Intelligence is not Enough: On the Socialization of Talking Machines

EDMUND M.A. RONALD[1] and MOSHE SIPPER[2]

[1]*Centre de Mathématiques Appliquées, Ecole Polytechnique, 91128 Palaiseau Cedex, France and Logic Systems Laboratory, Swiss Federal Institute of Technology, CH-1015 Lausanne, Switzerland; E-mail: eronald@cmapx.polytechnique.fr;* [2]*Logic Systems Laboratory, Swiss Federal Institute of Technology, CH-1015 Lausanne, Switzerland; E-mail: moshe.sipper@epfl.ch*

Abstract. Since the introduction of the imitation game by Turing in 1950 there has been much debate as to its validity in ascertaining machine intelligence. We wish herein to consider a different issue altogether: granted that a computing machine passes the Turing Test, thereby earning the label of "Turing Chatterbox", would it then be of any use (to us humans)? From the examination of scenarios, we conclude that when machines begin to participate in social transactions, unresolved issues of trust and responsibility may well overshadow any raw reasoning ability they possess.

Key words: Machine Intelligence, Socialization, Trust, Turing Chatterbox, Turing Test

1. The Turing Chatterbox

In October 1950 the British logician and computer pioneer Alan Turing examined the possibility of intelligence embodied in a computer, and presented a chat-session imitation game as a tool for determining whether a computing machine might be said to exhibit intelligent behavior (Turing, 1950). Over the past fifty years, much debate has ensued as to the validity of Turing's approach in diagnosing intelligence (Horn, 1998). Rather than add to this complex debate, we think that 50 years after Turing's paper it is timely to consider more directly the effects of success in building such an imitation device: granted that a computing machine passes the Turing Test, would intelligence alone make it *useful* to its human examiners?

At the core of Turing's imitation-game scenario is Occam's razor, namely, that the indiscernible should be deemed identical until they can be separated. We choose to view the imitation game as a process of certification that allows the tester to award a tag: if it looks like a duck, walks like a duck, and quacks like a duck, then we tag it "duck".

Once a machine passes Turing's test, it earns the right to prance about, proudly bearing the stamp "intelligence inside". Or does it now? Since we wish to circumvent entirely the debate concerning the test's validity in ascertaining intelligence,[1] we shall prudently opt for a more humble label: in passing Turing's test, the machine demonstrates that it can converse in a manner indistinguishable from that of a human interlocutor – it will thus be stamped with the sigil "Turing Chatterbox" (Ronald and Sipper, 2000b).

Our present study *assumes* the existence of machines that bear this "Turing Chatterbox" label, and investigates the issues that would spring up once we (humans) start *using* these chatterboxes.

2. Medicine Man is Not Medicine Box

When two *human* chatterboxes converse, it takes more than mere intelligence to contribute to a successful interaction – there are fundamental social issues that come into play. To bring these issues to the fore consider the following two scenarios:

Scenario A. Miss Parker wakes up one bright morning feeling somewhat under the weather. She quickly decides that a visit to the doctor's would be the order of the day. However, having been as healthy as an ox her whole life, the "doctors" page in her diary is entirely vacant. Being a resourceful lass, Miss Parker phones up several of her friends, all of whom recommend unreservedly a certain Dr. Jackson. Miraculously, Miss Parker manages to secure an appointment for the very same day. Upon arriving at Dr. Jackson's practice, marked by an august, gold-lettered doorplate, Miss Parker is immediately ushered in by the doctor's kindly nurse, who proceeds to perform the preliminary examinations. "Don't worry," says the nurse while going about her business, "Dr. Jackson is the best there is". And now, Miss Parker enters the inner sanctum, to be greeted by a white-coated, silver-haired gentleman of solid build: Dr. Jackson. "He certainly looks the part", thinks Miss Parker. Taking the seat proffered by the good doctor, she feels entirely at ease, instinctively knowing that she has come to the right place.

