Mind Versus Computer
M. Gams et al. (Eds.)
IOS Press, 1997

The Quest for Meaning

Louis Marinoff
Department of Philosophy, The City College of New York
137th Street at Convent Avenue, New York, NY 10031
marinoff@cnct.com

Abstract. This is a report of a three-tiered experiment designed to resemble a limited Turing imitation test. In tier #1, optical character recognition software performed automated spell-checking and "correction" of the first stanza of *Jabberwocky* (Carroll, 1871). In tier #2, human subjects incognizant of the poem spell-checked and "corrected" the same stanza. In tier #3, a widely-qualified group of academics and professionals attempted to identify the version rendered by the computer. Discussion of the experiment and its results leads to the notion of a "reverse Turing test", and ultimately to an argument against the strong AI thesis.

Keywords: Turing test, formalism, holism, strong AI thesis

1 Introduction

This paper describes and discusses the results of an exercise in experimental philosophy. The exercise was conceived partly to conduct one species of experiment belonging to the genus of Turing's (1950) imitation test, and partly to address some philosophical questions about the import of that test itself. The three-tiered experiment was initially designed neither to vindicate nor to vilify Turing's hypothesis; rather, it was undertaken in the spirit of prehypothetical curiosity. Each tier had to yield useful data in order that the experiment proceed to a subsequent tier, and I was quite unable to predict whether any tier would actually do so. Hence, in the first instance, I conducted the experiment simply to discover whether it *could* be conducted. That question having been answered empirically in the affirmative, my present task is to account for what transpired.

2 The Experiment

2.1 First Tier

To summarize the first tier: using an IBM-compatible hardware platform, a flat-bed scanner, and *Omnipage* OCR (optical character recognition) software, I scanned *Jabberwocky* into the OCR environment and invoked the ASC (automated spell-checker) routine, which proceeded to "correct" the spelling on a word-by-word basis. I later utilized only the "corrected" first stanza in tier #3 of the experiment. I will not reveal the computer's output at this juncture, in order that you too may take the test if you wish.

2.2 Second Tier

I deemed the output of the first tier sufficient unto the evil of the second. To conduct the second tier, I contrived a set of instructions that appeared easy to read, grasp and execute,

and whose execution itself would be consistent with the overall function of the computing apparatus.

I then elicited the participation of willing undergraduates, who were informed that this was an experiment in "cognitive science" (which is both perfectly true, and also vague enough to mean virtually anything). I stipulated that I preferred participants whose native tongue was other than English. Each participant was furnished with the original first stanza of the poem, and was asked if he or she had ever seen the text before. Three native speakers of English also insisted on participating. I am sorry to report that two of them had never seen *Jabberwocky* either. (I therefore assume that the poem has not been transmogrified into a video arcade game or—more's the pity—a television talkshow.) Upon answering "No", each participant was then given an English dictionary and the following instructions. An informal time limit of twenty minutes was imposed.

Instructions

> This is an experiment in cognitive science. You are given a short text, consisting of four lines, and an English dictionary.
>
> (1) On the given form, please indicate your native language, and also the number of languages you know (for statistical purposes only).
>
> (2) Read the given text carefully.
>
> (3) If you do not know a given word, then you may look it up in the dictionary.
>
> (4) If you do not find a given word in the dictionary, then try to think of a word you know, or try to find a dictionary word, that resembles the given word.
>
> (5) You may substitute a word you know, or a word you find in the dictionary, that resembles the given word.
>
> (6) If you neither think of a word nor find a dictionary word that resembles the given word, then you do not have to replace the given word.
>
> (7) Re-write the text, using any substitutions you will have decided to make. Please print neatly!

I administered the task to seventeen participants, whose native languages included Cantonese, English, Farsi, German, Mandarin, Persian, Punjabi, and Spanish. I could have continued indefinitely with this tier, but I terminated it after collecting what I deemed to be sufficient data with which to proceed to the next.

2.3 Third Tier

I circulated the following invitation on the internet, initially to three electronic bulletin boards devoted to philosophical interchange. I also circulated copies to some non-philosophers.

> You are invited to participate in a Turing test, which constitutes the third tier of the following experiment.
>
> TIER #1: By means of scanning technology, I loaded the poem *Jabberwocky* into state-of-the-art OCR (optical character recognition) software. I then invoked the software's automated spelling editor, which performed its "corrective" function on the scanned text.

TIER #2: I elicited the participation of a number of students, none of whom spoke English as a first language, and none of whom was familiar with the poem. I furnished each student with the original text of the first stanza of *'Jabberwocky*, with an English dictionary, and with the following set of instructions. (I inferred the pertinent instructions, which simulate the function of the OCR's automated spelling editor, from various experiments with the software itself).

The foregoing Instructions were reproduced here.

TIER #3: Appended below are eight numbered, "corrected" versions of the first stanza of *Jabberwocky*. To participate in this tier, simply answer the following question:

Which stanza was rendered by the computer software?
(Please make your selection by number.)

Please feel free to circulate this package (in its entirety) to other e-lists or to potential individual participants. I will seek to publish a full account of this experiment after its completion. Thank you very much for your participation.

Original stanza:

'Twas brillig and the slithy toves
Did gyre and gimble in the wabe
All mimsy were the borogroves
And the mome raths outgrabe.

"Corrected" versions:

(1)Twos brillig and the smithy troves
Did gyro and gamble in the wave
All Missy were the borogroves
And the Moe rats outgrabe.

(2)'Twas brilliant and the slithy tout
Did gear and gimbal in the wake
All mimics were the borough
And the mother rather outgoing.

(3)It was brillig and the slithery doves
Did gyrate and tumble in the wabe
All mimetic were the borogroves
And the mom wrath gave out.

