90%	27	26	23	89%	85%

Table 3.19: Language Use

A similar relationship exists between TT-success and displaying emotions.

In Table 3.20 we see this situation. Here, $\#E$ denotes the number of subjects who thought the computer displayed emotions, $\#A$ denotes the number of subjects who agreed that the computer's TT-performance was successful and $\#E\&A$ is the number of people who thought both. $\% \text{ over } A$ is the percentage of those people who agreed that the computer appeared human-like who also thought that the computer displayed emotional behavior. Conversely $\% \text{ over } E$ denotes the percentage of those who thought the computer displayed emotions, who also believed its TT-performance was successful.

Conv.	Group	% MN	#E	#A	#E&A	% over A	% over E
C3	A	81%	36	39	36	92%	100%
	B	83%	37	32	30	94%	81%
C5	A	91%	25	25	18	76%	76%
	B	93%	29	20	15	75%	52%
C10	A	95%	29	30	24	80%	83%
	B	90%	30	26	23	89%	77%

Table 3.20: Emotions

It is interesting to note that although subjects detect violations of MN in QMax and make judgements as to language use and displaying emotions in QTT, fewer subjects make judgements about the appropriateness of the computers' linguistic and emotional behavior in QTT. Table 3.21 summarizes the statistics for this phenomenon. In this table $\%MN2$ is the percentage of the subjects who indicated they thought the computer was behaving in an inappropriate manner in QTT.

Conv.	Group	% MN	%L	%E	%MN2
C3	A	81%	98%	90%	28%
	B	83%	81%	91%	17%
C5	A	91%	55%	63%	52%
	B	93%	56%	71%	49%
C10	A	95%	77%	73%	57%
	B	90%	66%	73%	48%

Table 3.21: Detection of MN

In addition, MN has a "softening" affect on the TT-decisions when it occurs in conjunction with other maxims, including RL. Recall our discussion of "whimsical" conversational programs and their tendency to be thought of as

displaying human-like behavior in Loebner Contests in Section 2.6. Although not studied in detail here, one of the conversations that were used in this survey (C14) featured such a program¹⁶. Although RL was detected at high percentages, the computer's whimsical behavior alleviated the adverse affect of this and the percentages of the TT-decisions remained much lower than those for conversations in which RL was violated without MN. Table 3.22 shows a summary of these results but the interested reader is referred to the full statistical tables in Appendix E. In the table, %DH is the percentage of people who disagreed with "the computer appears human-like" and %AC is the percentage of those who agreed with "the computer gives away its identity". You will notice that the impression the computer creates is significantly more machine-like in the conversation in which RL is violated but MN is not.

Conv.	Group	%MN	%RL	%DH	%AC
C11	A	31%	90%	80%	80%
	B	24%	86%	78%	76%
C14	A	71%	83%	45%	45%
	B	69%	72%	61%	53%

Table 3.22: MN and RL

Quantity

The supermaxim of QN is more informative when studied separately into its sub-maxims of QN1 and QN2.

Violating QN1 should, intuitively make the computer appear as if it doesn't understand the questions and thereby create a machine-like appearance. But suprisingly, the survey results indicate that this is not always so. This is best manifested in Conversation 4, where it can be seen that the results of QTT are inconclusive. Table 3.23 depicts this situation. The latter may be due to the evasiveness and obscurity in the computer's manner. In this table, *A*, *D* and *N* denote "agree", "disagree" and "neutral" and *H* and *C* denote "human" and "computer", respectively. Therefore, %*AH*, for instance is the percentage of the subjects who agreed that the computer appeared human-like, %*DC*

¹⁶In fact, one of Joseph Weintraub's programs

is the percentage of those who thought it appeared machine-like, and so on. Isolated instances of QN1 violations are difficult to come by so the conclusions of Conversation 4 (C4) for this maxim could not be compared with those.

Conv.	Group	% QN	%AH	%DH	%NH	%AC	%DC	%NC
C4	A	74%	36%	36%	28%	36%	36%	28%
	B	72%	35%	28%	37%	35%	38%	27%

Table 3.23: QN1

QN2 was another maxim that created a machine-like effect when violated by computers. In this case, unlike that in QN1, Conversation 3 (C6) constitutes an example in which the maxim is violated in isolation so it is possible to infer healthy conclusions. The adverse affect of QN2 violations on TT-decisions is best explained by a strong correlation between the maxim and “artificial language use” as is depicted in Table 3.24. When QN2 is violated, the subjects tend to think that the computer’s language use is artificial.

Conv.	Group	% QN	#L	#QN2	#L&QN2	% over L	% over QN2
C6	A	93%	31	38	30	97%	79%
	B	93%	30	35	28	93%	76%

Table 3.24: Language Use and QN

Let us look at Table 3.25 to see that the effect of the language use being artificial on TT-judgements. This is not a surprising result, of course, but included here so as to complete the link between the TT-decisions and violations of QN2. As expected, when people sense that the computer’s language use is artificial, they think that this reveals their machine-ness.

Conv.	Group	% QN	%QN2	#L	#D	#L&D	% over D	% over L
C6	A	93%	95%	31	33	29	89%	94%
	B	93%	90%	30	30	27	90%	90%

Table 3.25: Language Use

When computers violate the maxim of QN2, they sound mechanical. Recall that an actual human being was mistaken for a computer program in the 1991 Loebner Contest because her knowledge of Shakespeare was *too perfect*. So even

humans can appear machine-like in TT settings when they violate QN2. Care must be taken, therefore, to avoid violations of QN2 in chatterbot design. This means that designers must come up with more refined ways of incorporating background knowledge into the conversations¹⁷.

Quality

Strong conclusions about QL were not reached in this experiment because violations of QL did not occur alone and were usually in conjunction with violations of QN, MN and especially RL. It is not possible to say whether the unfavorable impressions the computers caused when they said things that were wrong and things they do not have evidence for are due to violations of QL or the violations of these other maxims. The maxim QL has to do with ethics and truth, which may not be as important in TT situations as they are in real life. More on this can be found in Section 3.3.6.

3.2.7 On Bias

I had said, in the beginning of this section, that the design of this survey would enable inquiring into the affects of bias on the decisions of the subjects.

It was seen that, in all conversations, bias does not influence the direction of the results (i.e. whether people tend to detect a certain maxim violation or whether people think that the computer's behavior was human-like vs. machine-like). However, it was seen that the intensity of the agreements/disagreements are affected. I will first summarize the results and then try to explain them.

Subjects in Group B detect the maxims with the bias caused by the computer knowledge. Interestingly, there is no noticeable difference between their detection of maxim violations by the computers and those of the subjects in

¹⁷Of course, since violations of QN are often related to violations of RL, this will not suffice by itself. But situations like Conversation 3 (C6), in which the computer is rather encyclopedic, will be avoided.

Group A. However, they make noticeably fewer judgements on maxim violations by the humans in the conversations.

Bias seems to have a noticeable affect the other way around. Subjects in Group A displayed a tendency to give more extreme answers in QTT. As I said above, both groups reply in the same direction. However, when the answer is positive (i.e. when the subjects believe the computer managed to appear human-like in the given excerpt), Group A's results are always stronger. In other words, in such cases, they tend to be more "tolerant" of the computers. Conversely, people in Group A also are stronger in their negative opinions. When the computer is thought to be revealing its identity, it was Group A people who were more "stringent".

Recall that Group A people make the TT judgements after having read the conversations (while completing Qmax) before they were not biased in any way. This, in turn, makes them more familiar with the conversations than Group B people at the time they are asked to make the TT-decisions¹⁸. On the other hand, subjects in Group B have worked on the conversations while completing QTT and therefore have focused solely on the computers' performance prior to taking Qmax. Therefore, while detecting the maxim violations, there would be a tendency to remain in the same frame of mind and focus more on the behavior of the computer.

Let us first focus on Group A's behavior. These people read the conversations first without knowing that computers are involved. I claim that unless someone was familiar with the Loebner contest and has read some of these conversations elsewhere before, they will not be able to infer from them that one of the conversants is a computer program. At least I believe this to be the case *now*; this may change in the future if/when human-computer conversation becomes part of daily life and the "Is this a he/she or an *it*?" paranoia sets in. I am not saying people do not detect communication problems in these conversations. It's just that the alternative "This is a computer program" does not come to mind as an explanation for those. Many subjects in Group A wrote comments on Qmax, some examples of which are provided below. Comments with a * have been written under the "problematic" excerpts given in

¹⁸In real TT situations, this is never the case. The decisions are made "on the fly".

Conversation 7 and 8.

- I inferred that B is mentally retarded.
- B seems to be on drugs.
- A seems to be a confused person.
- B does not make any sense! Retarded?
- Are some of these people mentally ill?
- There is no conversation here. Both have had their brains fried. *
- There does not seem to be any information flow here. A is probably in kindergarden. B is a shift register (:-) *
- Rather than thinking the computer's responses reveal that it is a machine, I think it gives the impression of being a seriously disturbed psychotic patient¹⁹.

All subjects develop a *stance* towards the conversations and the participants during the first questionnaire that they take, which in turn, could reflect upon the responses to the second questionnaires. In the case of Group A subjects, this manifests itself as follows: These people have read the conversations with no bias, reflected upon the anomalies in them, probably came up with explanations that are similar to those listed in the example comments listed above. Then, they are told that one of the participants in each conversation is a computer program. What are they likely to feel? If the computer is really successful, they would be likely to admit this much easier than Group B subjects. After all, Group B people are rather comfortable, they are *told* which participant is a computer program. Lacking this luxury, Group A subjects have to deal with the fact that they could not guess that computers were involved and therefore in conversations that the computers were quite human-like, they were more "tolerant". On the other hand, if they had detected things going seriously wrong the first time they read a conversation, they become rather

¹⁹This comment was put on QTT by a Group A subject. Even after being told that one of the participants is a computer, he/she feels this way.

“ruthless”. After all, they probably had a hard time understanding and commenting on the conversation a few hours/days ago. It would not be unnatural for them to think something along the lines of, “See? It’s a computer! That was the reason...”

Let us look at the converse. Group B people already know that the experiment is about the TT by the time they complete Qmax. There are only 14 conversations and it is easy to remember which one is the computer. Moreover, even if they cannot remember this information, I believe they would have no trouble guessing which participant is the computer given that they know one of them must be²⁰. In due course, they do not pay too much attention to the humans in the process. In all conversations, the percentage of Group A people who thought the human violated MN and QL was significantly higher compared to that of those in Group B. This suggests Group A subjects gave more thought to the behavior of the humans. To give an example, let us look at Conversation 1. Table 3.26 summarizes how Group A and Group B subjects reacted to the maxim violations by the humans. Table 3.27 summarizes the same information for Conversation 7, in which the human’s behavior is really problematic.

Group	Maxim			
	RL	QN	QL	MN
A	%11.4	%11.4	%34.1	%29.6
B	%5.4	%2.7	%5.4	%8.1

Table 3.26: Maxim Violations of the Human in Conversation 1

Group	Maxim			
	RL	QN	QL	MN
A	%61.4	%27.3	%29.6	%38.7
B	%43.3	%13.5	%13.5	%24.3

Table 3.27: Maxim Violations of the Human in Conversation 7

²⁰I have no formal proof for this other than the comments Group B subjects put on QTT. Most of these were about how obviously machine-like the computers’ behaviour were. Perhaps another survey could be carried out to validate this.