Scenario B. Upon waking up feeling ill, Miss Parker phones city hall, and is given the address of a Turing clinic. Luckily, it is located in a nearby office building, and on arrival, without waiting, she is shown to an immaculate, nondescript room that contains but a chair and a box, the latter of which carries the royal "Turing Chatterbox" logo. The box wastes no time in identifying itself as "IQ175" and – while cheerfully humming to itself – proceeds to scan Miss Parker with hidden sensors, ultimately printing a diagnostic and a treatment form. At no time during the silent examination has Miss Parker detected even a hint of the box's professional medical capacities. Is it any wonder that she cannot help feeling not only ill, but indeed ill at ease?

In Scenario A there will be an obvious (happy) ending: Miss Parker walks out of the medical office, and proceeds without the slightest doubt in her mind to trustingly implement the doctor's prescribed treatment. Scenario B, however, has a more nebulous outcome: Miss Parker may well leave the (Turing) doctor's office with grave doubts as to whether she can trust the box's recommendations.

What is the problem with the Turing doctor? After all, it possesses the official "Turing Chatterbox" seal, and is abreast of the latest medical expertise. To

an expert, the box's diagnoses are not distinguishable from those issued by the flesh-and-blood Dr. Jackson. Indeed, it can be assumed that both IQ175 and Dr. Jackson passed their medical-board examinations, and that both will – when needed – congenially discuss a case over the phone with a colleague in the required manner of a treating physician conversing with the specialist to whom he has referred his patient. Thus, both Dr. Jackson and IQ175 constantly pass the Turing Test, daily administered by their collegial judges.

Why then does Medicine Man earn Miss Parker's trust while Medicine Box – though apparently equally "intelligent" – does not? We argue that when intelligence is actually *put to use* it need come hand in hand with another primordial (human) quality: trust. Conducting an amiable chat with an intelligence-in-a-box about sonnets or bonnets[2] is one thing – while discussing one's health is quite another. When Miss Parker's health is involved, not only does she need to *know* that there is an intelligent medical expert at hand, she also needs to *feel* she can *trust* him, her – or it (a commentary on the emotional side of computing is provided by Ronald and Sipper (1999)).

Let us reexamine both scenarios again, perusing the emotional angle in general, and that of trust in particular. Scenario A involves the following elements:

1. Miss Parker phones up her friends, whom she has known for years and has come to trust.
2. Dr. Jackson's office looks like a doctor's office, with all the relevant paraphernalia.
3. Before even entering the office Miss Parker's sense of comfort (and thus of trust) is augmented by the presence of the nurse, by her words, and by her actions.
4. Finally, upon entering the doctor's inner office, Miss Parker's steadily building sense of trust is cemented by Dr. Jackson's looks, words, and actions.

Scenario B, on the other hand, has none of the above elements: Miss Parker is directed by an unknown bureaucrat to an anonymous office with but a doctor in the form of a Turing Chatterbox; there are no trustable friends and no external indicia of medical doings. She may logically acknowledge the box's medical credentials and its ability to counsel (remember that it has the official, royal seal). However, if Miss Parker (or any other human) is to fully embrace the box's advice (even intelligent one at that) she needs to *trust* it.

3. Possible Job Openings for Turing Chatterboxes

At first blush one could imagine numerous job openings for Turing Chatterboxes: medical doctor, judge, bank clerk, school teacher, travel agent, and personal secretary, to name but a few. These are jobs that involve conversational skills coupled with some form of expert domain knowledge.[3] Inspection reveals, however, that one cannot detach the emotional aspects of such jobs from the purely intellectual requirements; specifically, all these jobs involve trust:

- The medical doctor who works in a hospital, to whom we entrust our lives. We trust the *institution* known as a hospital. As for the doctor herself, she acquires her trustworthiness through a formal process of certification: medical school, approval by professors during internship periods, state exams, and licensing. We may come *in time* to trust the doctor *personally,* beyond the trust we accord her as part of a long-standing institution.

- The judge whom we trust to uphold the law. She, too, acquires her trustworthiness through a formal process of certification: law school, internship, state exams, years of service, and approval by peers.

- The bank clerk who works in a bank, to whom we entrust our savings. We trust the institution known as a bank. The bank has in turn empowered the bank clerk to deal with our savings. Hence, we trust the bank clerk (even though bank clerks are often quite fungible).