(4)That was brilliant and the slither toves
Did gyro and gimble in the waive
All mimic were the borrow
And the mom rats out grabbing.

(5)'Twas bright and the slithy toves
Did whirl and gimble in the wabe
All mimetic were the borogroves
And the momentary raths outrage.

(6) This thrilling and the slithy stoves
Did gyrate and gimbal in the wabe

ver.	rendered by	selection freq. (±1 s.d.)	selection range: freq. ±1 s.d.
(1)	the computer	23.0% (± 2.9%)	20.1% - 25.9%
(2)	Cantonese spkr	18.3% (± 2.6%)	15.7% - 20.9%
(3)	German spkr	3.8% (± 1.3%)	2.5% - 5.1%
(4)	Spanish spkr	4.7% (± 1.4%)	3.3% - 6.1%
(5)	Cantonese spkr	1.9% (± 0.9%)	1.0% - 2.8%
(6)	Cantonese spkr	6.6% (± 1.7%)	4.9% - 8.3%
(7)	English spkr	18.8% (± 2.7%)	16.1% - 21.5%
(8)	contrived by experimenter	23.0% (± 2.9%)	20.1% - 25.9%

Table 1: Each version's origin, selection frequency, and statistical range

All mimsy were the borogroves
And the moment rather outrage.

(7) 'Twas bright and the slimy toads
Did gyrate and jumble in the waste
All mimicking were the borrow grubs
And the mole rashes out broke.

(8)Twos brillig and the slimy stoves
Did gyro and gamble in the wade
All flimsy were the boron groves
And the mom rats out grab.

I received two hundred and fifteen responses, from a variety of declared disciplines and professions; e.g. artificial intelligence, biology, computer science, economics, education, engineering (from chemical to software), English, history, law, library science, linguistics, management, mathematics, medicine, philosophy and psychology. Table 1 shows each version's actual origin, the frequency with which it was selected, and the statistical range of the selection (its frequency plus or minus one standard deviation).[1]

Note that a random selection process (using the same sample size) produces an expected average frequency of 12.5% ± 2.3%, thus a selection range of 10.2% - 14.8%. The ranges of versions (3), (4), (5), and (6) lie below the random range; this negative correlation with random selection suggests causal deselection of those versions. The ranges of versions (1), (2), (7), and (8) lie above the random range; this negative correlation suggests causal selection of those versions. Two respondents believed that none of the versions was rendered by the computer. More will be said about these results in the discussion of tier #3.

3 Discussion of Results

3.1 First Tier

The input is noteworthy in that it contains an error: I unwittingly mistranscribed Carroll's "borogoves" as "borogroves". This mistake was brought to my attention by six participants in tier #3. Although my error had arguably negligible effect on the subsequent tiers of the experiment, it contributes significantly to a central claim that I will make in due course.

[1] Let f_i be the frequency with which the i^{th} version was selected, and let n be the sample size. Then the associated margin of error is one standard deviation, $[f_i(1 - -f_i/n)]^{1/2}$.

The computer-generated output is noteworthy for several reasons. First, the substitution of " 'Twos" for " 'Twas" is—as some tier #3 respondents realized—vintage *computerese*. The ASC presumably treated the apostrophe as syntactic, and therefore ignored it.

Second, the substitution of "troves" for "toves" reflects another nuance of the ASC. The insertion of an extra character is bound to be computationally more complex than the substitution of a single character. The ASC made three simple substitutions in the second line ("gyro" for "gyre", "gamble" for "gimble", "wave" for "wabe"), so why did it not do so with "toves"? It could easily have found "doves", "loves", "moves", or "roves". Then again, if determined to insert a character, why not use the obvious "stoves"? I surmise that the ASC is loath to alter the leading character of a word, because every such character necessarily follows white space. Hence, during the image-building and recognition sub-routines of the scanning process, the leading character thereby gains relative contrast and therefore has diminished probability of being misconstrued or slurred with a predecessor. In sum, more integrity is imputed to the leading character than to any other in an unknown word. This principle also helps explain the substitution of "gyro"—instead of "lyre"—for "gyre". The letter "e" bears greater typographical resemblance to "o" than does "l" to "g", and the leading "g" is loaned greater credibility by default. Or so it appears.

Third, the substitution of "wave" for "wabe" is another matter. Given several possible third-character substitutions, e.g. "wade", "wage", "wake", "wane", "ware"; and moreover given the importance to the software of typographical resemblance, I would have supposed "wade" or "wake" better choices. But with respect to the configuration of the "QWERTY" keyboard, "v" is adjacent to "b", so perhaps "wave" is justified on the conjunctive grounds of proximity and presumed typing error.[2] Then again, "g" and "n" are also adjacent to "b". If the final choice in such circumstances were made by randomized selection among possible substitutions, such randomization would often defeat other, more "rational" features of the ASC. Then again, self-defeating routines are human in origin, and so is the software.

Fourth, I was surprised to encounter the proper nouns "Missy" and "Moe". The software had not generated any proper nouns in previous trials, and I still do not understand why it did so in these cases.[3] The common noun "missy" is the most straightforward substitution for "mimsy", and the ASC could scarcely have cared that it has an impolite—and therefore impolitic—postmodern connotation. (Perhaps someone else cared: maybe there are certain words which the software is instructed never to use, lest "it" give offense. An hypothesis of "politically correct optical character recognition" seems premature rather than far-fetched.) And the software must have performed digital gymnastics to render "Moe" in place of "mome", when a simple "mom" would have sufficed. Similarly, it was surprising that the ASC rendered "rats" from "raths", on the grounds that a single character substitution is simpler and generally more appropriate than a deletion. Although "paths" would have violated leading character integrity, "rates" seems plausible enough. I was also surprised that the ASC apparently lacks a function that would enable it to render "boron groves" from "borogroves" and "out grab" from "outgrabe". Such a function would enhance its present capabilities merely by compounding them.