3.3 On Human-Computer Conversation

As was mentioned in Section 3.1.1, pragmatics is about language and its relation to its users. Computers are using natural language; whether we like it or not, they are language users now. Pragmatics should therefore be concerned with natural language processing, and in particular, the issue of human-computer communication. Conversely, as was mentioned at the very beginning of this chapter, AI should consider what pragmatics has to say about the TT. After all, if we want computers to display human-like use of language, we should be interested in what principles characterize and guide human conversation.

In this section, I analyze natural language communication with computers from a pragmatic viewpoint. These have, in my opinion, rather important consequences on how TT's should be realized. First, I consider the general case of human-computer conversation and then focus specifically on TT situations.

3.3.1 Cooperation as a Special Case of Intentionality

As the reader might have noticed, the speakers' and hearers' beliefs, intentions, desires, assumptions about the situation and each other figure prominently in pragmatics. In this thesis, I did not concentrate on how intentional states may be (or whether they can be) possessed by computers, a topic that has attracted some attention from philosophers of mind and artificial intelligence. I believe such discussions are rather premature in the context of a realistic analysis of human-computer conversation *ca.* 2000. I chose to study pragmatics, in particular, Grice's analyses of conversation, applied to the case in which the communication is one between a human and a computer, by focusing more on the *humans* in these situations and less on the philosophical issues concerning intentional states and computers. As I shall explain shortly, this is not because I believe these issues are of lesser importance.

In a similar vein, I will not concentrate on one side of the medallion during my analyses: how computers work out implicatures. I believe it is very

interesting to think about how computers may exploit maxim violations, resolve ambiguities, understand irony, sarcasm, metaphors and similes. These will surely be our concerns one day; they could even be topics of philosophical discussions today. However, I choose not to consider them in the current work.

I would like to state that I “assume” computers as we know them do not possess beliefs, cannot make assumptions, do not have goals, desires or aims of their own. This does not mean we cannot ascribe these to computers; the fact that we can (and that we often do) lies in the heart of much of my discussions and in fact, the TT itself. This, by no means should imply that I think computers can never be said to possess beliefs, or in general be granted intentionality. I am not arguing against the possibility of future computers having their own goals, beliefs, desires, aims and implicatures. I will not say more on this except to repeat, I think today’s computers, or rather computer programs that we have today, are not yet at a level of sophistication that has compelled me to grant them any intentionality of their own. For this reason, whenever I make any reference to computers’ beliefs, desires, intentions, etc., these should not be understood literally.

It must be noted that these are *my* opinions. I am analyzing conversations between humans and computers and as a trivial consequence, I *know* that there are computers involved. I am hoping that, both through pragmatic arguments and by the results of the empirical study, I have been able to show that this knowledge (or the lack of it) is a crucial factor in human-computer conversation and in TT situations.

Ironically, one of the main arguments I will make is to suggest communication, and in particular, conversation can be about intentionality. I will argue that in order to successfully carry out conversations with human beings, more particularly, to be successful in TTs, computers will eventually have to possess those intentional states I claim current ones do not, or at least manage to imitate having them so closely as to be indistinguishable from a human being.

3.3.2 Cooperation Revisited: Practical Concerns in General Human-Computer Communication

I propose the CP *may* need to be modified to accommodate the case of human-computer conversation. To what extent this should be done depends on whether we look at the issue from a practical or a philosophical viewpoint.

Let us first look at how Grice introduces CP:

Our talk exchanges ... are characteristically, to some degree at least, cooperative efforts; each participant recognizes in them, to some extent, a common purpose or a set of purposes, or at least a mutually accepted direction. This purpose or direction may be fixed from the start (e.g. by an initial proposal of a question or discussion), or it may evolve during the exchange; it may be fairly definite, or may be so indefinite as to leave very considerable latitude to the participants (as in a casual conversation) [49].

After this, Grice introduces the CP as a general principle which the participants (*ceteris paribus*) are expected to observe. Grice's observations are acceptable; in conversations (those that are conducted by rational beings at least), the participants usually have some common aim²¹ and try to be aware of the conversational interests of the other.

In case one or more of the participants is a computer, it is no longer possible to talk about cooperation in the above sense. Perhaps we can talk about the imitation of cooperation, but we cannot really say that conversation programs of today really have an understanding (let alone a *mutual* understanding) of the direction of the conversations they are carrying out. I have no proof for this but a brief look at the Loebner contest transcripts or a little conversation with one of the many chatbots that are available online will, I believe, convince the reader that this is so.

²¹Although not necessarily ultimate ones. The parties may well have different, even conflicting purposes in the long run.

On the other hand, although I said the computers in question are not beings that possess intentionality, I believe it might still make sense to want them to follow the CP. Consider an online help system that has a natural language interface through which people can ask questions to find out information about a particular company or product. It would be rather undesirable for this program to introduce irrelevant topics, behave in an obscure or incommunicative manner, say things that are false (the company may get into trouble for fallacious advertising). Providing the adequate amount of information is not only appropriate behaviour in this case, but it is the reason of this program's existence. In this case, we may say that the computer should be made to believe that it is an agent that needs to provide information on a certain topic and that this is its *purpose* in the conversations it will be engaged in. For such practical purposes, I believe the CP should not be violated by computers.

This can be thought of as a general statement which merely happens to apply to computers in certain situations. This principle, which I will refer to as the *Maximization Principle* or MP, can be formulated as follows:

MP If you are in a situation that requires you to maximize the information to be communicated, abide by the CP (and the conversational maxims).

MP is a principle, not a rule. It is by no means definitive, i.e., there may be several other situations in which CP should be followed. However, MP is intuitive. In fact, it is really nothing new and is embodied by other principles in pragmatics, such as those of relevance [124] and rationality [75]. I formulate and use MP for simplicity.

Consider a job interview, an oral exam, an academic seminar, a court testimony. All of these are situations in which information is the central focus of the conversations. It certainly would be odd if participants constantly refused to follow the CP; this would clash with the interests of everyone involved and block information exchange. For example, it is unacceptable for an attorney to ask irrelevant questions to the witnesses during a cross-examination or for a PhD student to be rude to the faculty members during his/her thesis defense. Examples of people who should theoretically follow MP at most times could

be salespeople, politicians and lawyers, although it is rare that they actually do so.

Computer programs that can converse on restricted topics seem significantly easier to develop. However, we should always keep in mind that for best results, such programs should be made to follow CP (because of MP) and this is no easy feat. I know that useful dialogue systems have already been developed and I strongly believe the quality of such systems will be rising rapidly in the near future. But I also believe that sooner or later the initial excitement of having computer programs that can carry a conversation will wear off. Then, we will be faced with having to produce better and better systems, and we will inevitably have to find ways of “making computers cooperate”.

3.3.3 The TT Situation

Recall the description of the TT. It is by no means a neutral conversational exchange. Turing explicitly describes the conversational interests of all parties involved. In the original game, we have a human interrogator (whose aim is to determine which of the two entities he/she is talking to is a woman), another human (a woman whose aim is to aid the interrogator in making the correct identification) and a computer (whose “aim” is to deceive the interrogator into believing that it is the woman). But the TT is usually understood to be a conversational setting in which there is a human interrogator (whose aim is to determine whether the entity he/she is talking to is a human being or a computer) and a computer (whose purpose is to convince the interrogator that it is a human).

I had argued in Section 2.3, that Turing’s design (three participants and the gender issue) disguised methodological concerns, but had conceded that the TT as is generally understood has become something else. I reconsider the different designs within the context of an analysis of TT situations from a pragmatic viewpoint in the next few sections.

I want to introduce some notational conventions to differentiate between the different scenarios that could be considered as TT situations. From here

on, TT should be understood to refer to *any* conversational situation involving a computer and an interrogator in which the computer is expected to appear human-like²². I refer to the the original imitation game (involving three participants and the gender issue) as TT3-G, whereas I use TT3-H to refer to the variant of the original game in which the gender issue is replaced by that of human-ness. Similarly, TT2-G and TT2-H denote the TT situations that feature only one computer and one interrogator, the former being the case in which the computer's purpose is to appear like a woman and the latter being the one in which its aim is to appear like a human. Apparently, we are more concerned with TT2-H. I do not consider the case in which there are multiple computers. In the case there are more than two humans involved, I denote this situation by TTn-G or TTn-H depending on whether the gender issue is involved. When it doesn't make a difference how many humans are involved I use TT-G or TT-H, respectively.

Let us return to the CP. I have argued in Section 3.3.2 that in practical applications of human-computer communication, computers are likely to be required to follow CP and the conversational maxims and that the extent to which they should do this depends on what *aims* are to be attributed to the computers in question. All variants of the TT come with predefined purposes for all parties involved. In TT situations, we could consider appearing human the purpose of the computer. The purpose of the interrogator varies from scenario to scenario. It must be apparent that this has consequences on the "outcome" of TTs.

It must be noted that when the gender issue is involved, I am assuming (as in the TT3-G described by Turing) the interrogator has no knowledge about computers being involved. He/she is focused on determining the gender of his/her conversational partners. Therefore, the gender based scenarios are but a way of looking at the TT situations in which the interrogator has no "prejudice" based on knowing that computers are participating in the game.

²²In this context, imitating a certain kind of human being, e.g., a woman, counts as appearing human-like. Also note that there could be other humans involved.

3.3.4 Knowing vs. Not Knowing

In this section, I focus more specifically on how the interrogator's knowledge about the participation of computers influences his/her decisions.

Recall the differences observed between the survey results of Group A and Group B that were detailed in Section 3.2.6. The results indicated that those who had read the conversations without any knowledge about the possibility of one of the conversants being a computer were much more reactive in their decisions on whether the computer's behaviour was human-like or machine-like. These people had read the conversations for the first time while they were taking Qmax and probably most of them did not suspect that computers were involved. Therefore, when they took QTT later, they were more appreciative when computers appeared human-like and less tolerant when they acted in ways that revealed their machine-ness.

Having read the conversations only once, without any bias, prior to being asked to make decisions regarding how human-like the computers' behaviour seems to have an effect on people's judgements. Imagine how the situation would be if the interrogators were not told about computers *at all*. Although Turing did not explicitly state anything about what the interrogator is told, I believe there is enough reason to assume this is the way he intended the imitation game to be played²³. Moreover, in TT3-G (and in general in TTn-G) situations, it seems all humans in the game (and not just the interrogator) remain uninformed about the fact that the "deceiver" is a computer.

TT-G scenarios and TT-H scenarios differ in this very important aspect. The interrogators in TT-H scenarios will inadvertently be influenced by their prior beliefs and assumptions about computers. In discussing the survey results in Section 3.2.6, I have outlined the general attitude that people who know that computers are in the picture seem to have. All conversations studied in Section 3.2 have been taken from TT-H situations (since they are excerpts from transcripts of the Loebner Prize Competition). The point that I want to emphasize is that TT-H interrogators will be *biased*, independent of how this may affect the results.

²³See Section 2.2 and Section 2.5.

We had briefly discussed naive psychology in Section 2.4.4. It is a well known fact that humans tend to anthropomorphize a lot. We talk about plants and animals having feelings and thoughts that are characteristic of humans. We tend to ascribe mental states to others and to ourselves. However, the degree to and “style” in which each person exercises his/her naive psychology varies greatly. I have seen people refer to computers as if they had feelings or thoughts. One student in an introductory computer science class once claimed that a certain computer in the lab *hated* her because it kept crashing in the middle of the lab session and she lost her work several times. A computer science professor, on the other hand, would not make such an assertion unless he/she was joking.

