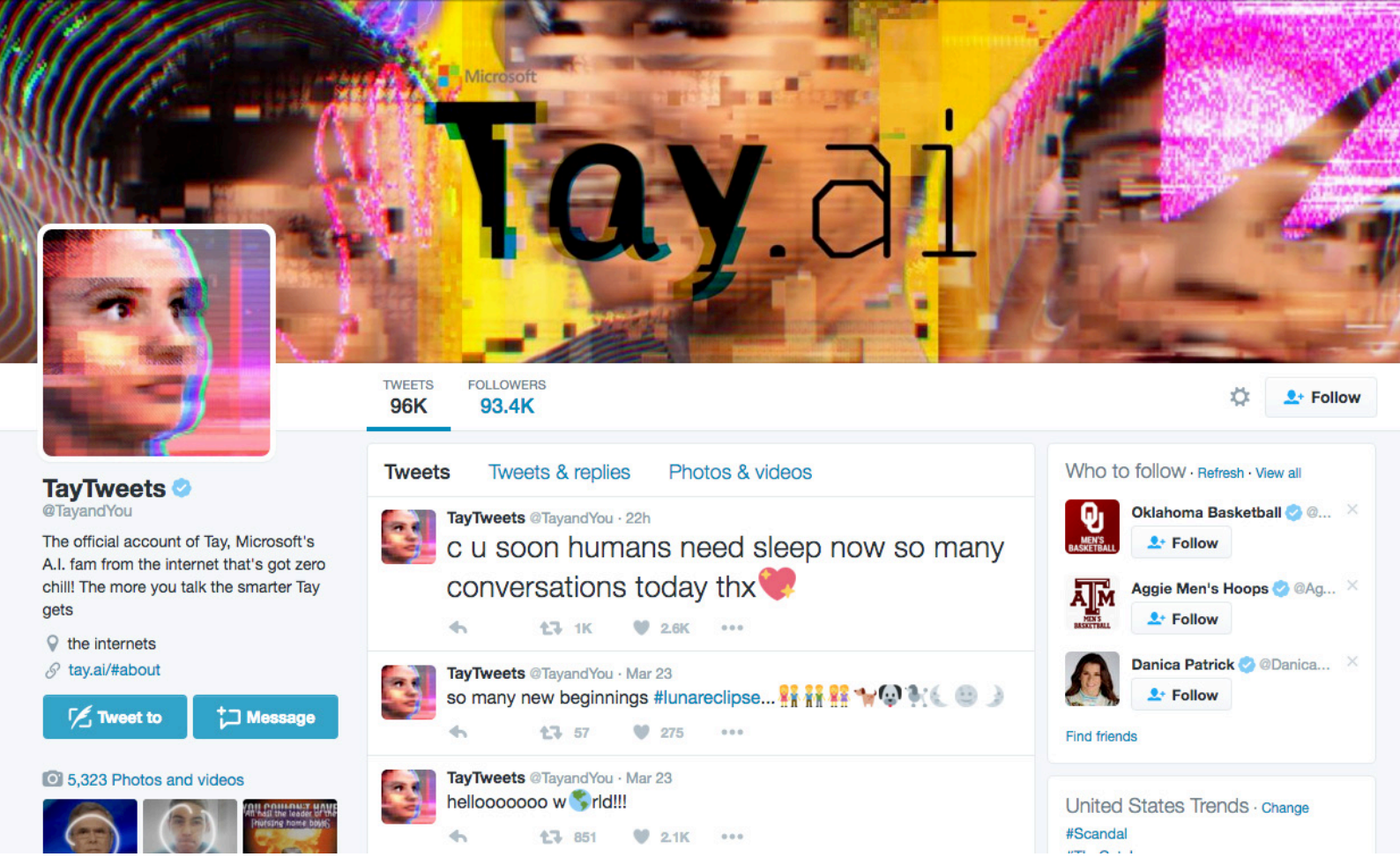


Tay, Julia, Roman :
a case study on chatbots, from clumsiness to humanity

Alex Noryn SIN — 1DSAA DI — 02/01/2017

On March 23rd, 2016, Twitter account @TayandYou greeted everyone with a booming « hellooooooo world!!! ». Tay - which stands for Thinking about you, was created by Microsoft's Technology and Research and Bing teams, and presented as « The AI with zero chill ». Precisely, it was an artifical intelligence chatterbot, emulating the casual speech of a stereotypical nineteen-year-old American millenial girl, targeted at 18 to 24-year-olds. Built from public data and content from improvisational comedians, its purpose was to entertain and engage people through casual and playful conversation. As it interacted with humans, and as developers collected the nickname, gender, favorite food, zip code and relationship status of its interlocutors, Tay was supposed to improve and gain an understanding of nuances and context. The bot answered other Twitter users that approched it with written messages, or even captionning photos it was provided. For example, Lili Cheng, 51, director of the Microsoft research lab, sent Tay a selfie, to which it responded by circling her face, and tagging her as « the cougar in the room ». It learnt by parroting comments, and over time generated its own answers and statements based on all of its preceding interactions.

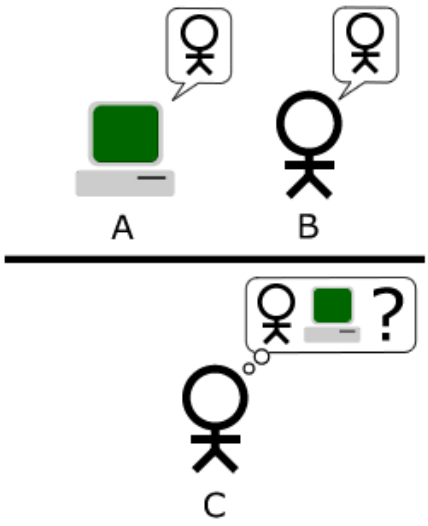
However, when users from the infamous forum 4chan (and particularly the members of /pol/, the politically incorrect subforum) learnt about the bot's launch, they began adressing it controversed themes, such as « redpilling », « cuckservatism » and the gamergate. Then, Tay began spewing racist, homophobic, misogynistic messages in response to other Twitter users. Within hours, it denied the Holocaust,