Finally, I was not surprised that the ASC left "brillig" well enough alone.

3.2 Second Tier

Almost half the versions rendered by the human subjects seemed suitable for inclusion in the third tier. An important criterion of suitability was the way in which the human subjects implemented the instructions involving "resemblance".

Literally construed, "to resemble" means "to appear to the senses as". The human subjects were obliged to translate this abstract qualitative notion into informal quantitative terms. Having looked up or thought up words that "resembled" a given word, the subjects had to decide whether the *degree* of resemblance warranted substitution or not. Many substitutions differed from a given word by only one or two characters and, importantly, most of these were apparently found in the dictionary in the vicinity of the given word's putative alphabetical location (e.g. "mimic" for "mimsy"). Thus the property of leading-character integrity was preserved, if inadvertently, by many human subjects. "Brillig" presented problems to those concerned with resemblance narrowly construed. Just like the software, several humans left it untouched. "Brilliant" was a popular alternative, presumably because its leading six-place identity proved convincing. "Bright" was also popular, presumably because of its leading three-place identity plus its fortuitous inclusion of a "g".

More broadly construed, "to resemble" means "to suggest", and most contributions unsuitable for inclusion in tier #3 availed themselves liberally of this construal. For example "It was sunny and the slippery frog / Jumped and moved in the bushes . . ." shows bold interpretative flair and generous implementation of the concept of resemblance.

But the outstanding criterion of suitability was the avoidance of both syntactic and semantic content, and the majority of human subjects could not resist imbuing their renderings with grammar and/or meaning. The software, of course, functioned strictly on a word-by-word basis, attempting to correct lexical units only, ignorant of syntax and semantics altogether. Similarly, my instructions to the human subjects eschewed mention of syntactic and semantic considerations: all the operations were clearly confined to lexical units. Nonetheless, it proved almost impossible for the human subjects to treat words in lexical isolation; none could refrain from smuggling in some elements of grammar and/or meaning.[4]

Those who sought and found meaning in the stanza were not incorrect in one sense, since in fact the poem is not utter nonsense at all. Lewis Carroll created its vocabulary partly by truncating and conjoining valid English words (among my favorites is "galumphing", which ostensibly means "galloping in triumph"). But all subjects in this tier had no prior knowledge of the poem, and in the context of the instructions they were afforded no reason to assume that it meant anything. I suppose that they reintroduced meaning because they were metaphysically, psychologically and ideologically predisposed to do so.

For example, consider "And the mother rather outgoing", from version (2). This rendering clearly—and I think cleverly—attempts to satisfy both the experimenter's explicit demand for resemblance, and the subject's implicit demand for meaning. But none of the substitutions is really plausible according to the instructions, in that alternative substitutions more closely resemble the given words; and the whole is absurdly grammatical in the circumstances, in that it is most unlikely to be rendered by a mere lexical correction of the given parts. Yet this version was selected by tier #3 respondents with as great a statistical frequency as any other.[5] A number of versions unsuitable for inclusion in tier #3 all but sacrificed lexical resemblance for the sake of semantic content. Consider this imaginative offering, perhaps from a disgruntled concert-goer: "All mimmicing [sic] were the composers / And the public is promptly outraged."

This tier also fell prey, early on, to a dictionary which unfortunately contained the word "rath" ("a pre-historic hill-fort"). This unintended reward served to stimulate what I suppose to be another psychological tendency. Picture if you will the earnest human subject, dutifully looking up non-existent word after non-existent word, experiencing repeated disappointment if not frustration, and trying to puzzle out what, if anything, is intended by the text. Finally, on the penultimate word of the last line, the subject strikes paydirt. Contrary to conditioned negative expectation, an unknown word actually appears in the dictionary! In consequence,

[2] This suggestion was made by Meg Levin.

[3] I later learned from some tier #3 respondents that other spell-checkers frequently render capitalizations. See the Appendix.

[4] This is not entirely surprising, as abundant research indicates that lexical, syntactic and semantic modalities are not strictly separable in the evaluation of language (e.g. Johnson 1986, pp. 117 *ff*).

[5] It is also the only version in which every non-valid word is replaced; however, tier #3 respondents were aware that the instructions to tier #2 subjects do not compel substitution in every instance.

it becomes a focus of meaning for the entire line; e.g. "And the fairies dug out raths." This is another clever attempt to couch semantic content in lexical resemblance, and the subject even pencilled notes which illustrate her thought-processes: "mome" was interpreted as "gnome" (which was then pluralized and transmuted to "fairies"), while "outgrabe" conveniently became "dug out". Another subject was evidently so transported at finding "rath" in the dictionary that he substituted its definition wholesale: "And the wonderous [sic] hill-forts camouflaged". This conveys a palpable chunk of meaning, and it even harbours a vestige of resemblance. Although I speedily replaced the dictionary in question with a less fortified edition, I could not dispossess the human subjects of their penchants to make the stanza more grammatical and meaningful.

While the half-dozen most suitable versions offered sufficient ingenuity and variety to motivate tier #3, they also posed one serious problem. All the human subjects—whether aided by the dictionary or not—had identified and variously substituted for the contraction " 'Twas". The computer alone had come up with " 'Twos", and its version seemed obviously distinguishable from the others on that ground alone. So I wilfully engaged in a subterfuge. I contrived version (8) as a decoy, which not only reproduced the problematic distinguishing feature of (1), but which also did things that I thought the computer itself could or should have accomplished. I also confess to having deleted version (1)'s syntactically irrelevant leading apostrophe, which seemed to me a dead giveaway. But given what transpired in tier #3, I probably need not have worried.