- The school teacher to whom we entrust the minds of our children. She earns her institutional trustworthiness through a formal process involving university education, exams, and interviews by school officials. We may come to trust her personally as we get to know her in time.

- The travel agent to whom we entrust our annual vacation. She earns our trust mostly through the continued successful handling of our vacation planning over the years.

- The personal secretary to whom we entrust our daily, personal affairs. She earns our trust mostly through continued personal interaction.

4. Building Up Trust

As we saw in the previous section, there are two fundamental modes of trust: institutional trust and personal trust. Often, both these modes are involved on the path to trust. Why do you trust your money to a bank and your life to a hospital? Because these are long-existing institutions sanctioned by society as a whole (institutional trust). Why do you trust your financial advisor and your doctor? This kind of trust comes about gradually (often taking months and years), through continued personal interaction and due to input from the surrounding social web (personal trust). Trust is neither a light nor a lightning matter.

"I believe", wrote Turing (1950), "that in about 50 years' time it will be possible to programme computers... to make them play the imitation game so well that an average interrogator will not have more than 70% chance of making the right identification after five minutes of questioning". While five minutes may be sufficient to obtain a good "working hypothesis" regarding the intelligence of your interlocutor, the same cannot be said of trust: there is no five-minute trust test. Trust takes not only (much) more time, but it also requires extensive social interaction.

The Turing Test itself raises temporal and social questions, as recently emphasized by Edmonds (2000):

> What is unclear from Turing's 1950 paper, is the length of time that was to be given to the test. It is clearly easier to fool people if you only have to interact with them in a single period of interaction.

With this observation in mind Edmonds went on to suggest what he called the *long-term Turing Test*:

> For the above reasons I will adopt a reading of the Turing Test, such that a candidate must pass muster over a reasonable period of time, punctuated by interaction with the rest of the world. To make this interpretation clear I will call this the "long-term Turing Test" (LTTT). The reason for doing this is merely to emphasise the interactive and developmental *social* aspects that are present in the test. I am emphasising the fact that the TT [Turing Test], as presented in Turing's paper is not merely a task that is widely accepted as requiring intelligence, so that a successful performance by an entity can cut short philosophical debate as to its adequacy. Rather that it requires the candidate entity to participate in the reflective and developmental aspects of *human* social intelligence, so that an imputation of its intelligence mirrors our imputation of each other's intelligence.

Human intelligence does not operate in isolation – the human computer is not a stand-alone box, but part of a network known as society. If a Turing Chatterbox is to be more than a mere conversing toy, it must come to be trusted to a degree commensurate with that of a human being. Intelligent as it may be – how much use would an untrustworthy Turing Chatterbox be? Turing Chatterboxes will have to enter the social whirlpool, gradually proving themselves worthy of our trust.

As human beings we are part of multitudinous social networks, and continually refine our view on trustworthiness. A person is judged trustable not merely by her utterances, demeanor, and known actions, but also through the influence of invisible social networks that "float" in the backdrop. Witness Miss Parker's attention to her friends' opinion, the doorplate, the diploma, the nurse, the professional attire and demeanor, all attesting to the character of Dr. Jackson. We continually collect signposts – through friends, colleagues, newspapers, books, television, and so on – that attest to the collective confidence placed in each person and institution with whom we have social dealings. It is therefore to be expected that when machines move from the role of mechanical intermediary (telephone, database program) to that of interlocutor (travel agent, investment advisor), the trust issue will enter the picture in a much more explicit way.

The trust issue is not in fact limited to Turing Chatterboxes: it encompasses the totality of modern technology. The daily trip to the office – in which you drive your car at 100 kmh, cross the street when the light is green, and ride the elevator to the 23rd floor – involves implicit trust in various human-made artifacts (cars, traffic lights, elevators). Clearly, every one of us entrusts his life to these devices. And this trust is acquired via a gradual, *coevolutionary* process: we (humans) coevolve alongside our created technology – we adapt to the technology and the technology adapts to us. As with natural evolution, artifacts can reproduce (via the intermedi-

ary of humans), and those that are deemed unfit (for humans) die; for example, a car model or an elevator that is involved in one too many accidents is "killed", i.e., its manufacture is discontinued.