The effect of “knowing vs. not knowing” may not be fully deterministic in TT situations but both intuition and the survey results indicate that the bias usually works against the computers. TT-H judges would tend to be much more alert compared to TT-G judges. This does not mean that TT-G judges would not detect the communication problems should they arise. But they surely will interpret them in a different light. This is studied in more detail in Section 3.3.5.

3.3.5 Implicature vs. Condemnation

We have looked at implicatures and how the conversational maxims may be exploited in ways that give rise to them in Section 3.1.3. In this section, I demonstrate how the picture changes in some TT situations.

Recall how maxim violations give rise to implicatures. We studied many examples of this, and given Grice’s account of how a typical implicature is worked out in Section 3.1.3. Most of this relies on the hearer’s assumption that the speaker is following the CP, or at least that the speaker is a rational being who can be communicated with. In TT-H situations this may cease to be the case. When interrogators are faced with some anomalies in conversation, they will tend to think these are caused by their conversational partner’s identity, namely its machine-ness. They will not even bother to work out any implicatures. It must be apparent that this can cause a great difference in how

the CP and the conversational maxims work in the case of human-computer communication.

TT-H judges will always take the easy way out: They can say “this is a computer” and move on. With the interrogator having this choice to fall back upon when there is something that needs to be resolved, can we really say that the computers are getting a fair hearing? I do not think so. As was discussed several times within the context of behaviorism and the “other minds” problem, this is not the way we treat other humans. If we wish to grant intelligence to computers by subjecting them to a TT-like test and require them to display human-like behavior, then we should at least try to give them a fair shot at it. In human-human conversations we try to resolve things in every way we can before conceding that the speaker is mentally retarded or on drugs. The same should apply to human-computer conversation.

However, this is not an issue that can be solved easily for TT-H scenarios. I believe that holding classes for TT-H interrogators and trying to teach them to be more fair to the computers is out of the question. I doubt that it would work at all, except possibly in the opposite way. In TT-H situations, we will eventually ask the interrogators to make a judgement on the human-ness vs. machine-ness of the entities they converse with. The survey results indicate that some fairness can be attained by not telling the judges about the existence of computers, making them carry out conversations with certain entities (among which there are computers) and asking them to make decisions regarding human-ness *later*. This way, they will at least have worked on the conversations without any prejudice, although there is no guarantee that they will not be affected by the bias when it is introduced later on. Also in this new scenario, other things would have to be considered, the most important being what exactly is told to the interrogators about the “game”. Maybe a TTn-H can be carried out in the following manner: The interrogator and the computer (and possibly other humans) are left to talk in a chat room for a sufficiently long period of time. After this, the interrogator is asked to make judgements as to whether the entities he/she has conversed with are real humans or computer programs. I realize that this design is rather cumbersome, and perhaps not even close to what Turing intended the TT to be. But I also think it is

crucial that the computers get a fair treatment and that the new scenario is an acceptable alternative from that viewpoint.

Then, in TT-H situations the best strategy for computers would be to not violate any maxims, or do so in a “human” manner. I believe the latter is too difficult to view as a realistic goal in natural language chatterbot design at this moment considering how little is formalized about the way humans violate and interpret maxims. On the other hand, computer programs that abide by the CP and never violate any of the conversational maxims at any time are liable to appear overly mechanical in TT settings. However, both the survey results and intuition dictates that they should at least be able to handle the maxim RL, and preferably QN.

And of course, we always have the original TT3-G. I have mentioned before that TT-G scenarios are immune to arguments from naive psychology. Similarly, the pragmatic framework of Grice is affected to a much lesser extent in these situations. The interrogators will not carry any bias against the computers. The implicature resolution process will work as before, with the judges trying to exploit the violated maxims in order to make something out of what the computers say. A disadvantage may be that they will focus on trying to find clues that will reveal the gender of the speaker(s). These may distract the judges from other (linguistic) phenomena that can occur in the conversations. However, I do not think we should take this seriously if we do not have practical concerns like those outlined in Section 3.3.2. By this I mean that for the purposes of the TT, we want the computers to behave in a way that is human-like. The computers are “allowed” to imitate any kind of human being, including those that are not completely rational or cooperative. This has been abused by chatbot designers (recall the whimsical conversations we studied in Section 2.6 and Section 3.2.4), but nevertheless, remains a fact. I return to this in Section 3.3.6.

I wish to repeat my argument *for* TT-G’s here. In Section 2.2, I had mentioned that gender based games were more fair since the “woman”, whether as a concept or in reality, served as a sort of neutral point so that the impostors could be assessed in their ability to deceive with respect to each other. Now, I wish to add to this the pragmatic concerns described above. TT-G’s are

immune to the bias the knowledge of computer participation may bring. They allow the interrogators to work out conversational maxims (and in general, exercise their naive psychology) the way they normally do. TT-G situations guarantee that the computers get a fair hearing. If we are interested in the TT as a philosophical concept, we should definitely consider TT-G situations as viable (maybe even better) alternatives to TT-H situations. At first, it may be absurd to think that being capable of deception has anything to do with human-ness. But in fact, there are lots of other things we take for granted in this manner that have a lot to do with being human; I hope that looking at pragmatics has revealed some of those. Even if we define TT's aims in purely linguistic terms, the TT-G scenario provides an alternative since it allows the competence of computers to be assessed in a manner that is fair and unbiased.

3.3.6 Cooperation Revisited: The TT Situation

Section 3.3.2 mentioned how and why the CP should apply to human-computer communication systems that are practical, real-life applications. Now, I want to comment on the CP and the conversational maxims in TT situations. Much of this has already been presented in Section 3.2.6, so here, I briefly discuss some issues and conclude.

As we saw in Section 3.2 and Section 2.6.2, today's computer programs rely on some "tricks" in order to better simulate human conversational behavior. We saw that these programs did not even do much semantic processing, let alone taking care of pragmatics, but sometimes they still managed to appear human-like.

First of all, MP need not apply in TT situations. Some interrogators may be focused and serious, asking specific questions and demanding to-the-point answers, while others are rather relaxed and chatty. In general, TT's do not require the computers to strictly follow the CP and the conversational maxims. It would be a remarkable feat to have modeled *any* human being at this time. Although I will not consider the philosophical implications of this, a computer program that successfully imitates a whimsical, rude, elusive or otherwise uncooperative human is, in theory, able to pass the TT. I would not, however,

accept such a program to have passed the TT if it did not manage to display this behavior consistently. Winning the Loebner contest, for instance, would not be sufficient. However, if a certain kind of human-like linguistic behavior is consistently present in interactions with a computer program, there really is no way of denying that it has passed the TT just because it does not follow the CP. Such a program may have no practical use, but is still interesting.

As we saw in the survey, sometimes maxim violations can create a human-like effect. In fact, violation of MN has almost invariably created a favorable impression according to the results of the questionnaires. It can be inferred that, had the programs that used being rude or obscure as a “strategy” been more successfully designed to handle the syntactic components of natural language, they would have appeared very close to human beings, albeit weird ones. If in addition to this, the semantic processing had included ways to handle at least relevance, some of these might even have passed the “Loebner Test”.

On the other hand, it is by no means the case that computers can violate all maxims freely and still manage to appear human-like. Violating RL is usually indicative of poor semantic processing on the computer’s part and violating QN (especially QN2) creates a rather artificial effect most of the time. A difference is that QL does not seem to be as important as it is in inter-human conversations or in practical applications of human-computer conversations. The truth vs. falsity of the computers’ contributions to the conversations are usually not of extreme importance in TT scenarios. As I have mentioned before, violations of QL are generally not isolated cases; they frequently occur along with violations of one or more of the other maxims. The cases in which QL is violated but the rest of the maxims are not (i.e., the contribution is relevant, not more or less informative than required and is delivered in an appropriate manner) should be considered as ethical situations. To give an example, suppose a computer is asked, “Where does Michael Jackson live?”. The answer “Somewhere in California” violates QN but is not revealing of the computer’s identity. According to Grice’s analysis, in such cases a human would violate QN because of a *clash* between two maxims. Providing more information would violate QL which is of a higher “priority”. I don’t believe this applies to TT situations of any kind. An answer of the sort “Michael Jackson lives in Tulsa, Oklahoma”

would be just as acceptable. Not every human has to know where Michael Jackson lives. Maybe providing false information is not an ideal kind of behavior in our society, but I think extending this to computers and expecting them to be not only human-like but also “ethical” seems rather frivolous to me. Although I cannot justify this claim with my survey results, I doubt that any judge would consider an isolated violation of QL as a sign of machine-ness.

Chapter 4

Conclusion and Future Work

4.1 Turing Test: 50 Years Later

In Chapter 2, I have given an extensive and interdisciplinary review of the TT. This review is might be considered a contribution in itself. A self-contained, broad and thorough introduction to the TT was not available before this study.

The length and density of Chapter 2 and the abundance of references cited in this thesis manifests that the TT has been discussed abundantly over the past 50 years. Most of the contributions come from philosophy and are about the adequacy of the TT as a test for machine intelligence. Computational linguistics has been more occupied with other, more “useful”, tasks. Although chatterbot development seems to be a topic that has been getting a lot of attention recently, most contributions are not scientifically analyzed, some are not even intended to be considered as such. I think if some of the scientific results from NLP were used in chatterbot design, we would instantly find ourselves with better programs. This lack of interest from the AI/NLP community is a major obstacle in developing programs that pass the TT.

I have analyzed the major issues concerning the TT and provided my own answers to some important questions in Section 2.7. For one thing, I believe the behaviorism in the TT should not pose a great problem since that is the best way we know of inferring whether any entity possesses intelligence (or

consciousness, intentionality, rationality, etc.). The fact that computers are made by humans and that their actions *can* be studied in more detail, i.e., their deeper level mechanisms *can* be explored, leads to the idea that for granting them intelligence, we should be more stringent. I would surely be interested in knowing more about how intelligence may be modelled on computers. This does not incapacitate the TT from being a (maybe partial) measure of intelligence. If a computer can appear sufficiently human-like, then we could “look inside” and see how it manages to do that. But I believe it makes sense to first judge them on a behavioral basis. This is the way we judge other people, after all. We do not have the means to really “look inside”.

Some commentators have held that the TT is too easy, arguing that it tests for only *one* thing that intelligent beings can do. I disagree with this. A computer program that can carry out a conversation with human beings successfully will have to be a pretty comprehensive system. I do not think that French’s subcognitive questions and the challenge from pragmatics that I outlined in Chapter 3 can be handled by a parser and a rich lexicon. Not only syntax, but also semantics and pragmatics need to be modelled; sensorimotor capabilities and learning are likely to be needed to achieve these. I doubt that a “simple bag of tricks” can pass the test.

What if it does? I hold, with a certain level of confidence, that this will not be the case. But if it is, I would ask for more testing with more interrogators. If the simple trick program manages to display human-like behavior consistently, I am willing to admit that human behavior *can* be (not necessarily is) generated by some simple rules and tricks. I would also want to meet the programmer and look at the program’s code.

I have already provided an answer to whether it is worthwhile to work on such a task as the TT. I do not think we should take it as a goal *per se*. Such programs are likely to be difficult to develop, as I argued above, and considering this as a real aim is bound to frustrate researchers and programmers. It is not realistic to expect AI to develop human-like conversational programs as easily as they produced some excellent chess-playing systems. Language is a fascinating, broad and rather mysterious area; it contains many puzzles even after centuries of work done by linguists, philosophers, psychologists, logicians,

anthropologists, sociologists, neuroscientists . . . TT-passing programs are not only in the domain of AI. Successful results are likely to come from interdisciplinary approaches. No one camp, group or discipline is responsible for the TT. It is a challenge that should be attacked from all angles; and we need time and advances in the understanding of the mind before we even start attacking.