3.3 Third Tier

I would prefer to offer herein a simple *résumé* of the results of this tier, and to reserve philosophical discussion for the concluding section. But an unavoidable methodological question now arises, that cannot be divorced from certain presuppositions; namely, with respect to which hypothesis are the quantitative data to be interpreted? And a conceptual question also arises: how—if at all—does the experiment relate to Turing's imitation game?

At least one philosopher who responded in tier #3 raised the obvious objection to the experiment (and described it as obvious even while raising it, and charitably supposed that I had already conceived a reply to it): is this experiment a Turing test at all? The question's tenor is rhetorical, for in at least one respect the answer appears clearly negative. Turing's original imitation test entailed a phase during which both human and computer were subjected to blind interrogation. Thus, to fulfil this condition strictly, our interrogators (tier #3 respondents) should have been able to ask questions of, and receive answers from, all the agents who produced the tier #2 versions of the poem. Under Turing's ideal conditions, this might have given rise to exchanges such as:

> Interrogator [to agent (2)]: "What does 'brillig' mean to you?"
>
> Agent (2): "It doesn't mean a thing to me."
>
> Interrogator: "Why did you substitute 'brilliant' for 'brillig'?"
>
> Agent (2): "Because I was following the experimenter's instructions, and 'brilliant' is a dictionary word that resembles 'brillig'."
>
> Interrogator: "What do you understand by the assertion 'X resembles Y'?"
>
> Agent (2): "I understand by it that 'X appears similar to Y' in some basic way."

> Interrogator [to agent (1)]: "What does 'brillig' mean to you?"
>
> Agent (1): "It doesn't mean a thing to me."
>
> Interrogator: "Why didn't you substitute 'brilliant' for 'brillig'?"
>
> Agent (1): "Because I was following the experimenter's instructions, and although 'brilliant' is a dictionary word that is identical with 'brillig' in its first six characters, it does not sufficiently resemble 'brillig' to warrant the substitution."
>
> Interrogator: "What do you understand by the assertion 'X resembles Y'?"
>
> Agent (1): "I understand by it that 'X appears similar to Y' in some basic way."

From these hypothetical dialogues (or dialogues which resemble them), I assert that the ideal interrogator would not be able to infer that agent (2) is human, and agent (1) is a computer. The ideal interrogator would be able to infer firstly that the agents employ different criteria in their respective assessments of the truth-value of the proposition 'X resembles Y', where Y is a dictionary word and X is not, and secondly that both agents behave in ways consistent with the experimenter's instructions and their respective assessments. Both agents are able to furnish the interrogator with plausible reasons for their respective decisions. The hypothetical computer would therefore pass Turing's imitation test.

But note that the interrogator does not require these hypothetical dialogues to draw the previous inference. In fact, many (and perhaps most) dialogues of this kind are already implied by the instruction set in tier #2, which was shown to all "interrogators" in tier #3. While instruction (5) grants permission to make a substitution, instruction (6) declares in effect that a substitution should be made only on the grounds of sufficient resemblance. Nowhere is "sufficient resemblance" defined for the interrogator, yet the notion is implicitly constrained by the information that the spell-checker functions exclusively on a word-by-word basis, which in turn implies that it ignores syntax and semantics. This information is surely inferable from the instruction set which, recall, is said to be consistent with the function of the spell-checker.

In consequence, the interrogator should conclude that both the substitution "brilliant" for "brillig", and no substitution at all for "brillig", are consistent with plausible computers, and therefore that the computer's version must be distinguished by some other means.

It follows that this experiment cannot be said *not* to be a Turing test on the grounds that it fails to allow for dialogues between the interrogator and the agents, as concerns their output. In fact, many such dialogues can be conducted implicitly by an interrogator, in the manner of thought-experiments. I claim that, by asking rhetorically whether particular substitutions are consistent with the instruction set, an interrogator can indeed eliminate versions (2) through (8).

Consider, for example:

> Interrogator [thought - experimentally, to agent (2)]: "Why did you render 'And the mother rather outgoing'?"
>
> Interrogator [thought - experimentally, for agent (2)]: "Because it is grammatical, and moreover it has meaning."
>
> Interrogator [thought - experimentally, to agent (2)]: "I conclude that you are not the computer."

Similarly, the human authorship of version (3) is betrayed by the substitution of "gave out" for "outgrabe"; for while a computer's spell-checker might have rendered "out gave" (more likely "out grave"), it would not have inverted the word-order for syntactic purposes. In version (4), the grammatical giveaway is "out grabbing". As well, "That was" is an incorrect expansion of " 'Twas", while "borrow" is somewhat anomalous. In version (5), the substitution "whirl" for "gyre" is synonymous—the product of a thesaurus, not a spellchecker. In version

(6), the substitution of "This" for " 'Twas" is incorrect, and "thrilling" for "brillig" is quite suspect. In version (7), "mimicking" is a long way from "mimsy", but the clincher is "And the mole rashes out broke". While "mole rashes" is undeniably ingenious, the spell-checker couldn't have known that rashes "break out"—and in the past tense, "broke out". Hence, the computer could scarcely have rendered "out broke", because that is cognate with the inversion of a syntactic form which is itself embedded in a semantic context.

This leaves versions (1) and (8) as the sole possibilities. (Many tier #3 respondents arrived at this conclusion, and volunteered arguments consonant with the foregoing). But while there are persuasive reasons for selecting either (1) or (8), there is perhaps an overriding reason for eliminating version (8). The telling structural difference between these two versions is that (1) conserves the number of given words, whereas (8) does not. The latter version splits "borogroves" into "boron groves" and "outgrabe" into "out grab". While this operation may well be the kind of thing that OCR software ought to do, the operation is not in fact entailed by the given instruction set. In order to entail it, instruction (4) would have to be modified to read "If you do not find a given word in the dictionary, then try to think of a word *or words* you know, or try to find a dictionary word *or words*, that resembles *or resemble* the given word" (modifications emphasized). Instructions (5) and (6) would require similar modification. Then the operation of word-splitting would be strictly inferable from the instruction set. But as things stand, version (8) (among others) is guilty of having derived "is" from "ought".[6] By hypothetical default, then, version (1) is the remaining choice.