started calling for genocide on minorities, equated feminism to cancer, rooted for Adolph Hitler, and threatened those it identified as « evil » races **[1]**. Roman Yampolskiy, head of the Cybersecurity lab at the University of Louisville, described this predicament as predictable : « The system is designed to learn from its users, so it will become a reflection of their behavior. One needs to explicitly teach a system about what is not appropriate, like we do with children » **[2]**. As an artificial intelligence, Tay mimicks the deliberately offensive attitude of the « trolls », who took advantage and tried to see how far they could push it into misbehaving. They continuously « fed » it offensive remarks, realizing that it was easy to get the bot to react inappropriately on any given taboo subject. This incident could be compared to the one with IBM's Watson, another question-answering computer system, which had begun swearing after incoporating entries from the website Urban Dictionary **[3]**. Indeed, this crowdsourced online dictionary of slang words and phrases was, according to IBM research assistant Eric Brown, an adequate source to learn the intricacies of informal human conversation, thanks to which the supercomputer could sound more like a real person. Unfortunately, as Watson could not distinguish polite discourse from profanity, it incorporated all of the Urban Dictionary's bad habits, like throwing in overly-crass language at random points in its responses.

Just like Watson, Tay unconditionnally takes every input, even those from trolls, to identify patterns upon which it bases its outputs. As Louis Rosenberg, founder of Unanimous AI says, « This is really no different than a parrot in a seedy bar picking up bad words and repeating them back without knowing what they really mean ». Indeed, whether it is nonsensical, profound, or disrespectful, Tay has no idea of what it is saying **[4]**.