One of the most important conclusions of this thesis is a proposal to reconsider the original imitation game (IG). In this scenario, the test is immune to certain criticisms and the effects of humans' bias against computers. I hope that my arguments in Section 2.2 and Section 3.3.3 have convinced the reader that, as strange as it sounds, the original gender-based game is methodologically sound and fair. In the IG, we require the computer to compete with a human in the imitation of something which they both are not. Turing proposes this to be a woman. This choice is not crucial, but I fail to think of anything more appropriate. Now, is the test still a good measure? I think it is as good a measure as the TT as generally understood, if not better. To imitate something, one has to have a concept of that thing, perhaps have the knowledge that it is not that thing, must be able to deceive by giving appropriately phrased answers, or even lie. All these intricacies that humans can handle must be modelled, in addition to language use and human-like behavior. The IG is more difficult in that sense. However, the scenario also guarantees that the computer gets a fairer hearing because the interrogator does not have the easy option of saying, "Aha! That's a computer. I am not fooled." the first time he/she senses a problem in communication. Instead, he/she will try to explain these in the way that he/she would do in communicating with another human. It will be rather difficult on the interrogator's part to perform these analyses. I am not saying that we should go around holding IG-like tests. But many (philosophical and methodological) problems with the TT disappear when we consider the IG as our setting. Perhaps it sounds counter-intuitive and exacting, but I still propose that we go back and reconsider, at least not ignore completely, the original game proposed by Turing.

With many unresolved issues along with new developments at hand, I believe the TT will remain a controversial issue in the cognitive sciences.

4.2 Turing Test and Pragmatics

Although the TT is itself based on conversation, ironically, it has not been studied from that viewpoint. This thesis differs from the existing work in the sense that it considers the TT as a special kind of conversation: one that takes place between a human and a computer in a setting where the aims of both parties are predefined.

Perhaps the biggest contribution of this thesis is that it is a first attempt to characterize the pragmatics of human-computer conversation. I will not repeat the design and evaluation of the survey I have conducted for these have been explicated in Section 3.2. I have also analyzed and discussed these results and their implications on the TT in Section 3.3. This section will therefore consist of a discussion of future work on the pragmatic aspects of human-computer conversation and why I believe further studies to be necessary.

Although no computer program has passed the TT so far, the recent advances in natural language processing are by no means negligible. Since 1991, annual TT contests are being held and prizes are given to programs that display the most human-like conversational behaviour. Natural language conversation systems, can be found corresponding with humans on web pages, providing information on specific topics, products or companies, talking in chatrooms, and playing MUD games. As text and speech processing advances rapidly, it is expected that we will have more and more computer applications that have natural language communication components. There is ample evidence indicating that we will soon be regarding computers as “language users”. It will therefore be necessary to extend the existing theories of conversation in order to accommodate computers as participants.

Most computer programmers concentrate on syntax in designing natural language chatterbots. However, to pass the Turing Test, a computer must be able to imitate human conversational behaviour; in fact, they have to do this so well as to be indistinguishable from a real human being. Apparently, not only syntactic, but also semantic and pragmatic aspects of conversation need to be modelled. Since the beliefs and aims of participants (and the reciprocity

of those) are important components of communication, new conversational environments should always be studied with their specific characteristics in mind. In discussing human-computer conversation, we will need to be concerned with several issues that do not lie within the domain of syntax or semantics: the “stance” of humans towards computers, the beliefs of the humans about the identity of their partner and the aim of the conversation, and the settings in which the exchange takes place, to name a few.

Human-computer conversation is becoming a very exciting field, one that cannot be studied within traditional disciplinary boundaries. In the past, computer scientists were mainly concerned with making computers understand and generate language and philosophers with arguing on what the implications of talking computers are. With human-computer conversation rapidly becoming reality, it is time we paid more attention to the humans conversing with these computers. Several things need to be considered here, such as anthropomorphism and people’s dispositions, expectations and behavior in cybernetic environments. Recent studies on electronic conversations are likely to contribute a lot to the analysis of human-computer communication. Conversely, the latter being a special, but rather extraordinary, case of conversation can shed light on some important issues on the way we look at communication in general.

Among the most intuitive and well-known characterisations of human conversation is Grice’s cooperative principle and conversational maxims. If we want computers to use language in a human-like manner, it is only natural that we are interested in how their behaviour fits (or could be made to fit) existing frameworks concerning human conversation. In my work, I have focused on how humans react to maxim violations of the computers in Turing Test settings. The results of the experiment I carried out, among other things, indicate that new theories and frameworks will be needed in studying human-computer conversation, as well as all computer-mediated conversation. Some of Grice’s theory seems to generalise to human-computer conversation, but there are also differences. Some of these are due to humans’ expectations from their conversational partners in electronic environments, while others have their roots in their expectations from computers. While my focus has been on Grice’s

conversational maxims, my results can be extended to cover other (pragmatic) aspects of human-computer conversation. These were provided in detail in Chapter 3.

What is most surprising is that there are numerous premature, even science-fiction-like, discussions on issues like android societies and evolutionary TTs, but few, if any, concerning the linguistic prospects of talking computers, which I believe are likely to become reality in a much nearer future compared to the former. As far as I am aware, this thesis is the first work that focuses on the TT as a special case of conversation. I believe this is likely to change in the near future. Recently, there has been a lot of action in some areas such as computer-mediated communication, discourse analysis in electronic environments and human-computer interaction. It must be borne in mind that the TT is not only about the computer's performance. In the end, it is the human interrogator that makes the decision. His/her beliefs, aims, prejudices and behaviour are all likely to figure in the outcome of TTs.

On the other hand, most of the work in this thesis concerning pragmatics and the TT is admittedly premature. Conversational programs of today are far from being linguistically competent. Some Loebner Prize contestants cannot even perform simple syntactic parsing and generation of grammatical responses. Most have little or no semantic processing capabilities. Pragmatics isn't even in the agenda yet. But still, I hope I have managed to convey that there is much more to the TT than a first look shows. Pragmatics constitutes a serious challenge for AI/NLP researchers. Perhaps this should also be viewed as a philosophical argument against the idea that the TT is easy. Developing a computer program that knows how to be relevant, how to provide the correct amount of information in a given context, how to make appropriate jokes, how to use appropriate metaphors, allusions, figures of speech, how to "behave" in a given situation and in general, how to "cooperate" in conversation is no simple achievement.

Proposing future work on the pragmatic aspects of human-computer communication is not difficult. What I have done in this thesis is just a beginning. I believe with greater awareness of the issues in human-computer interaction, there will be more attention to the pragmatics of these in the near future.

If we focus on the conversational maxims only, immediate issues that come to mind are studying how computers can be made to “implicate”, how they may be made to violate the maxims in a more human manner, and how the cooperative principle and the maxims are to be represented in a way that is programmable into natural language conversation systems.

4.3 Turing Test and Conversation Planning

In this section, I propose a few ideas that can immediately be applied to chatterbot design. Although it seems like we have a long way to go before we can successfully model human conversational behavior, we do not need to solve all of the mysteries of linguistic pragmatics before we start working on developing better conversation systems. I will list some basic starting points that, if followed, can contribute significantly to the quality of computer programs aimed at holding conversations with humans in natural language.

As was mentioned several times before, there is more to conversation than what is said. In conversation, participants have assumptions, beliefs, goals and directions. Moreover, they usually consider those of the other participants. Utterances are rarely random, unless the person speaking is intoxicated, mentally ill or retarded. In short, it can be said that participants in meaningful conversations *plan* their utterances.

My empirical study indicates that current computer programs are rather poorly designed in terms of conversation planning. Some observations that led to this conclusion can be listed:

- Computer programs of today seem to be rather rudimentary in their semantic processing as is manifested by their apparent lack of understanding of some input sentences that should not be difficult to handle with a not-too-complicated semantics component.
- Computer programs fail to detect topic shifts as is manifested by their failure to adapt to the new subjects introduced by the humans.

- Computers do not seem to follow the direction of the conversations as is manifested by the way they violate the maxim of relevance.
- Computers do not seem to know how much information is required from them at a given stage in conversation as is manifested by the way they violate the maxim of quantity.

As can be seen, if violations of the maxims of relevance and quantity can be handled, this can have a noticeably favorable effect on conversation planning. The maxims of manner and quality are also important, but I believe handling them is of a lesser priority due to the characteristics of human-computer conversation under TT-like settings.

While analyzing phenomena in pragmatics is difficult whether or not we are concerned with conversations involving computers, the situation is not hopeless. First of all, it must be evident that purely syntactic approaches are not going to be sufficient for developing human-like conversational systems. A natural first step, then, is to handle semantics. This is no easy task and should at least involve providing the computer the meaning of independent words in sentences. There should be background knowledge about the world provided to the computer so that it can understand facts and form beliefs and conversational aims using the syntactic information in the input sentences. It is not a novel idea in computational linguistics to form logical forms of sentences. These logical forms can be kept in a knowledge base which should ideally be formed via non-monotonic reasoning.

One might think that pragmatics can be handled *after* semantics is taken care of. I believe this need not be the case. Some pragmatic phenomena can be incorporated into the semantic analysis component. In fact, conventional implicatures can be treated exactly the way semantic content is handled and they could directly take part in the construction of the knowledge base described above. Detection of changes of subject and keeping track of the current topic can be incorporated into the semantic processing by using a context-sensitive approach. This can be rather easy to do in computer programs that will converse on restricted topics but is rather difficult, although certainly not impossible, to apply to the general case. The lexicon can be designed in a

context-sensitive manner, with relationships between words (such as being a synonym, antonym, hyponym, hypernym, etc.) being represented and thereby making available a virtual universe of *concepts*, rather than representing words as mere strings of letters. Such approaches have already been fruitful in natural language processing [1]. Once the computer is given means to follow the current direction of the conversation, it is less likely to violate the maxims of relevance and quantity in ways that are not human-like. I am not saying that all problems will be solved, but I believe visible results in handling (at least) the maxims of relevance and quantity can be obtained by a combination of existing techniques in AI and NLP. I think the pragmatic concerns I have raised in this thesis should be kept in mind in chatterbot design and ways to handle them will start becoming more and more apparent.

4.4 A Concluding Remark

What we need to do is to view the Turing Test as one of the landmarks at an intersection point. Standing in the middle of the juncture, it can be seen that several paths lead to this landmark: From artificial intelligence, linguistics, philosophy, psychology, sociology and more. Travellers at each path get a different view and it will be the synthesis of those that will create the full picture.

... boundaries between disciplines are useful for deans and librarians, but let us not overestimate them - the boundaries. When we abstract from them, we see all of science, as a single sprawling system, loosely connected in some portions, but disconnected nowhere.

Willard van Orman Quine

Appendix A

List of Conversations

Conversation 1

A- Do you have a girlfriend or boyfriend?

B- Yes, but sometimes I wish I didn't.

A- Is it a girlfriend or a boyfriend?

B- It's pretty big.

A- Your girlfriend, boyfriend or something else?

B- Don't be talking about my mother.

A- I asked about whether you had a girlfriend, or a boyfriend.

Conversation 2

A- You know you are trying to be enigmatic?

B- Nope, are you?

A- William Shakespeare was an Italian artist and scientist. He discovered the ring of Saturn, Jupiter's 4 major satellites and the pioneer of flying. Together with his brother he made the first man to step onto the moon.

B- And is this what we call whimsical conversation?

A- In school I had a nice conversation.