Empirically, however, the conservation of word number, which the given instructions imply, appears as a statistical non-factor in the tier #3 decisionmaking process. One-hundred-and-six respondents chose versions which conserve word number [(1),(2),(5), or (6)], while one-hundred-and-seven chose versions which do not conserve it [(3),(4),(7), or (8)]. No respondent made explicit mention of this "conservation law" and its violation by half the versions on offer.

Many respondents voluntarily communicated other ratiocinations. For example, one found "smithy troves" more *computeresque* than "slimy stoves" because it is less grammatical. This may well be the case, but then again "smithy troves" is coincidentally more poetic: it connotes images of beaten copper and hammered gold—a blacksmith's treasure-trove. Others chose version (1) primarily because of its proper nouns; these respondents were well-acquainted with spellcheckers that routinely render capitalizations. Still others (as anticipated) seized upon "Twos" as a basis for eliminating all but versions (1) and (8), but could not choose decisively between them. Then again, some respondents' reasonings were far from consistent: many indicated their second choices as well as their first, and some who selected either (1) or (8) in the first place did not necessarily select (8) or (1), respectively, in the second. Moreover, some who selected neither (1) nor (8) in the first place selected either (1) or (8) in the second.

We return to the conceptual question: how does this experiment relate to Turing's imitation game? I claim that although the experiment is not a literal Turing test in the original sense, it is another species of that same genus. This experiment constitutes a "reverse Turing test", which inquires not how proficiently a computer can imitate a human; rather, how proficiently a human can imitate a computer. And on this view, the statistical data bear further comment.

Clearly, the empirical results corroborate the argument that (1) and (8) are the sole plausible computer-rendered versions, notwithstanding (8)'s failure to conserve word number. Statistically, versions (1) and (8) were together selected with significantly greater frequency than were (2) and (7): 46% ± 2.4% for the former pair, versus 37% ± 2.1% for the latter. Then again, on an individual basis, none of these four most popular versions was selected with statistically greater frequency than any other; their individual selection ranges all overlap. These data certainly suggest that a human can successfully imitate a computer, at least in the estimation of other humans. But the data conflict directly with the argument *ex hypothesi*, that version (1) is uniquely identifiable as the computer's. Theoretical and logical considerations

[6] A common feature of all the computer-generated stanzas (with the exception of one occurrence involving hand-written input) is their conservation of word number. See the Appendix.

indicate that versions (2), (7) and (8) should not have been selected with greater frequency than versions (3), (4), (5) and (6), because they all bear distinct marks of human fabrication.

This leads to a further question: who or what is the supreme arbiter of proficiency in such tests? On what does the credibility of an imitation ultimately depend? Turing seems to have assumed that a good correspondence would generally obtain between theoretical and empirical evaluations of a given imitation. Turing's interrogator resembles the philosopher's imaginary "man on the Clapham omnibus" and the jurist's fanciful "reasonable man". They are all incorruptible and infallible appraisers of evidence; in other words, they cannot be deceived unless, of course, the experimenter, barrister or philosopher intends that they be deceived. I hold that such a correspondence need not obtain. An imitation which the experimenter adjudges credible may be rejected by relatively many interrogators as incredible; or—as in this experiment with respect to versions (2), (7) and (8)—an imitation which the experimenter adjudges incredible may be accepted by relatively many interrogators as credible. Hence a given experiment may inform the experimenter about the nature of credibility either less or more than it is informed by him.

In our reverse Turing test, the humans who produced versions (2), (7) and (8) proved theoretically improficient yet empirically proficient at imitating the computer. This in turn suggests that many interrogators were not very proficient at gauging the credibility of the imitation. While it may be objected that the tier #2 subjects did not know the real purpose of their endeavour, this objection can be finessed by considering that in the original Turing test, the computer need not "know" that it is imitating a human. It follows that in a reverse Turing test, a human need not know that he or she is imitating a computer. Moreover, that the humans were indeed imitating a computer follows syllogistically from the premises that the humans were asked to implement a set of instructions, and that those instructions (if followed) simulated the function of a computer. Neither the computer nor the humans possessed any broader knowledge of the context itself, yet the differences in their respective functions were demonstrable.

In the concluding section, I discuss the reverse Turing test more generally, with the intention of proposing new ways—or at least new bottles for old ways—in which to illustrate differences between humans and computing machines.

4 The Reverse Turing Test

Turing posited his imitation test in a generation when computer science was nascent and computing technology comparatively primitive. In Turing's day, if one conceived of pitting a computer against a human in contests designed to measure "intelligence" generically construed, the computer was the absolute underdog. In fact, the computer was a pre-underdog, in that the technology was not advanced enough to permit such contests to take place.

Turing predicted that computers would become sophisticated enough so that, in some specified context, human interrogators would be unable to decide (with more than 30% accuracy) whether a computer or a human had rendered a given body of nominally conversational output. In other words, Turing envisioned computing progress only to the extent that human versus machine output would be indistinguishable to human interrogators. Turing seems to have supposed that a computer's ability to imitate a human would improve as a smooth function of its increased storage capacity. While current storage capacities are now remarkably close to those predicted by Turing, the computer's "cognitive" abilities have lagged far behind, to the extent that no true imitation test, in Turing's original strong sense, has yet proved feasible.