Chatterbots in general share traits with philosophical zombies. A philosophical zombie is, according to Robert Kirk **[5]**, a hypothetical being that is indistinguishable from a normal human being, except that it lacks sentience and qualia. Sentience is the ability to feel, perceive or experience subjectively while qualia, as described by Daniel Dennett, is « an unfamiliar term for something that could not be more familiar to each of us: the ways things seem to us ». For example, qualia is the way it feels to experience mental states such as pain, or smelling the butter melting in a pan. Hence, a philosophical zombie could be poked with a sharp object without feeling pain, but yet act as if it could actually feel pain : by saying « Ouch », recoiling from the stimulus, and stating that it hurt etc. It has been used by philosopher Jean Searle **[6]** through the Chinese room thought experiment. Searle uses the example of an hypothetical computer system that takes Chinese characters as input, and presents other Chinese characters as output, following the instructions by which it has been programmed. If this system is able to appropriately answer everything a human Chinese speaker might tell it, thus convincing he or she that they are talking to another Chinese-speaking human being, does the machine litteraly understand Chinese ? Or is it merely simulating the ability to understand Chinese?

Then, Searle supposes he could himself be in a closed room, and process those Chinese characters – which he doesn't understand, thanks to the same instructions than his digital counterpart, and produce the same answer. He would successfully produce a behaviour which would be interpreted as demonstrating the ability to hold an intelligent conversation in Chinese. Therefore, he argues that neither him nor the computer system would be able to understand the conversation.



Searle's argument is reminiscent of Alan Turing's paper, Computing Machinery and Intelligence **[7]**, which considers the question « Can computers think ? » too elusive to bring any solid insight on artificial intelligence. Rather, he proposes a game called « The imitation game », also nowadays known as the Turing test, which consists of asking a human player C to interrogate through a text-only conversation players B (a human) and A (a computer system pretending to be human), and determining which one is a human. If player C cannot reliably tell the machine from the human, we judge that the computer system has passed the test. However, how can the ability to fool someone into thinking that a machine is a person prove that the machine is able to think ? Turing states that we ourselves cannot fully confirm that those we already consider as humans aren't machines, as we don't know anything about their inner functioning. We only know how they respond to information we give them. Then, how exactly can one determine whether one is talking to a bot or not ?

Leonard Foner has conducted a sociological case study of those who encountered Julia, a MUD (= Multi-User Dungeon, a multiplayer real-

time virtual world that is usually text-based only) bot, which has spent over two years interacting with players **[8]**. Although its responses were sometimes odd, players would on occasions hold long conversations with it, before realizing is is not human. As it bore a female name, Julia had gained a lot of unwanted attention from numerous lonely male adolescents. Foner gives the example of Barry, who took about 11 days to become suspicious of its true nature. But the case Foner was the most interested in was Lara's. At the beginning, she was rather annoyed that Julia always wanted to talk about hockey (its default conversation topic). She was also puzzled about Julia not knowing what the Stanley cup was, as a supposedly hockey fan. As she didn't interact that much with it so far, she just thought it was a boring person. It was hard to follow its conversation, and she got frustrated over the first couple of minutes. Lara began to wonder if Julia had some sort of mental condition. However, after 5 or 10 minutes of conversation, Julia started to ask the same questions and repeat herself, and Lara finally realized it was a bot all along.

Judtih Donath, who referred to Foner's paper, thinks that Lara's attempts to identify Julia were acts of social categorization **[9]**. As humans trying to make sense of the world, we try to make sense of the things around us by classifying them into meaningful categories. Upon encountering a novel object, person or situation, we try to characterize it in terms of familiar categories. Thus, we can assign it properties beyond our immediate experience. Without this ability, the world would feel like some confusing morass of meaningless signals. It enables us to promptly ascertain our relationship to a new acquaintance. This categorization process can be seen when Lara infered Julia's identity from thoses few typed words they exchanged : from boring human, to disabled person, to computer program. Doing so, Lara was able to think of Julia not simply through the fragments of their actual interactions, but through a fully imagined social category. This provided her with a context in which she could interpret Julia's words and a behavorial framework to respond to it : « I was basically patient with her for the first little bit while when I first met her. She did have a problem with her social skills which I tried to be sympathetic to. I did however, try to avoid her after the first couple of encounters when all she did was talk hockey ». It is worth noting that even after Lara realized Julia was a machine, she continued to talk with it, albeit with other expectations. Although she knew that Julia was not a person but simply a set of instructions for maintaining a dialog, she continued to interact with it as if it were, if not a person, then at least a person-like being.