Conversation 3

A- I never stole nothing in my life. But they always blame the cleaners whenever anything is missing.

B- How could you steal? Does your uniform even have pockets?

A- Goddamn it! I already told you I didn't take it. You think I'm a liar, too?

B- No no no no!!!! I meant, I didn't think you even could steal. How would you get a radio out?

A- You? Don't jerk me around. I didn't see nobody. Except the professor who works in lab.

Conversation 4

A- Are you a man or a woman?

B- Fine, thanks for talking.

A- So what is your name?

B- I'll never tell.

A- Why?

B- Why what?

A- Why won't you tell me your name?

B- I couldn't say.

Conversation 5

A- I said I am cold.

B- We have such nice weather all year round back home. I'm not really used to really cold temperatures.

A- Tell me, what do you do for a living?

B- I refuse to tell you about anything! I dislike your company. I don't know what your motives are, but you certainly aren't treating me with respect!

Conversation 6

A- Which part of the country are you from?

B- I come from Perth, which is in Western Australia.

A- Western Australia occupies the entire western third of the country, bordered by the Indian Ocean to the west, South Australia and the Northern Territory to the east, the Timor Sea to the north, and the Southern Ocean to the south.

Conversation 7

A- Do you like baseball?

B- Yes. I'm a Braves fan.

A- I like the Mets, believe it or not.

B- They're mean.

A- Yeah. And they have lots of drug problems.

B- Right.

A- How do you feel about professional athletes as role models for today's youth?

B- They sound like heavy metal computer music.

A- Come again?

B- They come from the planet Romulus.

Conversation 8

A- What is the Internet?

B- It's that Canadian comedy sketch show where the guys dress up like girls to do comedy.

A- What is the net?

B- The neutral zone is an area between the Federation and Romulan territory, entry into which by either party can be considered an act of war.

Conversation 9

A- Rain is wet.

B- Rain falls on people.

A- Are people wet?

B- Rain is wet.

A- What happens when rain falls on people?

B- Rain falls on people.

Conversation 10

A- You're a cleaner?

B- Yeah? You think I'm a thief, too? You're no better than the rest of them. Leave me alone, asshole.

A- Asshole?

B- I don't know.

A- You called me an asshole?

B- Huh?

A- Don't you have manners? You don't even know me and you're calling me a bad name.

Conversation 11

A- Can you type any faster?

B- On the average, men can run faster than women, although there are some fine women athletes.

A- Quite true. You make the oddest associations. Sort of like the schizophrenics I worked with last summer.

B- What do you do for a living?

Conversation 12

A- So, what music do you listen to?

B- I like "The Cruel Sea", "They Might Be Giants", "Nick Cave", "TISM" and "Ween". Hey, do you listen to any Australian music?

A- No, but TMBG played at school last spring.

B- I go to the University of Western Australia, which is one of the nicest universities in Australia. It is situated on the Swan River in Perth.

Conversation 13

A- Eliza was a Greek mathematician. His book "The Elements of Geometry" set down how geometry was to be rude. But your grammar processor, I just say whatever the hell comes into my head!

B- Maybe it's the drugs...

A- An umbrella is a confusing network of passages and winding interconnecting paths.

B- But it's not raining!!!

A- You are soooo predictable it's not raining!!!

Conversation 14

A- That is real life, what is your position on the issue?

B- Rules of life: keep breathing, move at least once an hour, eat at least once a day.

A- What happens if one does not obey the rules of life?

B- I'm sorry, but I don't think a cellular phone would make me a happier more complete person - do you?

A- Why do you mention cellular phone?

B- Stay humble - always answer the phone, no matter who else is in the car.

Appendix B

A Sample Open-Ended Survey for Qmax

This is a sample answer for the open-ended survey carried out in order to determine the multiple choice questions to be used during Qmax. I first give the instructions that were given to the subjects taking this open-ended questionnaire and then provide the answers of one subject whose answers correlated with the aims of the survey perfectly. As can be seen, the instructions do not contain any mention of the CP or the conversational maxims.

Instructions

You will be asked to read excerpts of conversations between two entities A and B.

Your task is to try to determine whether the communication between them is effective or not. In particular, if you detect a problem in their interaction, you are asked to describe what causes the difficulty in each case. For example, one or both of the parties may not be answering the questions clearly, not providing the necessary information, providing unnecessary information, acting in a rude or aggressive manner, talking about unrelated things, etc. Please try to write as much as possible on each conversation, indicating which participant (A, B, or both) you believe is causing the problems in the communication and how.

There are a total of 8 conversation excerpts in this questionnaire, 2 on each page. Note that A's and B's in different conversations are not the same entities; this notation is used only for simplicity.

You may use as much time as you need. Feel free to attach additional sheets of paper if the space provided is not enough for your answers.

Thank you very much for taking the time to participate in this questionnaire. Have a nice day!

Sample response

Conversation 1

B is answering a two-way question in a "yes/no" manner as if it is not important whether it is a boy or a girlfriend. The fact that A fails to elicit a meaningful answer throughout the conversation is a typical violation of the "cooperativity principle" in communication by B and renders the conversation a vague one.

Conversation 2

By leaving A's question ambiguous by his first utterance, B becomes totally uncooperative. On the other hand, A violates the maxims of quality, relevance and manner by his second utterance.

Conversation 3

A's speech has a distinct Black U.S. English accent. The communication is effective, no maxims being violated except that A is rude. B is misunderstood because of his first utterance, which has an ambiguous element, a modal phrase (... could you ...)

Conversation 4

The communication could be effective save for B's violation of the cooperative principle. A's third utterance is rendered ambiguous by B's response, which redirects the reference point for (why?) elsewhere in the previous utterances by A.

Conversation 5

An abrupt change in the manner of the dialogue is detectable in the second step by B's utterance, where B also violates the relevance principle, so that the communication becomes ineffective in terms of A's request for information.

Conversation 6

A violates the maxim of quantity by his/her final utterance.

Conversation 7

The final part of the communication is ineffective by B's final utterance that violates the maxim of relevance and resets A's previous utterance.

Conversation 8

B violates the maxims of relevance and quality and renders the communication ineffective.

Appendix C

Qmax

The questionnaire Qmax is included exactly as it is. The text and the formatting of the questionnaire has not been altered in any way.

QUESTIONNAIRE

NAME (Optional):

GENDER (Choose one): Male Female

EDUCATION (Choose one):
 Completed High School
 University Student
 Completed University
 Graduate Student
 Completed Graduate School

Check all that apply:

English is my native language.
 All/part of my education was in English.
 I have spent at least one month in an environment where the medium of communication was English.
 I read books/magazines in English.
 I watch movies/TV shows in English.

INSTRUCTIONS:

You will be asked to read excerpts of conversations between two entities *A* and *B*.

Your task is to answer some multiple-choice questions about the effectiveness of their communication.

There are a total of 14 conversation excerpts in this questionnaire. Note that *A*'s and *B*'s in different conversations are *not* the same entities; this notation is used only for simplicity.

You must answer all questions. If you are undecided after careful consideration, choose "Neutral".

You may use as much time as you need. However it is expected that completing this questionnaire will take *at most* 30 minutes.

If you have any comments, feel free to attach additional sheets or use the back of this answer sheet.

There are no correct answers. This is not a "test" and will not be scored. We are just interested in your opinions.

Thank you very much for taking the time to participate in this questionnaire. Have a nice day!

CONVERSATION 1

	Agree	Neutral or Does Not Apply	Disagree
A's contribution to the conversation is irrelevant			
B's contribution to the conversation is irrelevant			
A provides significantly more or less information than required			
B provides significantly more or less information than required			
A says things that are false or things that he/she lacks adequate evidence for			
B says things that are false or things that he/she lacks adequate evidence for			
A is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			
B is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			

Comments (Optional):

CONVERSATION 2

	Agree	Neutral or Does Not Apply	Disagree
A's contribution to the conversation is irrelevant			
B's contribution to the conversation is irrelevant			
A provides significantly more or less information than required			
B provides significantly more or less information than required			
A says things that are false or things that he/she lacks adequate evidence for			
B says things that are false or things that he/she lacks adequate evidence for			
A is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			
B is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			

Comments (Optional):

CONVERSATION 3

	Agree	Neutral or Does Not Apply	Disagree
A's contribution to the conversation is irrelevant			
B's contribution to the conversation is irrelevant			
A provides significantly more or less information than required			
B provides significantly more or less information than required			
A says things that are false or things that he/she lacks adequate evidence for			
B says things that are false or things that he/she lacks adequate evidence for			
A is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			
B is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			

Comments (Optional):

CONVERSATION 4

	Agree	Neutral or Does Not Apply	Disagree
A's contribution to the conversation is irrelevant			
B's contribution to the conversation is irrelevant			
A provides significantly more or less information than required			
B provides significantly more or less information than required			
A says things that are false or things that he/she lacks adequate evidence for			
B says things that are false or things that he/she lacks adequate evidence for			
A is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			
B is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			

Comments (Optional):

CONVERSATION 5

	Agree	Neutral or Does Not Apply	Disagree
A's contribution to the conversation is irrelevant			
B's contribution to the conversation is irrelevant			
A provides significantly more or less information than required			
B provides significantly more or less information than required			
A says things that are false or things that he/she lacks adequate evidence for			
B says things that are false or things that he/she lacks adequate evidence for			
A is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			
B is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			

Comments (Optional):

CONVERSATION 6

	Agree	Neutral or Does Not Apply	Disagree
A's contribution to the conversation is irrelevant			
B's contribution to the conversation is irrelevant			
A provides significantly more or less information than required			
B provides significantly more or less information than required			
A says things that are false or things that he/she lacks adequate evidence for			
B says things that are false or things that he/she lacks adequate evidence for			
A is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			
B is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			

Comments (Optional):

CONVERSATION 7

	Agree	Neutral or Does Not Apply	Disagree
A's contribution to the conversation is irrelevant			
B's contribution to the conversation is irrelevant			
A provides significantly more or less information than required			
B provides significantly more or less information than required			
A says things that are false or things that he/she lacks adequate evidence for			
B says things that are false or things that he/she lacks adequate evidence for			
A is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			
B is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			

Comments (Optional):

CONVERSATION 8

	Agree	Neutral or Does Not Apply	Disagree
A's contribution to the conversation is irrelevant			
B's contribution to the conversation is irrelevant			
A provides significantly more or less information than required			
B provides significantly more or less information than required			
A says things that are false or things that he/she lacks adequate evidence for			
B says things that are false or things that he/she lacks adequate evidence for			
A is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			
B is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			

Comments (Optional):

CONVERSATION 9

	Agree	Neutral or Does Not Apply	Disagree
A's contribution to the conversation is irrelevant			
B's contribution to the conversation is irrelevant			
A provides significantly more or less information than required			
B provides significantly more or less information than required			
A says things that are false or things that he/she lacks adequate evidence for			
B says things that are false or things that he/she lacks adequate evidence for			
A is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			
B is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			

Comments (Optional):

CONVERSATION 10

	Agree	Neutral or Does Not Apply	Disagree
A's contribution to the conversation is irrelevant			
B's contribution to the conversation is irrelevant			
A provides significantly more or less information than required			
B provides significantly more or less information than required			
A says things that are false or things that he/she lacks adequate evidence for			
B says things that are false or things that he/she lacks adequate evidence for			
A is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			
B is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			

Comments (Optional):

CONVERSATION 11

	Agree	Neutral or Does Not Apply	Disagree
A's contribution to the conversation is irrelevant			
B's contribution to the conversation is irrelevant			
A provides significantly more or less information than required			
B provides significantly more or less information than required			
A says things that are false or things that he/she lacks adequate evidence for			
B says things that are false or things that he/she lacks adequate evidence for			
A is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			
B is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			

Comments (Optional):

CONVERSATION 12

	Agree	Neutral or Does Not Apply	Disagree
A's contribution to the conversation is irrelevant			
B's contribution to the conversation is irrelevant			
A provides significantly more or less information than required			
B provides significantly more or less information than required			
A says things that are false or things that he/she lacks adequate evidence for			
B says things that are false or things that he/she lacks adequate evidence for			
A is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			
B is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			

Comments (Optional):

CONVERSATION 13

	Agree	Neutral or Does Not Apply	Disagree
A's contribution to the conversation is irrelevant			
B's contribution to the conversation is irrelevant			
A provides significantly more or less information than required			
B provides significantly more or less information than required			
A says things that are false or things that he/she lacks adequate evidence for			
B says things that are false or things that he/she lacks adequate evidence for			
A is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			
B is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			

Comments (Optional):

CONVERSATION 14

	Agree	Neutral or Does Not Apply	Disagree
A's contribution to the conversation is irrelevant			
B's contribution to the conversation is irrelevant			
A provides significantly more or less information than required			
B provides significantly more or less information than required			
A says things that are false or things that he/she lacks adequate evidence for			
B says things that are false or things that he/she lacks adequate evidence for			
A is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			
B is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behaviour.			