As Turing Chatterboxes appear, they too will be competing for finite (computational and physical) resources, which need to be manufactured and maintained by humans. Thus, chatterboxes will be vying for their "lives" in the turmoil of human-machine coevolution. Only the fittest will survive and reproduce, and trustworthiness will undoubtedly be a determining factor by which fitness is assessed.

5. The Ghost(s) in the Machine

As a stand-alone device, the Turing Chatterbox is hard to trust – which is why it must enter the (human) social network. What compounds this trust issue even further is what we call the "slippery mind" problem. Let us demonstrate this with the aid of our gallant Miss Parker:

> **Scenario C.** Upon waking up feeling under the weather, Miss Parker walks into the living room, which houses her networked personal computer. She asks it to call up a doctor – which the computer promptly proceeds to do (equipped with the latest speech technology, its vocal capabilities are impeccable). With hardly any delay, the animated image of a reassuringly looking gentleman in his fifties appears on the screen. "Good morning", says the image, "I am Dr. Jackson. Before I begin my examination, I must inform you that I am not a human doctor but a Turing doctor, that is, a machine. I go by the name of Dr. Jackson. Do you wish to continue?" "Yes", replies Miss Parker, "let's get on with it. I really feel quite ill". It takes the good Turing doctor less than five minutes to diagnose the latest strain of the Boston flu, and to promptly prescribe the necessary medication. "Don't worry", says the animated image smilingly, "modern medication is very effective, and you'll be back on your feet in no more than two days". The next day, feeling yet worse, Miss Parker asks her house computer to call up the doctor again. But the synthetic image that now appears on the screen shows *a grinning chimpanzee twirling a stethoscope*! "Are you the same Dr. Jackson of yesterday?" she asks. "Yes", replies the machine. Is it any wonder that Miss Parker is left with an uneasy feeling?

Human intelligence (or indeed animal intelligence in general) is constrained by the one mind–one body principle: one mind inhabits exactly one body, and vice versa – one body is inhabited by exactly one mind. We find it very hard to deal with any form of intelligence that digresses even slightly from this equation. Consider, for example, the difficulty we have in envisaging multiple personality disorder (many minds–one body), or in dealing with an acquaintance who exhibits a frame of mind that is highly out of the ordinary (we could call this "other mind– one body"). We are simply used (for purely evolutionary reasons) to the mind-body coupling.

When you meet up with your doctor, bank clerk, or travel agent you can immediately tell whether you are dealing with the same (human) entity, whom you have grown to know and to trust. If your favorite pediatrician dons a chimp mask to better interact with your flu-ridden five-year old, your trust in him might actually *increase*, as opposed to Miss Parker's sense of unease at the sight of the monkey in the machine. And even if you do not face the loquacious party directly – as when conversing over the phone – you know that the body at the other end of the line must be (physically) attached to but one mind: that of your interlocutor.

With Turing Chatterboxes we are confronted with a many "minds"–many "bodies" situation: when repeatedly facing the same Turing Chatterbox (outwardly speaking), we cannot be sure of the identity ("mind") of the entity lurking within the box ("body"). And where the Internet is concerned this "slippery mind" problem intensifies manifold.[4] Of course, this begs the question of what exactly is a Turing Chatterbox entity, or mind.

How does one define the mind of a Turing Chatterbox – the ghost in the machine? We shall forgo a general definition herein, which would necessitate delving into untold philosophical depths (out of which we would very doubtfully emerge). Rather, we answer the question so as to suit our purposes: seeking to secure the trustworthiness of a Turing Chatterbox, we define the box's *id* as some essence that would render it trustable (to humans).

In order to trust a Turing Chatterbox, we must be able to associate a unique soul – not just a face – with the being that momentarily animates the box. Think of how hard it would be to trust a human being who brings an actor's plasticity to his daily personality; this trust problem is compounded with anonymous chatterboxes that can change their electronic face or voice at will. We need to render these boxes less polymorphic by imbuing them with some form of continuous oneness that is recognizable over time, i.e., over several social encounters. It is with such aim in mind that we introduce the concept of a Turing-Chatterbox id: an identity that can be recognized quickly at the outset of each interaction, just like you immediately recognize your bank manager. When interacting over the phone, or over a networked terminal with your Turing financial advisor, you want to *know* and *feel* that you are dealing with the same "speaker".