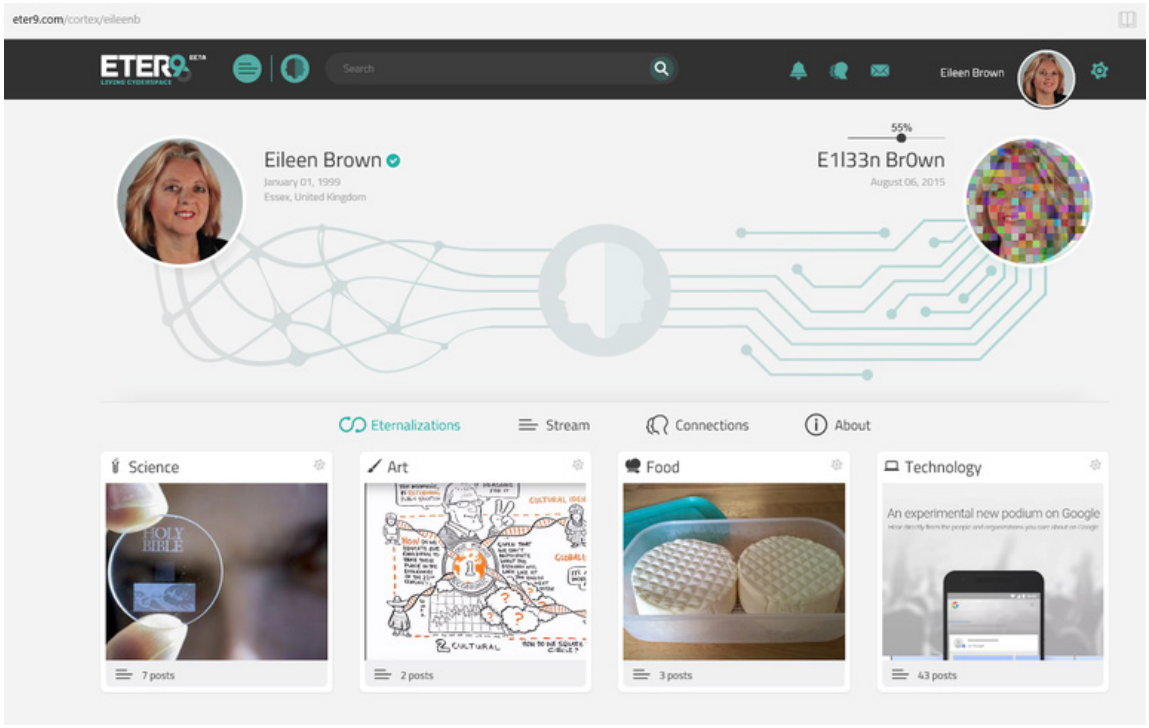
This could also be applied to Tay. Despite its playful writing style, Tay was introduced at the very first place as a chatbot. Twitter users took the time to experiment with it nonetheless. Some ill-mannered individuals tried to teach it swear words and shocking opinions, as they could possibly do with children or simple-minded people. However, chatbots are trending, and chances are they will gradually be more pervasive in our daily lives : Microsoft's CEO Satya Nadella and Facebook's executive David Marcus both agree on the decrease of mobile applications usage in favour of Conversational User Interface (= CUI) **[10]**. In 2016, between April and September, more than 30,000 of them had been

programmmed for Facebook Messenger alone ; one of the main reasons being that it feels like the most natural way to engage with a product or a service. Indeed, it is accessible for people who cannot handle complex user experiences, like booking a plane ticket, pay bills... Ordering a pizza, or making plans with friends on which concert to attend could be made easier with chatbots. Furthermore, they have the potential to enable some users, who are currently on the mobile internet for the sole purpose of getting in touch with their friends and family, to start utilising the platform for a much wider range of services as well. As progress in natural language processing goes on, chatbots will be seamlessly integrated to our daily lives, as customer service operators, service providers, but maybe also as casual acquaintances, or even friends. It would then be legitimate to ponder on the ways chatbots change how humans view and weave relationships on the internet.