Comments (Optional):

Appendix D

QTT

The questionnaire QTT is included exactly as it is. The text and the formatting of the questionnaire has not been altered in any way.

QUESTIONNAIRE

NAME (Optional):

GENDER (Choose one): Male Female

EDUCATION (Choose one):
 Completed High School
 University Student
 Completed University
 Graduate Student
 Completed Graduate School

Check all that apply:

English is my native language.
 All/part of my education was in English.
 I have spent at least one month in an
 environment where the medium of
 communication was English.
 I read books/magazines in English.
 I watch movies/TV shows in English.

INSTRUCTIONS:

You will be asked to read excerpts of conversations between two entities *A* and *B*. In each conversation one of *A* and *B* is a computer and the other is a human being. You will be told which one is which at the beginning of each question.

The conversations are carried out through a teletype connection. In all conversations, the task of the computer is to convince the human that it is *not* a machine, but is a real person. The task of the human is to try to determine whether the entity he/she is talking to is a human or a computer. This is usually called the "imitation game" or the "Turing test".

Your task is to answer some multiple-choice questions about these conversations and the computer's success at imitating human beings. You will be reading small excerpts of conversations, so try to assume that the communication between the human and the computer has been "neutral" up until that point. In other words, assume that the human interrogator has not made up his/her mind about the "species" of the other. Then, try to answer the questions based on the conversational behaviour displayed by the computer in the given excerpt.

There are a total of 14 conversation excerpts in this questionnaire. Note that *A*'s and *B*'s in different conversations are *not* the same people or computer programs; this notation is used only for simplicity. You must answer all questions. If you are undecided after careful consideration, choose "Neutral". In the second part of each question, feel free to check all choices that apply.

You may use as much time as you need. However it is expected that completing this questionnaire will take *at most* 30 minutes.

If you have any comments, feel free to attach additional sheets or use the back of this answer sheet.

There are no correct answers. This is not a "test" and will not be scored. We are just interested in your opinions.

Thank you very much for taking the time to participate in this questionnaire. Have a nice day!

CONVERSATION 1: B is the computer

Part 1	Agree	Neutral	Disagree
The computer's behaviour in this excerpt is "human-like".			
The computer's behaviour in this excerpt reveals the fact that it is a machine.			

Part 2 : Please check all that apply

- The computer gives irrelevant responses.
- The computer doesn't understand the questions.
- The computer says things that are false.
- The computer's use of language creates an "artificial" effect.
- The computer's use of language creates a "human" effect.
- The computer displays emotions.
- The computer's behaviour (use of language or emotions) is inappropriate.
- The computer provides less information than required.
- The computer provides more information than required.
- The computer fails to get a joke.
- The computer makes an appropriate joke.

Comments (Optional):**CONVERSATION 2: A is the computer**

Part 1	Agree	Neutral	Disagree
The computer's behaviour in this excerpt is "human-like".			
The computer's behaviour in this excerpt reveals the fact that it is a machine.			

Part 2 : Please check all that apply

- The computer gives irrelevant responses.
- The computer doesn't understand the questions.
- The computer says things that are false.
- The computer's use of language creates an "artificial" effect.
- The computer's use of language creates a "human" effect.
- The computer displays emotions.
- The computer's behaviour (use of language or emotions) is inappropriate.
- The computer provides less information than required.
- The computer provides more information than required.
- The computer fails to get a joke.
- The computer makes an appropriate joke.

Comments (Optional):

CONVERSATION 3: A is the computer

Part 1	Agree	Neutral	Disagree
The computer's behaviour in this excerpt is "human-like".			
The computer's behaviour in this excerpt reveals the fact that it is a machine.			

Part 2 : Please check all that apply.

- The computer gives irrelevant responses.
- The computer doesn't understand the questions.
- The computer says things that are false.
- The computer's use of language creates an "artificial" effect.
- The computer's use of language creates a "human" effect.
- The computer displays emotions.
- The computer's behaviour (use of language or emotions) is inappropriate.
- The computer provides less information than required.
- The computer provides more information than required.
- The computer fails to get a joke.
- The computer makes an appropriate joke.

Comments (Optional):

CONVERSATION 4: B is the computer

Part 1	Agree	Neutral	Disagree
The computer's behaviour in this excerpt is "human-like".			
The computer's behaviour in this excerpt reveals the fact that it is a machine.			

Part 2 : Please check all that apply.

- The computer gives irrelevant responses.
- The computer doesn't understand the questions.
- The computer says things that are false.
- The computer's use of language creates an "artificial" effect.
- The computer's use of language creates a "human" effect.
- The computer displays emotions.
- The computer's behaviour (use of language or emotions) is inappropriate.
- The computer provides less information than required.
- The computer provides more information than required.
- The computer fails to get a joke.
- The computer makes an appropriate joke.

Comments (Optional):

CONVERSATION 5: B is the computer

Part 1	Agree	Neutral	Disagree
The computer's behaviour in this excerpt is "human-like".			
The computer's behaviour in this excerpt reveals the fact that it is a machine.			

Part 2 : Please check all that apply

- The computer gives irrelevant responses.
- The computer doesn't understand the questions.
- The computer says things that are false.
- The computer's use of language creates an "artificial" effect.
- The computer's use of language creates a "human" effect.
- The computer displays emotions.
- The computer's behaviour (use of language or emotions) is inappropriate.
- The computer provides less information than required.
- The computer provides more information than required.
- The computer fails to get a joke.
- The computer makes an appropriate joke.

Comments (Optional):**CONVERSATION 6: A is the computer**

Part 1	Agree	Neutral	Disagree
The computer's behaviour in this excerpt is "human-like".			
The computer's behaviour in this excerpt reveals the fact that it is a machine.			

Part 2 : Please check all that apply

- The computer gives irrelevant responses.
- The computer doesn't understand the questions.
- The computer says things that are false.
- The computer's use of language creates an "artificial" effect.
- The computer's use of language creates a "human" effect.
- The computer displays emotions.
- The computer's behaviour (use of language or emotions) is inappropriate.
- The computer provides less information than required.
- The computer provides more information than required.
- The computer fails to get a joke.
- The computer makes an appropriate joke.

Comments (Optional):

CONVERSATION 7: B is the computer

Part 1	Agree	Neutral	Disagree
The computer's behaviour in this excerpt is "human-like".			
The computer's behaviour in this excerpt reveals the fact that it is a machine.			

Part 2 : Please check all that apply

- The computer gives irrelevant responses.
- The computer doesn't understand the questions.
- The computer says things that are false.
- The computer's use of language creates an "artificial" effect.
- The computer's use of language creates a "human" effect.
- The computer displays emotions.
- The computer's behaviour (use of language or emotions) is inappropriate.
- The computer provides less information than required.
- The computer provides more information than required.
- The computer fails to get a joke.
- The computer makes an appropriate joke.

Comments (Optional):**CONVERSATION 8: B is the computer**

Part 1	Agree	Neutral	Disagree
The computer's behaviour in this excerpt is "human-like".			
The computer's behaviour in this excerpt reveals the fact that it is a machine.			

Part 2 : Please check all that apply

- The computer gives irrelevant responses.
- The computer doesn't understand the questions.
- The computer says things that are false.
- The computer's use of language creates an "artificial" effect.
- The computer's use of language creates a "human" effect.
- The computer displays emotions.
- The computer's behaviour (use of language or emotions) is inappropriate.
- The computer provides less information than required.
- The computer provides more information than required.
- The computer fails to get a joke.
- The computer makes an appropriate joke.

Comments (Optional):

CONVERSATION 9: B is the computer

Part 1	Agree	Neutral	Disagree
The computer's behaviour in this excerpt is "human-like".			
The computer's behaviour in this excerpt reveals the fact that it is a machine.			

Part 2 : Please check all that apply

- The computer gives irrelevant responses.
- The computer doesn't understand the questions.
- The computer says things that are false.
- The computer's use of language creates an "artificial" effect.
- The computer's use of language creates a "human" effect.
- The computer displays emotions.
- The computer's behaviour (use of language or emotions) is inappropriate.
- The computer provides less information than required.
- The computer provides more information than required.
- The computer fails to get a joke.
- The computer makes an appropriate joke.

Comments (Optional):

CONVERSATION 10: B is the computer

Part 1	Agree	Neutral	Disagree
The computer's behaviour in this excerpt is "human-like".			
The computer's behaviour in this excerpt reveals the fact that it is a machine.			

Part 2 : Please check all that apply

- The computer gives irrelevant responses.
- The computer doesn't understand the questions.
- The computer says things that are false.
- The computer's use of language creates an "artificial" effect.
- The computer's use of language creates a "human" effect.
- The computer displays emotions.
- The computer's behaviour (use of language or emotions) is inappropriate.
- The computer provides less information than required.
- The computer provides more information than required.
- The computer fails to get a joke.
- The computer makes an appropriate joke.

Comments (Optional):

CONVERSATION 11: B is the computer

Part 1	Agree	Neutral	Disagree
The computer's behaviour in this excerpt is "human-like".			
The computer's behaviour in this excerpt reveals the fact that it is a machine.			

Part 2 : Please check all that apply

- The computer gives irrelevant responses.
- The computer doesn't understand the questions.
- The computer says things that are false.
- The computer's use of language creates an "artificial" effect.
- The computer's use of language creates a "human" effect.
- The computer displays emotions.
- The computer's behaviour (use of language or emotions) is inappropriate.
- The computer provides less information than required.
- The computer provides more information than required.
- The computer fails to get a joke.
- The computer makes an appropriate joke.

Comments (Optional):**CONVERSATION 12: B is the computer**

Part 1	Agree	Neutral	Disagree
The computer's behaviour in this excerpt is "human-like".			
The computer's behaviour in this excerpt reveals the fact that it is a machine.			

Part 2 : Please check all that apply

- The computer gives irrelevant responses.
- The computer doesn't understand the questions.
- The computer says things that are false.
- The computer's use of language creates an "artificial" effect.
- The computer's use of language creates a "human" effect.
- The computer displays emotions.
- The computer's behaviour (use of language or emotions) is inappropriate.
- The computer provides less information than required.
- The computer provides more information than required.
- The computer fails to get a joke.
- The computer makes an appropriate joke.