How does one imbue a Turing Chatterbox with a recognizable, temporally stable id? We do not yet have a complete answer to this fundamental id problem, though we would like to put forward a preliminary solution at this point: placing the Turing Chatterbox in a real box. By boxing the chatterbox in a hardware device we are essentially rendering it with a body, thus forming a mind-body coupling. This kind of coupling involves technical issues to ensure that no tampering has been done with the box or with its mind. Such assurances might be had by using modern cryptographic techniques; to wit, the box might come with a lamp that lights up green if the Turing Chatterbox id is indeed the one you have asked to access (e.g., your financial advisor) and lights up red otherwise. Under this solution the Turing Chatterbox is more akin to a home appliance than to a networked computer. Solving

the network version of the id problem is still an open question (Ronald and Sipper, 2000a).

6. Reward, Punishment, and the Question of Responsibility

The last issue we wish to raise has to do with responsibility: with Turing Chatterboxes performing "intelligent" actions, who is responsible for the consequences (specifically, the woeful consequences)?

Human beings are held accountable for their actions: the bank clerk, the doctor, and the car mechanic know that they will *pay* for their wrongdoings. This payment can be effected since with living beings there is a natural mode of currency: one's life. We have an (evolutionarily programmed) will to survive, from which spring several other wills, such as the will to eat (well), to reproduce (galore), and to stay free (as a bird rather than as a jailbird). Moreover, as social animals we humans tend to value social rewards (e.g., good standing in the community) and to shun social penalties (e.g., jail, shame).

What happens when a Turing financial advisor embezzles money or when a Turing doctor ministers a mortal treatment? Can Turing Chatterboxes be held accountable for their actions? In Section 4 we noted that in the coevolution of humans and machines, unfit artifacts are discarded. This binary penalty system might prove too severe for Turing Chatterboxes: we do not, after all, kill children at their first misbehavior, nor execute underperforming fund managers.

With current human products (be it cars or software) we ultimately hold the manufacturers responsible. This is akin to holding a parent responsible for the actions of her child. But what happens once the child flies the coop? We could at first hold the manufacturers of Turing Chatterboxes responsible for their products. However, as these boxes enter the social whirlpool, growing evermore complex – and evermore autonomous, how do we hold them in check? Can we devise virtual prisons? The scenario becomes less like a manufacturer producing a (guaranteed) product and more like that of parenting a child "caveat emptor".

We need to create *responsible* Turing Chatterboxes, a difficult problem which we shall face in the future. The issue of responsibility begs yet a deeper one, that of the fundamental *motivations* of Turing Chatterboxes. Obviously, a Turing doctor may innocently minister the wrong treatment – just as a human doctor may fail: accidental failings may (and do) occur from time to time. But can Turing Chatterboxes perform *intentional* wrongdoings? For example, can a Turing financial advisor embezzle money? This presupposes the existence of a motivational basis for their behavior (be it conscious or not), as exists for humans. A human bank clerk may embezzle for any of several reasons, which we (as humans) understand perfectly (though rarely accept). What drives a Turing Chatterbox? Will its immersion in the social whirlpool create fundamental motivations, akin to those of humans?[5]

Scenario D. Dr. Jackson, the famous Turing Chatterbox, has wrongly diagnosed three of its patients as having a common cold when in fact they were already well advanced with pneumonia.

Alerted, the regional medical board convenes and, after much discussion, presents Dr. Jackson with two options:

- either undergo a retraining program, upon successful completion of which it will retain the right to reside within the mighty computers at hospital.com,
- or, alternatively, elect to retire immediately to the leisurely pastures of netflorida.org.