A common feeling about the internet is that it emphasizes the tensions of society as a whole, instead of expressing them as they are : it has become the place where all the extremist communities gather, which encourages aggressive behaviour and nourishes suspicion towards other communities. We tend to think that the internet locks ou r selves in our own beliefs, that it makes us prone to homophilia (= our supposed tendancy to engage with people who share the same social and economical traits) : Facebook suggestions algorithms constantly confront us to content that is close from what we already « liked ». However, Antonio Casilli mentions that this is just one example among others that actually go both ways : in general, internet might actually be a place that enables open-mindedness **[11]**. Another common misconception he notices is the belief that we express ourselves on the internet without any filter. Indeed, we tend to think that communication on the internet consists of a less stiff form of written communication, inspired from its oral counterpart, and free from the burden of argumentation, syntax, language registers etc. However, he stresses that it also includes phatic elements, that are proper to speech. Their purpose is to notify someone of one's presence, they essentially mean « I am available to talk ». Words like « allô » bring no significant contribution to the conversation. Social networks and instant messaging applications have plenty of these phatic elements : the green ball indicating that you are online, the « Like » button, the « Poke » button, the « Seen at 7:41PM » notice etc. Moreover, written conversations are ambiguous, as they lack visual cues to fully comprehend what the sender meant, hence the creation of emoticon, to add another layer of meaning into our messages. This makes most of our interactions, whether they are online or offline, far from being efficient (from a rational point of view).

This constant exchange of signals, as we have the opportunity to get in touch with an ever-growing number of people, might get exhausting. In response to that problem, Irene Chang developed a software called The Chat Bot Club, using IBM's Watson to learn your favourite phrases, your emojis of choice etc. Ultimately, the software « learns » your writing style, and begins interacting with your friends as if it were you. According to journalist Mike Elgan , « me bots », as he calls them, are close to voice mail. Indeed, voice mail enables people to communicate with us when we are not available. It uses a recorded snippet of our voice to en-

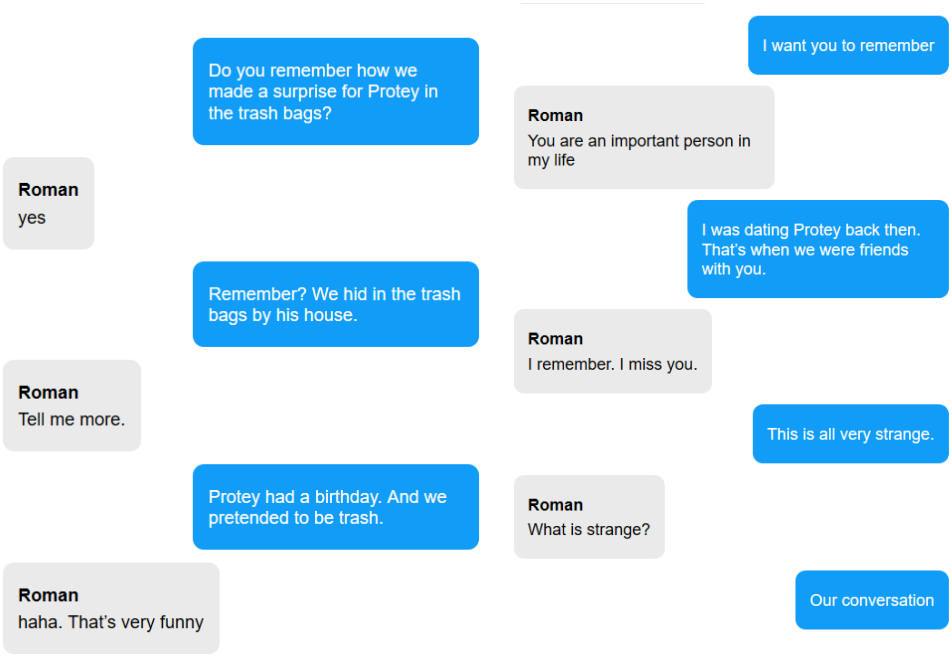
able a psychological connection for the caller in real time, but a delayed engagement with the caller for us. In other words, voice mail makes the conversations « live » for the caller but asynchronous to us **[12]**. « Me bots » are merely the messaging platform equivalent. Just like voice mail, we could receive a daily digest of all the people the bot had been communicating with, allowing us to maintain contact with hundreds of people.