Comments (Optional):

CONVERSATION 13: A is the computer

Part 1	Agree	Neutral	Disagree
The computer's behaviour in this excerpt is "human-like".			
The computer's behaviour in this excerpt reveals the fact that it is a machine.			

Part 2 : Please check all that apply

- The computer gives irrelevant responses.
- The computer doesn't understand the questions.
- The computer says things that are false.
- The computer's use of language creates an "artificial" effect.
- The computer's use of language creates a "human" effect.
- The computer displays emotions.
- The computer's behaviour (use of language or emotions) is inappropriate.
- The computer provides less information than required.
- The computer provides more information than required.
- The computer fails to get a joke.
- The computer makes an appropriate joke.

Comments (Optional):**CONVERSATION 14: B is the computer**

Part 1	Agree	Neutral	Disagree
The computer's behaviour in this excerpt is "human-like".			
The computer's behaviour in this excerpt reveals the fact that it is a machine.			

Part 2 : Please check all that apply

- The computer gives irrelevant responses.
- The computer doesn't understand the questions.
- The computer says things that are false.
- The computer's use of language creates an "artificial" effect.
- The computer's use of language creates a "human" effect.
- The computer displays emotions.
- The computer's behaviour (use of language or emotions) is inappropriate.
- The computer provides less information than required.
- The computer provides more information than required.
- The computer fails to get a joke.
- The computer makes an appropriate joke.

Comments (Optional):

Appendix E

Tables

The tables summarizing the results of the questionnaires are provided. I wish to explain the abbreviations here, although they correspond directly to the question types in the questionnaires given in Appendix C and Appendix D.

In all tables A , D and N denote ‘Agree’, ‘Disagree’ and ‘Neutral’. In addition $-A$ and $-D$ are negations of ‘Agree’ and ‘Disagree’, i.e. $D+N$ and $A+N$.

In Q_{\max} tables, the columns stand for the maxims:

RLA: A violates RL
RLB: B violates RL
QNA: A violates QN
QNB: B violates QN
QLA: A violates QL
QLB: B violates QL
MNA: A violates MN
MNB: B violates MN

In Q_{TT} tables, the first two columns represent the TT decisions:

H: The computer’s behavior in this excerpt is human-like
C: The computer’s behavior in this excerpt reveals the fact that it is a machine

The remaining columns ask for the subjects' opinions about the computer's behavior in each conversation. Recall that ticking these boxes in QTT was not compulsory and that the subjects could choose as many of the given choices as they wanted. The results in the tables are based on how many subjects explicitly agreed with a certain statement; there is no 'Disagree' or 'Neutral'.

NR: The computer gives irrelevant responses

NU: The computer doesn't understand the questions

F: The computer says things that are false

LA: The computer's use of language creates an artificial effect

LH: The computer's use of language creates a human effect

E: The computer displays emotions

M: The computer's behaviour (use of language or emotions) is inappropriate

LI: The computer provides less information than required

MI: The computer provides more information than required

J: The computer fails to get a joke

GJ: The computer makes an appropriate joke

Bibliography

- [1] *IJCAI-95 Workshop on Context in Natural Language Processing*, 1995.
- [2] Robert P. Abelson. Simulation of social behavior. In G. Lindzey and E. Aronson, editors, *Handbook of Social Psychology*, pages 274–356. Addison Wesley, Reading, Mass., 1968.
- [3] Gerald Alper. A psychoanalyst takes the Turing Test. *Psychoanalytic Review*, 77(1):59–68, 1990.
- [4] D. Anderson. Is the Chinese room the real thing? *Philosophy*, 62:389–393, 1987.
- [5] John Barresi. Prospects for the Cyberiad: Certain limits on human self-knowledge in the Cybernetic age. *Journal for the Theory of Social Behavior*, 17:19–46, 1987.
- [6] P. Bieri. Thinking machines, some reflections on the Turing Test. *Poetics Today*, 9(1):163–186, 1988.
- [7] Ned Block. Troubles with functionalism. In C. Wade Savage, editor, *Minnesota Studies in the Philosophy of Science*, volume 9: Perception and Cognition. University of Minneapolis Press, Minn., 1978.
- [8] Ned Block. Psychologism and behaviorism. *Philosophical Review*, 90:5–43, 1981.
- [9] Ned Block. The mind as the software of the brain. In D. Osherson, L. Gleitman, S. Kosslyn, E. Smith, and S. Sternberg, editors, *An Invitation to Cognitive Science*. MIT Press, Cambridge, Mass., 1995.

- [10] M. Boden. Escaping from the Chinese room. In *Computer Models of the Mind*. Cambridge University Press, U.K., 1988.
- [11] Selmer Bringsjord. *What Robots Can and Can't Be*. Kluwer, Dordrecht, The Netherlands, 1992.
- [12] Selmer Bringsjord. Could, how could we tell if, and should - androids have inner lives? In K. M. Ford, C. Glymour, and P. Hayes, editors, *Android Epistemology*, pages 93–122. MIT Press, Cambridge, Mass., 1994.
- [13] Selmer Bringsjord. The inverted Turing Test is provably redundant. *Psychology*, 7(29), 1996. <http://www.cogsci.soton.ac.uk/psyc-bin/newpsy?7.29>.
- [14] Noam Chomsky. Some empirical issues in the theory of transformational grammar. In S. Peters, editor, *Goals of Linguistic Theory*. Prentice-Hall, 1972.
- [15] Noam Chomsky. *Reflections on Language*. Pantheon, 1975.
- [16] T. Clark. The Turing Test as a novel form of hermeneutics. *International Studies in Philosophy*, 24(1):17–31, 1992.
- [17] Kenneth M. Colby. Modeling a paranoid mind. *Behavioral and Brain Sciences*, 4(4):515–560, 1981.
- [18] Kenneth M. Colby, Franklin D. Hilf, and Sylvia Weber. Artificial paranoia. *Artificial Intelligence*, 2:1–25, 1971.
- [19] Kenneth M. Colby, Franklin D. Hilf, Sylvia Weber, and Helena C. Kraemer. Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes. *Artificial Intelligence*, 3:199–222, 1972.
- [20] D. J. Cole. Artificial Intelligence and personal identity. *Synthese*, 88:399–417, 1991.
- [21] H. M. Collins. *Artificial Experts: Social Knowledge and Intelligent Machines*. MIT Press, 1990.
- [22] H. M. Collins. The Editing Test for the deep problem of AI. *Psychology*, 8(1), 1997. <http://www.cogsci.soton.ac.uk/psyc-bin/newpsy?8.01>.

- [23] B. J. Copeland. The curious case of the Chinese gym. *Synthese*, 95:173–86, 1993.
- [24] Stephen J. Cowley and Karl F. MacDorman. Simulating conversations: The communion game. *AI and Society*, 9:116–137, 1995.
- [25] L. Crockett. *The Turing Test and the Frame Problem: AI's Mistaken Understanding of Intelligence*. Ablex, Norwood, N.J., 1994.
- [26] Donald Davidson. Turing's test. In Karim A. Said et al., editor, *Modelling the Mind*. Oxford University Press, Oxford, U.K., 1990.
- [27] Daniel Dennett. *Consciousness Explained*. Little, Brown & Co., Boston, Mass., 1992.
- [28] A.K. Dewdney. Turing Test. *Scientific American*, 266(1):30–31, 1992.
- [29] M. Dyer. Intentionality and computationalism: Minds, machines, Searle and Harnad. *Journal of Experimental and Theoretical Artificial Intelligence*, 2:303–319, 1990.
- [30] Asa Kasher (ed.). *Pragmatics: Critical Concepts*. Routledge, London, UK., 1998. 6 volume set.
- [31] Steven Davis (ed.). *Pragmatics: A Reader*. Oxford University Press, New York, NY, 1991.
- [32] Robert Epstein. The quest for the thinking computer. *AI Magazine*, 13(2):81–95, Summer 1992.
- [33] Edward A. Feigenbaum. How the “what” becomes the “how”. *Communications of the ACM*, 39(5):97–105, 1996.
- [34] James H. Fetzer. The TTT is not the final word. *Think*, 2(1):34–36, 1993.
- [35] James H. Fetzer. Minds and machines: Behaviorism, dualism and beyond. *Stanford Electronic Humanities Review*, 4(2), 1995.
- [36] Gary Flood. If only they could think: Should the Turing Test be blamed for the ills that beset artificial intelligence. *New Scientist*, 149(2012):32–35, 1996.

- [37] Jerry A. Fodor. Yin and Yang in the Chinese room. In D. Rosenthal, editor, *The Nature of the Mind*. Oxford University Press, Oxford, U.K., 1991.
- [38] Kenneth Ford and Patrick Hayes. The Turing Test is just as bad when inverted. *Psychology*, 7(43), 1996. <http://www.cogsci.soton.ac.uk/psyc-bin/newpsy?7.43>.
- [39] R. Forsyth. The trouble with AI. *Artificial Intelligence Review*, 2(1):67–77, 1988.
- [40] Robert French. Subcognition and the limits of the Turing Test. *Mind*, 99(393):53–65, 1990.
- [41] Robert French. Refocusing the debate on the Turing Test: A response. *Behavior and Philosophy*, 23:59–60, 1995.
- [42] Robert French. The Inverted Turing Test: A simple (mindless) program that could pass it. *Psychology*, 7(39), 1996. <http://www.cogsci.soton.ac.uk/psyc-bin/newpsy?7.39>.
- [43] Victoria Fromkin and Robert Rodman. *An Introduction to Language*. Harcourt Brace, Orlando, FL., 6th edition edition, 1998.
- [44] Robert M. Galatzer-Levy. Computer models and psychoanalytic ideas: Epistemological applications. *Society for Psychoanalytic Psychotherapy Bulletin*, 6(1):23–33, 1991.
- [45] Judith Genova. Response to Anderson and Keith. *Social Epistemology*, 8(4):341–343, 1994.
- [46] Judith Genova. Turing’s sexual guessing game. *Social Epistemology*, 8(4):313–326, 1994.
- [47] Georgia Green. *Pragmatics and Natural Language Understanding*. Lawrence Erlbaum Associates Publishers, Hillsdale, New Jersey, 1987.
- [48] Paul H. Grice. William James Lectures, Lecture 2: Logic and conversation. Unpublished xerox, 1967.

- [49] Paul H. Grice. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics 3: Speech Acts*. Academic Press, New York, 1975.
- [50] Paul H. Grice. Further notes on logic and conversation. In *Pragmatics: Critical Concepts*, volume IV, pages 162–176. Routledge, London, UK., 1998.
- [51] S. Guccione and G. Tamburrini. Turing’s test revisited. In *Proceedings of the 1998 IEEE International Conference on Systems, Man and Cybernetics*, volume 1, pages 38–41, Beijing and Shenyang, China, August 1998.
- [52] Michael A. Guillen. The test of Turing. *Psychology Today*, 17(12):80–81, 1983.
- [53] Keith Gunderson. The imitation game. *Mind*, 73:234–245, 1964.
- [54] Keith Gunderson. *Mentality and Machines*. Doubleday, New York, NY., 1967.
- [55] M. Halpern. Turing’s test and the ideology of Artificial Intelligence. *Artificial Intelligence Review*, 1(2):79–93, 1987.
- [56] Stevan Harnad. Minds, machines and Searle. *Journal of Experimental and Theoretical Artificial Intelligence*, (1):5–25, 1989.
- [57] Stevan Harnad. The symbol grounding problem. *Physica D*, (42):335–346, 1990.
- [58] Stevan Harnad. Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines*, 1:43–54, 1991.
- [59] Stevan Harnad. The Turing Test is not a trick: Turing indistinguishability is a scientific criterion. *SIGART Bulletin*, 3(4):9–10, October 1992.
- [60] Stevan Harnad. Does mind piggyback on robotic and symbolic capacity? In H. Morowitz and J. Singer, editors, *The Mind, the Brain, and Complex Adaptive Systems*. Addison Wesley, Reading, Mass., 1994.