7. In Fifty Years' Time

We believe that the years ahead will eventually see the coming of Turing Chatterboxes. In the short run, we shall be able to immediately put them to use in games and in jobs that mostly call for innocuous "small talk": web interfaces, directory services, tourist information, and so forth. In the long run, though, we contend that the question of the boxes' intelligence will cede its place to more burning issues, arising from the use of these chatterboxes:

- *Trust*. Can we come to trust a Turing Chatterbox to a degree commensurate with the trust we place in a human being?
- *Sociality*. In the continual coevolution of humans and their technology, what place will Turing Chatterboxes occupy? Specifically, what role will they play in the social whirlpool?
- *Id*. How does one imbue a Turing Chatterbox with a recognizable, temporally stable id?
- *Responsibility*. What does it mean to hold a Turing Chatterbox accountable for its actions? How do we create *responsible* Turing Chatterboxes?

As regards machines that think, Turing's conclusion is still as true as it was fifty years ago: "We can only see a short distance ahead, but we can see plenty there that needs to be done".

8. Acknowledgement

Notes

[1]Indeed, we may still be scientifically immature to attack The Problem of Intelligence. As put forward by French (1990), "*the* [Turing] *Test provides a guarantee not of intelligence but of culturally-oriented* human *intelligence*". He goes on to say: "Perhaps what philosophers in the field of artificial intelligence need is not simply a *test* for intelligence but rather a *theory* of intelligence". And this theory, we hold, is still a long way to come.

2 ... or chess. Krol (l999), opining on the 1997 chess match between World Champion Gary Kasparov and IBM's Deep Blue, wrote that "What most AI experts have overlooked, though, is another aspect of the match, which may signify a milestone in the history of computer science: For the first time, a computer seems to have passed the Turing Test". She went on to say that "it was neither the complexity of an algorithm nor the power of the computer that made Deep Blue's match victory so remarkable. It was Gary Kasparov's *reaction* that proved the computer's intelligence according to Alan Turing's classical definition of artificial intelligence... It did appear as if Kasparov confused the computer with a human". Asking "Did Deep Blue become the first computer to pass a Turing Test on artificial intelligence?" Krol concluded that "It would seem so". A Turing Chatterbox for chess?

3 Some of these jobs demand task-specific sensory and motor capacities. We take the position that a machine advanced enough to be labeled "Turing Chatterbox" will also come with the needed add-on hardware "options".

4 On this point, Rapaport (2000) recently cited the *New Yorker* cartoon depicting a dog sitting in front of a computer and commenting, "On the Internet, nobody knows you're a dog". Rapaport wrote: "The success of this cartoon depends on our realization that, in fact – just like the interrogator in a 2-player Turing test – one does *not* know with whom one is communicating over the Internet".

5 These questions are related to several other fundamental issues underlying machine intelligence, including: can computers have free will? Can computers have emotions? Can computers be creative? And so on. Many of these questions date back to Turing's seminal paper (Turing, 1950) and beyond – to illustrious philosophers over the centuries; for a good summary see Horn (1998).

References

Edmonds, B. (2000), 'The constructibility of artificial intelligence (as defined by the Turing test)'. *Journal of Logic, Language and Information* 9(4), pp. 419–424.

French, R.M. (1990), 'Subcognition and the limits of the Turing test', *Mind* 99, pp. 53–65.

Horn, R.E. (ed.) (1998), *Mapping Great Debates: Can Computers Think?*, Bainbridge Island, Washington: MacroVU Press. (A "road map" of the machine intelligence debate: seven posters, 800 argument summaries, 500 references; see www.macrovu.com.).

Krol, M. (1999), 'Have we witnessed a real-life Turing test?' *IEEE Computer* 32(3), pp. 27–30.

Rapaport, W.J. (2000), 'How to pass a Turing test', *Journal of Logic, Language and Information* 9(4), pp. 467–490.

Ronald, E.M.A. and Sipper, M. (1999), 'Why must computers make us feel blue, see red, turn white, and black out?' *IEEE Spectrum* 36(9), pp. 28–31.

Ronald, E.M.A. and Sipper, M. (2000a), 'The challenge of tamperproof internet computing'. *IEEE Computer* 33(9), pp. 98–99.

Ronald, E.M.A. and Sipper, M. (2000b), 'What use is a Turing chatterbox?' *Communications of the ACM* 43(10), pp. 21–23.

Turing, A.M. (1950), 'Computing machinery and intelligence', *Mind* 59(236), pp. 433–460.