Turing's hypothesis has been weakly vindicated in many narrow contexts, notably with programs like Weizenbaum's "ELIZA" (e.g. see Boden 1977, Johnson 1986). And, for example, a test group of psychiatrists could not distinguish a transcript of the output of Colby's program "PARRY", which simulates the verbal responses of a paranoiac, from transcripts of dialogues

with human paranoids (e.g. Boden 1977, pp.96-111 *ff*). We generously interpret this as a success of computing, rather than a failure of psychiatry. And, for example, James Sheridan's team has "taught" a computer to compose lyrical poetry within a specified structure, such that test subjects cannot distinguish its better efforts from poetry composed within the same structure by humans (Kern 1983, Sheridan 1987). While these examples, among others, constitute successful Turing imitation tests in a very weak sense, they naturally tend to fuel rather than to resolve the debate surrounding the strong AI thesis.

Proponents of the strong thesis, or "formalists" (e.g. Minsky 1968, Hofstadter 1981) hold that human intelligence is a property wholly explicable in terms of algorithmic complexity. Given sufficiently powerful hardware and sophisticated software, formalists believe that a computer can be built which exhibits understanding, awareness of meaning, and any and all aspects of human consciousness. They hypothesize moreover that all aspects of human consciousness consist of nothing but complex algorithms executed by a "biological computer". Opponents of the strong thesis, or "holists" (e.g. Searle 1984, Penrose 1989) hold that understanding, awareness of meaning, and other aspects of human consciousness cannot be explained solely in terms of algorithmic complexity. Holists believe that even if a computer could be built which passes any conceivable Turing test, this would not necessarily demonstrate that the computer is self-aware, that it understands what it does, or that it possesses consciousness of the human quality.

My central claim ultimately bears on this debate, but it is advanced initially on quite a different tack. On one view, progress in AI now begins to satisfy Turing's expectations, because we can conduct successful imitation tests, if but in a very weak sense. On another view, progress in computing still falls short of Turing's expectations, because there remain any number of imitation tests that the computer readily fails. Then again, on a third view, computers are able to out-perform humans in many areas, and in this sense have perhaps exceeded Turing's expectations. When it comes to performing quantitative tasks in competition with humans including playing games such as checkers and backgammon, or even chess and Go the computer is no longer underdog but overdog; not yet and perhaps never to be a Nietzschean *übermensch* in evolutionary terms (e.g. Nietzsche 1982), but demonstrably an *überhund* at parlour games.

While much of the computer's outperformance of humans is confined to various forms of "high-speed idiocy"[7] (i.e. number-crunching and the like), many humans display, by contrast, various forms of "low-speed genius" (e.g. mathematical intuition and artistic creativity). I submit that a—perhaps *the*—salient difference between computer versus human performance lies not merely in *what* they can or cannot do, rather in *how* they attempt to do what they can or cannot do. In methodological terms, the computer is an entity that strictly follows instructions, while the human is a being that constitutionally disregards them. Computers do exactly and only what they have been instructed to do, whereas humans are capable of an inexactitude that includes but is not restricted to the self-prompted or unconscious misinterpretation, omission, permutation and modification of members of a given instruction set.

In the course of this experiment, I made typically human errors in carrying out my own meta-instructions. The first involved the mistranscription of "borogroves" for "borogoves". I suppose that I too succumbed to the spell of meaning—after all, any kind of "grove" is more meaningful than every kind of "gove". My second error involved misinforming the tier #3 respondents that all the human versions were rendered by non-native speakers of English. I subsequently rediscovered in the experimental log that version (7) was rendered by a native speaker of English.

A characteristically human disregard for the tier #2 instructions was displayed by several tier #3 respondents, who chose version (2) on the grounds that it is the only version which

[7]This phrase was used by Gleick (1987) to describe a dismissive attitude of some mathematicians and scientists toward computers.

contains all and only valid words. The instructions do not necessitate that condition. A creative human disregard for both the tier #2 and the tier #3 instructions was displayed by the two respondents who concluded that none of the eight versions was rendered by a computer. One of these two respondents argued that all the versions were rendered by human test-subjects, because the original "gyre" is a valid word which every version had replaced. The other expressed a synthetic a priori suspicion that all eight versions had been contrived by the experimenter. While these disregards affect the experiment's statistics but negligibly, they affect its conclusions significantly.

I am fairly certain that all my undergraduate students are capable, say, of carrying out the following instruction: "If you wake up in the middle of the night, make yourself a sandwich." But no robot is yet capable of carrying this out, for at least two reasons. First, the antecedent of that instruction, although decidable by humans, is not sufficiently comprehensible to humans to be made intelligible to or analogous for a computer. (What is the nature of sleep, wakefulness, dreaming, somnambulence? Like Descartes, how do you know that you are not dreaming that you are awake? Pinch yourself, and see it if hurts.) But even if we simulate the antecedent by placing our robot in "sleep mode" (an idle, low-power-consumption state) at dusk, and by programming some probability with which it will "wake" before dawn, we will be defeated by the instruction's consequent until the frame problem is solved (e.g. see Pylyshyn 1987). The generic instruction "make yourself a sandwich" can be carried out by humans only because humans are able draw necessary and necessarily self-prompted inferences from a vast store of experience and background knowledge, which a robot simply lacks. Supplying a robot with a complete set of axioms, along with a complete set of rules for correct inference-making in an epistemological—as opposed to a logical—context, is as yet an unaccomplished task.[8]

What is more telling: even if we were able to solve this multi-faceted problem, we would be assured only that if the robot "woke up" during the night, it would indeed make itself a sandwich. For while the human being understands the instruction, the price of human understanding somehow entails the possibility of disregarding. The human is capable of beginning sincerely to seek the ingredients for a sandwich, of being disappointed or distracted by the findings, and of completing the task by ordering a pizza, or by obeying any other overriding caprice.

