From: *The Sherry Turkle Miracle*. Papers from a Hungarian Academy of Sciences online workshop, May 27, 2021

Barry Smith

Neural Chitchat

Introducing Little Bing

A constant theme in Sherry Turkle's work is the idea that computers shape our social and psychological lives. This idea is of course in a sense trivial, as can be observed when walking down any city street and noting how many of the passers-by have their heads buried in screens. In *The Second Self*, however, Turkle makes a stronger claim to the effect that where people confront machines that seem to think this suggests a new way for us to think – about human thought, emotion, memory, and understanding and thereby affects the way we think and see ourselves as humans.

I will attempt here[1] to throw a new light on claims of this sort by examining the Chinese chatbot 小冰 (pronounced "Xiǎoice", and loosely translated as "Little Bing"). Xiǎoice is a neural chatbot introduced by Microsoft in 2014,[2] and it is described in Zhou et al.[3] as "the most popular social chatbot in the world".

Zhou and his collaborators report that XiaoIce was "designed as an AI companion with an emotional connection to satisfy the human need for communication, affection, and social belonging". Their paper claims that XiaoIce "dynamically recognizes human feelings and states, understands user intents, and responds to user needs throughout long conversations". We are told further that since its re-

[1] This work is co-authored by Jobst Landgrebe, and some of the material within it is derived from a book manuscript entitled *There Will Be No Singularity* by Landgrebe and Smith.

[2] A visual impression of one of her achievements is here: https://www.youtube.com/watch?v=ihfbyvCzErw&t=199s.

[3] Li Zhou et al., "The Design and Implementation of XiaoIce, an Empathetic Social Chatbot", *Computational Linguistics*, vol. 46, issue 1, 2020, pp. 53–93.

lease in 2014, XiaoIce has "communicated with over 660 million users and succeeded in establishing long-term relationships with many of them".

Double Blandness

Like other "neural" chitchat applications, however, XiaoIce displays two major flaws, either of which will cause any interlocutor to realize immediately that they are not dealing with a human being and which will prevent any sane user from "establishing a long-term relationship" with the algorithm.

This is because such applications often create repetitive, generic, deflective, and bland responses, such as "I don't know" or "I'm OK", at least in longer conversations. This is because the training corpora which are used as training samples for algorithms of this sort contain many such answers, and so the likelihood that such an answer might somehow fit is rated by the system as high. Several attempts have been made to improve answer quality in this respect, but the utterances produced by the algorithms are still very poor.

"Bland" has two meanings: 1. the use of commonly repeated expressions, 2. the lack of any sort of creative step forward in the dialogue of a sort that would be of genuine interest or utility to the user. The reason for both of these effects is the method underlying how XiaoIce is built.

In this XiaoIce is analogous to a machine translation engine of the sort which merely reproduces sentence pairs from existing training sets. The translation corpus for the translation engine uses tuples of the form <l1s, l2s>, where l1s is the sentence to be translated in language 1, and l2s is some translation of l1s in language l2. XiaoIce uses a collection of tuples of the form <s1, s2>, which are pairs of sentences succeeding each other in one or other of the many dialogues stored in XiaoIce's large dialogue corpus.

Both google translate and XioaIce use statistical methods to generate inputs from outputs. And both merely mimic existing input-to-output-tuples without *interpreting* the specific utterance the sys-

tem is reacting to, and without taking into account the *context* in which the input was originated. Hence the double blandness.

Everything Depends on Context

To see why context is important, consider the sentence

> After Paris we need to get to Abbeville before nightfall.

This sentence might be used, in a first context, as part of a conversation between two British tourists planning a trip from Paris to Normandy, where they are discussing the closing times on Somme battlefield memorial sites. On the other hand, it might be used in a second context as part of a conversation between two Oklahoma truck drivers, discussing potential traffic holdups on Interstate 49 on the way from Paris, Texas to Abbeville, Louisiana. In both cases, the utterance in question involves multiple spatial and temporal contexts, including in both cases spatial and temporal contexts embedded inside each other. In the one case it is set in a social context determined by British speakers of a military tourism idiolect using a dialect of British English. In the other case its social context is associated with the use of a trucker idiolect by speakers of a dialect of American English. In both cases we have in addition a planning context determined by the intentions of the speakers involved, giving the dialogue in each case an immediate relevance and utility. In an urgent planning context (one of the speakers has discovered that there has been a large pile-up on the road from Paris to Abbeville) this may add a moment of urgency to the dialogue, resulting in one or both speakers adopting an urgent or angry or pleading tone. Or adding new gestures, or facial expressions, or attempts to grab his interlocutor and shake him round the shoulders, leading in turn to new contexts: of protesting on the part of the one who is grabbed, or of attempts to calm down the one who is doing the grabbing.

In any case, not bland.

Ungroundedness

It is context that gives ground to dialogue, sets the scene for interpretation by each dialogue partner of what the other has said and for both dialogue partners to use the dialogue as a means to realize their intentions.

In XiaoIce and in all similar applications no attempt is made to *interpret* utterance inputs. Interpretation is indeed impossible in the absence of any consideration of context. Rather, the machine simply tries to copy in its responses those utterances in the training set which have immediately followed syntactically and morphologically similar input symbol sequences in the past. Because utterances are decoupled from context, responses appear ungrounded.

Attempts to improve matters by the developers of XiaoIce using what are called "Grounded Conversation Models", which try to include background or context-specific knowledge, have not solved the problem. For the attempt to take context into account faces a sampling problem. While we can gather large amounts of data for contexts in general, as soon as we attempt to collect a representative sample of data relating to dialogue in some specific context, we find that this is impossible. [4] Available samples that could be used to train the algorithm are both too sparse and unable to represent the variance in the sorts of genuine human conversation that take place in that context.

A further problem faced by neural chitchat applications is that they create ever more incoherent utterances as a dialogue develops over time. This is first of all because they cannot keep track of the dialogue as it becomes its own context – for example when the grabbee, in the above scenario, tells the grabber that their conversation is at an end.

---

[4] With a few exceptions. See the appendix to Jobst Landgrebe and Barry Smith, "Making AI Meaningful Again", *Synthese* 198 (March 2021), pp. 2061–2081.

And secondly it is because the datasets they are trained from are actually models of inconsistency due to the fact that they are created as mere collections of fragments drawn from large numbers of different dialogues. Attempts to alleviate the problem using "speaker" embeddings or "persona"-based response-generation models are able to improve the situation slightly,[5] but they do not come close to ensuring realistic, convincing conversations.[6]

Given that machines of the mentioned sorts can neither interpret utterances by taking into account the sources of variance, nor produce utterances on the basis of such interpretations, the approach cannot be seen as promising when it comes to conducting convincing conversations.

Therefore when Turkle writes that where people confront machines that seem to think this suggests a new way for us to think, she is wrong on two fronts: first, neural chitchat algorithms do not seem to think; what they *do* is compute output behaviour generated to optimize a measure of a certain sort; for what they *seem to do* in the eyes of their users we need a whole new word. And second: what they do *not* do is to suggest new ways for us to think.

[5] Jianfeng Gao, Michel Galley, and Lihong Li, *Neural Approaches to Conversational AI*, 2018, arXiv abs/1809.08267, section 5.3.
[6] Li Zhou et al., *op. cit*.