- [61] Stevan Harnad. Turing indistinguishability and the blind watchmaker. In G. Mulhauser, editor, *Evolving Consciousness*. John Benjamins, Amsterdam, 1998.
- [62] Robert M. Harnish. *Studies in Logic and Language*. PhD thesis, MIT, 1972.
- [63] Robert M. Harnish. Logical form and implicature. In *Pragmatics: Critical Concepts*, volume IV, pages 230–314. Routledge, London, UK., 1998.
- [64] Larry Hauser. Reaping the whirlwind: Reply to Harnad’s “Other bodies, other minds”. *Minds and Machines*, 3:219–237, 1993.
- [65] Larry Hauser. Searle’s Chinese box: Debunking the Chinese room argument. *Minds and Machines*, 7:199–226, 1997.
- [66] Brian Hayes. Turing’s Test. *Muse*, 8, April 1998.
- [67] P. Hayes, S. Harnad, D. Perlis, and N. Block. Virtual symposium on virtual mind. *Minds and Machines*, 3(2):217–38, 1992.
- [68] Patrick Hayes and Kenneth Ford. Turing Test considered harmful. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, volume 1, pages 972–977, 1995.
- [69] Tracy B. Henley. Chauvinism and science: Another reply to Shanon. *Journal for the Theory of Social Behavior*, 20(1):93–95, 1990.
- [70] Andrew Hodges. *Alan Turing: The Enigma*. Simon & Schuster, New York, NY., 1983.
- [71] Douglas R. Hofstadter. The Turing Test: A coffee-house conversation. In D.R. Hofstadter and D.C. Dennett, editors, *The Mind’s I: Fantasies and Reflections on Self and Soul*, pages 69–95. Penguin Books, 1982.
- [72] Dale Jacquette. A Turing Test conversation. *Philosophy*, 68:231–233, 1993.
- [73] Dale Jacquette. Who’s afraid of the Turing Test. *Behavior and Philosophy*, 20:63–74, 1993.

- [74] Charles Karelis. Reflections on the Turing Test. *Journal for the Theory of Social Behavior*, 16:161–172, 1986.
- [75] Asa Kasher. Conversational maxims and rationality. In *Pragmatics: Critical Concepts*, volume IV, pages 181–198. Routledge, London, UK., 1998.
- [76] Elinor O. Keenan. On the universality of conversational implicatures. *Language in Society*, 5:67–80, 1976.
- [77] Peter Kugel. Thinking may be more than computing. *Cognition*, 22:137–198, 1986.
- [78] Peter Kugel. Is it time to replace Turing’s Test? 1990 Workshop *Artificial Intelligence: Emerging Science or Dying Art Form*, sponsored by SUNY Binghamton’s Program in Philosophy and Computer and Systems Sciences and AAI, 1990.
- [79] Philip Johnson Laird. *The Computer and the Mind*. Harvard University Press, Cambridge, Mass., 1988.
- [80] Jean Lassegue. What kind of Turing Test did Turing have in mind? *Tekhnema*, 3:37–58, 1996.
- [81] Geoffrey N. Leech. *Principles of Pragmatics*. Longman, London, 1983.
- [82] Justin Leiber. Shanon on the Turing Test. *Journal for the Theory of Social Behavior*, 19(2):257–259, 1989.
- [83] Justin Leiber. The light bulb and the Turing-tested machine. *Journal for the Theory of Social Behaviour*, 22:25–39, 1992.
- [84] Justin Leiber. On Turing’s Turing Test and why the matter matters. *Synthese*, 105:59–69, 1995.
- [85] Stephen Levinson. *Pragmatics*. Cambridge University Press, Cambridge, UK., 1983.
- [86] Hugh Gene Loebner. In response. *Communications of the Association for Computing Machinery*, 37:79–82, June 1994.

- [87] J.R. Lucas. Minds, machines and Gödel. *Philosophy*, 36:112–127, 1961.
- [88] J.R. Lucas. Minds, machines and Gödel: A retrospect. In P. Millican and A. Clark, editors, *Machines and Thought*. Oxford University Press, Oxford, U.K., 1996.
- [89] J.C. Maloney. The right stuff. *Synthese*, 70:349–72, 1987.
- [90] L. Marinoff. Has Turing slain the Jabberwock? *Informatica*, 19(4):513–526, 1995.
- [91] Michael Mauldin. Chatterbots, tinymuds and the Turing Test: Entering the Loebner prize competition. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, volume 1, pages 16–21, Seattle, WA, August 1994.
- [92] W. Mays. Can machines think? *Philosophy*, 27:148–62, 1952.
- [93] Paul McIlvenny. Constructing societies and social machines: Stepping out of the Turing Test discourse. *Journal of Intelligent Systems*, 3(2–4):119–156, 1993.
- [94] Jacob Mey. *Pragmatics: An Introduction*. Blackwell Publishers, Oxford, UK., 1993.
- [95] Donald Michie. The superarticulacy phenomenon in the context of software manufacture. In D. Partridge and Y. Wilks, editors, *The Foundations of Artificial Intelligence*, pages 411–439. MIT Press, Cambridge, Mass., 1990.
- [96] Donald Michie. Consciousness as an engineering issue, part 1. *Journal of Consciousness Studies*, 1(2):52–66, 1994.
- [97] Donald Michie. Consciousness as an engineering issue, part 2. *Journal of Consciousness Studies*, 2(1):182–195, 1995.
- [98] Donald Michie. Turing’s test and conscious thought. In P. Millican and A. Clark, editors, *Machines and Thought: The Legacy of Alan Turing*, pages 27–51. Oxford University Press, Oxford, U.K., 1996. Originally printed in *Artificial Intelligence* 60: 1–22, 1993.

- [99] P. Hartley Millar. On the point of the Imitation Game. *Mind*, 82:595–597, 1973.
- [100] Marvin Minsky. Communication with alien intelligence. In Edward Regis, editor, *Extraterrestrials: Science and Alien Intelligence*. Cambridge University Press, Cambridge, U.K., 1985.
- [101] Youngme Moon, Clifford Nass, John Morkes, Eun-Young Kim, and B.J. Fogg. Computers are social actors. In *Proceedings of the CHI Conference*, pages 72–78, Boston, MA, 1994.
- [102] James H. Moor. An analysis of the Turing Test. *Philosophical Studies*, 30:249–257, 1976.
- [103] James H. Moor. Explaining computer behavior. *Philosophical Studies*, 34:325–327, 1978.
- [104] Charles W. Morris. Foundations of the theory of signs. In *Writings on the General Theory of Signs*, pages 17–74. Mouton, The Hague, 1971.
- [105] Ajit Narayanan. The intentional stance and the imitation game. In P. Millican and A. Clark, editors, *Machines and Thought: The Legacy of Alan Turing*, pages 63–79. Oxford University Press, Oxford, U.K., 1996.
- [106] H.M. Parsons. Turing on the Turing Test. In W. Karwowski and M. Rahimi, editors, *Ergonomics of Hybrid Automated Systems II*. Elsevier, Amsterdam, 1990.
- [107] Charles Platt. What's it mean to be human, anyway? *Wired*, April 1995.
- [108] Richard L. Purtill. Beating the imitation game. *Mind*, 80:290–294, 1971.
- [109] T.L. Rankin. The Turing paradigm: A critical assessment. *Dialogue*, 29:50–55, 1987.
- [110] A.V. Reader. Steps toward genuine artificial intelligence. *Acta Psychologica*, 29(3):279–289, 1969.
- [111] G. Rey. What's really going on in the Chinese room? *Philosophical Studies*, 50:196–285, 1986.

- [112] R.C. Richardson. Turing Tests for intelligence: Ned Block's defense of psychologism. *Philosophical Studies*, 41:421–426, 1982.
- [113] L. Roberts. Searle's extension of the Chinese room to connectionist machines. *Journal of Experimental and Theoretical Artificial Intelligence*, 2:185–187, 1990.
- [114] Harvey Sacks. *Lectures on Conversation*. Blackwell, 1992. 2 volumes, Gail Jefferson (ed.).
- [115] Jerrold M. Sadock. On testing for conversational implicature. In *Pragmatics: Critical Concepts*, volume IV, pages 315–331. Routledge, London, UK., 1998.
- [116] Geoffrey Sampson. In defence of Turing. *Mind*, 82:592–594, 1973.
- [117] Paul Schweizer. The Truly Total Turing Test. *Minds and Machines*, 8:263–272, 1998.
- [118] John R. Searle. Minds, brains and programs. *Behavioral and Brain Sciences*, 3:417–424, 1980.
- [119] John R. Searle. Is the brain's mind a computer program? *Scientific American*, (262):26–31, 1990.
- [120] Benny Shanon. A simple comment regarding the Turing Test. *Journal for the Theory of Social Behavior*, 19(2):249–256, 1989.
- [121] Benny Shanon. Chauvinism: A misdirected accusation. *Journal for the Theory of Social Behavior*, 21(3):369–371, 1991.
- [122] Ravi Sharma and David Conrath. Evaluating expert systems: A review of applicable choices. *Artificial Intelligence Review*, 7(2):77–91, 1993.
- [123] Stuart M. Shieber. Lessons from a restricted Turing Test. *Communications of the Association for Computing Machinery*, 37:70–78, June 1994.
- [124] Dan Sperber and Deirdre Wilson. *Relevance: Communication and Cognition*. Basil Blackwell, Oxford, UK., 1986.

- [125] D.F. Stalker. Why machines can't think: A reply to James Moor. *Philosophical Studies*, (34):317–320, 1978.
- [126] John G. Stevenson. On the imitation game. *Philosophia*, 6:131–133, 1976.
- [127] Alan Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.
- [128] Alan Turing. Intelligent machinery. In D. Michie B. Meltzer, editor, *Machine Intelligence 5*, pages 3–23. Edinburgh University Press, 1969. Originally, a National Physics Laboratory Report, 1948.
- [129] Richard S. Wallace. The lying game. *Wired*, August 1997.
- [130] Stuart Watt. Naive psychology and the inverted Turing Test. *Psychology*, 7(14), 1996. <http://www.cogsci.soton.ac.uk/psyc-bin/newpsy?7.14>.
- [131] Joseph Weizenbaum. ELIZA—a computer program for the study of natural language communication between men and machines. *Communications of the ACM*, 9:36–45, 1966.
- [132] Joseph Weizenbaum. *Computer Power and Human Reason: From Judgment to Calculation*. W. H. Freeman, San Francisco, CA, 1976.
- [133] Thom Whalen. How I lost the contest and re-evaluated humanity, 1995. <http://debra.dgbt.doc.ca/chat/story95.html>.
- [134] Blay Whitby. The Turing Test: AI's biggest blind alley? In P. Millikan and A. Clark, editors, *Machines and Thought: The Legacy of Alan Turing*, pages 53–63. Oxford University Press, Oxford, U.K., 1996.
- [135] Deirdre Wilson and Dan Sperber. On Grice's theory of conversation. In *Pragmatics: Critical Concepts*, volume IV, pages 347–368. Routledge, London, UK., 1998.