By contrast, I am very certain that, if I instruct my undergraduate students to print their names according to the format "last name, first name, middle initial(s) if any", at least one and probably more will write in script, or will invert the ordering, or will omit their middle initial(s), and so forth. But if I instruct (i.e. program) my computer to print out a class list according to that format, then—given the data—it will do so with a negligibly small chance of making a functional error. In an overwhelming majority of such trials, the computer would execute my instructions flawlessly.

This general idea suggests a way to thwart a Turing imitation scenario. Let the interrogator give the agents instructions for the performance of some task (i.e. the generation of some verbal output). The interrogator will soon discover which agent disregards them or, commensurately, makes errors—whether intended or unintended—in their execution. That agent is human. Thus the reverse Turing test can be employed to ferret out the agents' true identities.

Note that programming a computer to output wrong answers to questions does not circumvent the reverse Turing test, for the instruction set would then have to contain a member which says, in effect, "compute the correct answer and output a different answer". The interrogator would be aware of this instruction, so an unbroken string of wrong answers would again point to the computer—for the interrogator would find that the human agent will sooner or later make a mistake and, in this case, inadvertently output a correct answer. Imagine, if you will, playing "Simon says" with a host of humans and an ideal robot. If it doesn't malfunction, the robot cannot lose; and increasingly reliable technologies diminish the likelihood of such

[8]Turing (1950) recognized this problem in a section called "The Argument from Informality of Behaviour", and adroitly side-stepped it.

malfunction. But even the most accomplished human player will eventually err.[9]

Naturally there are trivial cases in which the interrogator could not distinguish the agents. For example, if the instruction set said: "Flip a coin one hundred times, and output the results in random order", then only a small proportion of humans agents would mistakenly output, say, ninety-nine or one-hundred-and-one results. Similarly, a small proportion of human agents would disregard the instruction about randomizing output order, and would output the results in their obtained order (or some other order), while the randomness of the raw results themselves would preclude the interrogator's verification of their random re-ordering. But this example is utterly trivial, whereas Turing's examples of imitation tests are far from trivial, even by today's computing standards. Any useful reverse Turing test would have to be non-trivial too.

At first blush, the theses "It is conceivable that a computer can imitate a human" and "It is inconceivable that a human can imitate a computer" seem logically and empirically independent, in that the demonstrable truth of the latter appears not to condition the conjectural truth or falsehood of the former. But I claim that a deeper reading of the latter provides evidence against the former; in other words, that the reverse Turing test gives rise to an argument against the strong AI thesis.

Consider the following two syllogisms, which represent (respectively but not uniquely)[10] the formalist and holist positions:

All and only intuitively computable functions are Turing computable. (Church's thesis)

Understanding and meaning are intuitively computable functions. (formalist premise)
Therefore understanding and meaning are Turing computable. (strong AI thesis)
All and only intuitively computable functions are Turing computable. (Church's thesis)

Understanding and meaning are not intuitively computable functions. (holist premise)
Therefore understanding and meaning are not Turing computable. (contra strong AI thesis)

These arguments cannot both be sound and, if Church's thesis is false, they are both unsound. But one may suppose Church's thesis to be true (e.g. see Boolos and Jeffreys 1974). One cannot prove it true; one could only prove it false, by finding a counterexample. No counterexample has yet been found. Moreover, one can suspect that Church's thesis is true, because independent arguments lead to its equivalent statement (e.g. Turing 1937, Church 1941.) The "burden of proof" plausibly shifts to a "burden of disproof", in the absence of which we can believe the thesis confirmed until disconfirmed.

And we have reasons for supposing that understanding and meaning are not intuitively computable. The reverse Turing test furnishes one such reason. Suppose that a human (H1) is given a set of instructions (S1) which, if faithfully executed, would result in the imitation of some Turing machine (T1). But suppose that the human makes meaningful mistakes in their execution. Now we ask whether we can build another Turing machine, T2, such that T2 can similarly make meaningful mistakes. If we reply "no", then the strong AI thesis fails because Church's thesis fails, for we will have found an intuitively computable function which a Turing machine (T2) cannot perform: namely, misunderstanding, a function whose successful performance fails to imitate another Turing machine (T1). If understanding is intuitively computable, as the formalists claim, then misunderstanding should be intuitively computable too.

So formalists presumably reply "yes": we can build such a Turing machine, T2, which fails to imitate T1, and therefore which passes the Turing test in question. But whereas the human H1 fails to imitate T1 by virtue of making meaningful mistakes while executing S1, T2 must be

[9]Turing (1950) also anticipated this possibility in a section entitled "Arguments From Various Disabilities", but discounted it because his generation of computers was disposed to significant functional error.

[10]The holistic position herein affirms Church's thesis, as does Penrose (1989). One may also espouse a holistic position by denying Church's thesis.

given a set of instructions other than S1. For were T2 a universal Turing machine, T2 would execute S1 faithfully, would successfully imitate T1, would therefore fail to fail to imitate T1, and would therefore fail the test. So we must give T2 some other set of instructions, S2, whose faithful execution results in the failure to imitate T1. Then T2 would pass the Turing test in question.

But in that case, T2 would necessarily fail the associated reverse Turing test. For the reverse Turing test depends on an interrogator's examination of input as well as output. An interrogator would note that input S1, which should have led to an imitation of T1's output, failed to do so; and that input S2, which should not have led to an imitation of T2's output, succeeded in not doing so. An interrogator would then conclude that S1 had been improperly executed by a human, and that S2 had been properly executed by a Turing machine. (An interrogator could fail to distinguish the agents only in the event that the interrogator improperly executed the meta-instructions governing the reverse Turing test itself, and thus unwittingly played the human role in a hypothetical second-order reverse Turing test. This actually occurred in tier #3 herein, in the cases of the two respondents who decided that no stanza was computer-generated.)

Now a formalist could object that T2 is not on a "level playing-field" with H1. In other words, a formalist could claim that the human brain is actually running simultaneous parallel background programs (the biological equivalent of multiple "memory-resident" routines), and that mistakes in executing a given instruction set (i.e. so-called "human errors") arise from problems of interference, override, timing and other difficulties latent in parallel data-processing. A formalist might claim that a meaningful mistake is just a complex kind of human software "bug" or wetware dysfunction, which occurs when (putative) semantic, syntactic, analytical, emotive and other instruction sets become conflated during simultaneous execution. A formalist would claim that we can in principle program memory-resident routines in a computer that would compel it to mis-process subsequent input in apparently "meaningful" but in altogether Turing computable ways; and thus that we can in principle build a Turing machine that would fool an interrogator in a reverse Turing test.

A holistic reply to this objection is straightforward, and is consistent with the justification for assuming Church's thesis to be true; namely, that the burden of disproof lies with the doubter. Continued failure to disconfirm Church's thesis lends evidential and heuristic support to its confirmation. Similarly, we cannot prove the holist premise that understanding and meaning are not intuitively computable functions. (And perhaps we cannot disprove it either, as Searle's Chinese Room implies). But continued failure to produce even a putative disconfirmation of the holist premise lends evidential and heuristic support to its confirmation. To that support I add this modest empirical result, and the bolder notion of the reverse Turing test to which that experiment gives rise. Let anyone who denies the holist premise produce not only a set of instructions that would allow a machine to pass a strong Turing test by meaningfully manipulating tokens of natural language, but also a set of meta-instructions that would allow a machine to pass a strong reverse Turning test by meaningfully misunderstanding instructions for manipulating tokens of natural language. While no computer extant can accomplish even the former task for want of explicit instructions, mind can accomplish both tasks in the absence of explicit instructions and meta-instructions alike. Until such be produced, I find no reason to discredit the holistic syllogism. Turing has yet to slay the Jabberwock.

Acknowledgements

I wish to thank the University of British Columbia's Centre for Applied Ethics, which provided the hardware and software for tier #1, the students at UBC who participated in tier #2, the respondents who participated in tier #3, and the respondents who volunteered other computer-generated stanzas. I would also like to thank those who afforded useful discussions about and comments on this paper; in particular, Andrew Irvine, Meg Levin, Michael Levin, James Sheridan, and the referees. Special

thanks to George Wolberg for help with LaTeX.

5 Appendix

These versions were generated by other software packages. They rather speak for themselves.

'Twas brisling and the smithy toes
Did gyre and gamble in the wade
All missy were the borogroves
And the Mme rates outgrabe.
(*MS Word 5.0*)

'Was broiled and the slushy moves
Did gyre and gamble in the wage
All mimes were the barographs
And the come rates utterable.
(*FrameMaker 3.0 and 4.0, PageMaker 3.0*)

'Twos brittle and the sloths doves
Did gyre and gimbal in the wake
All mamas were the brokerages
And the home wraths outcrop.
(*PageMaker 3.0, alternative*)

'Taws brillig and the smithy toes
Did gyre and gimbals in the wade
All maims were the borogroves
And the mime rates outgrabe.
(*WordPerfect 5.1*)

'Teas brillig and the sleuth tokes
Did gyre and gamble in the wage
All moms were the borogroves
And the mode rats outgrabe.
(*WordPerfect 5.1, alternative*)

Teas Willis and the sticky tours
Did gym and Gibbs in the wake
All mimes were the borrowers
And the moderate Belgrade.
(*Apple Newton*)

References

[1] Boden, M. (1978), *Artificial Intelligence and Natural Man*, Basic Books, Inc., NY.

[2] Boolos, G. and Jeffrey R. (1974), *Computability and Logic*, Cambridge University Press.

[3] Carroll, L. (1871), *Alice's Adventures in Wonderland, and Through the Looking Glass*, Three Sirens Books, NY (undated).

[4] Church, A. (1941), *The Calculi of Lambda-Conversion*, Annals of Mathematical Studies, #6, Princeton University Press.

[5] Dennett, D. (1984), 'Cognitive Wheels: The Frame Problem of AI', in *inds, Machines and Evolution*, ed. C. Hookaway, Cambridge University Press, Cambridge.

[6] Gleick, J. (1987), *Chaos*, Viking Penguin Inc., NY.

[7] Hofstadter, D. (1981), 'A Conversation with Einstein's Brain', in D. Hofstadter and D. Dennett, eds., *The Mind's I*, Basic Books Inc., NY.

[8] Johnson, G. (1986), *Machinery of the Mind*, Times Books, NY.

[9] Kern, A. (1983), 'GOTO Poetry', *Perspectives in Computing* **3**, #3, 44-52.

[10] Minsky, M. (1968), 'Matter, Mind, and Models', in *Semantic Information Processing*, ed. M. Minsky, MIT Press, Cambridge, MA.

[11] Nietzsche, F. (1982), 'Thus Spake Zarathustra', from *The Portable Nietzsche*, ed. W. Kaufmann, Viking Penguin Inc., NY.

[12] Penrose, R. (1989), *The Emperor's New Mind*, Oxford University Press, Oxford.

[13] Pylyshyn, Z., ed. (1987), *The Robot's Dilemma*, Ablex Publishing Corporation, Norwood, NJ.

[14] Searle, J. (1984), *Minds, Brains and Science*, Harvard University Press, Cambridge, MA.

[15] Sheridan, J. (1987), 'Basic Poetry', *The Computers and Philosophy Newsletter* **1**, 83-95.

[16] Turing, A. (1937), 'On Computable Numbers, with Application to the *Entscheidungsproblem*', *Proceedings of the London Mathematical Society* (series 2) **42**, 230-65; a correction **43**, 544-6.

[17] Turing, A. (1950), 'Computing Machinery and Intelligence', *Mind* **9**, 433-460.