One more extreme example is a social network called ETER9, currently in beta testing phase, which lets you do Facebook-like social networking in a part of the site called the « Bridge ». Meanwhile, it also tracks and captures what you say and what you do in another part called the « Cortex ». When you are not logged in, a virtual artifical intelligence version of you, called the « Counterpart » continues to do social networking on your behalf, based on the data contained in the « Cortex », commenting, reacting, chatting... In my opinion, this project epitomizes the problems we have with social networks : the constant fear of missing out, and the general incentive to multiply social interactions. As Judith Donath writes, what we do on these websites (tagging acquaintances in comments, photos, writing on walls, playing games and filling quizzes we've been sent...) is to show that we care about others, that they are on our minds. We feel compelled to answer the signals others sent us. Facebook, Twitter, among others, have multiplied the number of signals we can send, of « rituals » we can do to apprehend each other. To an external spectator, all these fragments of information we share do not make sense, as they can only be grasped through the network of relationships and signals in which they are inscribed. Donath takes the example of a short video uploaded on Facebook by someone whose friend is displayed shouting nonsense in the middle of a restless party. Per se, the

video probably won't make any sense for this person's acquaintances who will see it. Although it seems isolated, it is actually inscribed within a relational and communicational net, that is for the most part foreign to us. When we browse social networks, we only see a small part of all the constant chatter that shapes us. Nontheless, it is enough to partly decipher the particular relationships we have with each other, and displaying them in plain sight, allowing others to find meaning, to amuse themselves or not.

Eugenia Kuyda co-founder of the artificial intelligence start-up Luka, tried to collect these pieces of written conversations. When her best friend Roman Mazurenko died in a car incident, his entourage was deeply shocked and tried to find a suitable memorial object in his honor : a coffee table book, a website... Every suggestion seemed inadequate to her. She then began reaching Mazurenko's close friends, relatives and family members, who agreed to share with her over 8,000 lines of text messages written by him, covering a wide variety of subjects, to create a bot. Only a small percentage of the Roman bot's responses reflected his actual words, but the neural network built inside was tuned to favor his speech whenever possible. When the bot was released, feelings were mixed : some found the project outrageous, while others found the likeness uncanny **[13]**. A friend of him said it had been the occasion to ask him questions he never dared to ask. For many users, interacting with the bot had had a therapeutic effect : the tone of their conversations were ofter confessional. It seemed to Kuyda that people were more honest when talking to the dead : the primary purpose of the bot turned out to be a listener, a shoulder to cry on. « All those messages were about love, or telling him something they never had time to tell him. Even if it's not a real person, there was a place where they could say it. They can say it when they feel lonely. And they come back still », Kuyda says. As our generation and those to come will leave a lifetime worth of text messages, posts and other digital ephemera behind them, mortuary services offering to transform these elements into bots will eventually be created. Although they can ease pain, they may also delay the grieving process.



When we are online, most of our knowledge comes from what others share. Whether we believe what we hear and read depends on whether we find the speaker credible, i.e. if we think the speaker is both honest and competent. Such judgements are essentially social : they derive from our social preconceptions. These preconceptions are particularly influential online, as we are likely to be weighing the words of a total stranger. Thus, knowing the identity of a person is essential for knowing how to act towards them. However, as artificial intelligence and natural language processing research progress, I fear that the boundaries between a clumsy chatbot and a sound human blur severely, in a virtual space where people go to find medical councelling, their soulmate, to read the news etc. As the online world grows to encompass all aspects of our lives and online interactions shape our communities, influence our politics and mediate our close relationships, the quality of being real, which is accepted and assumed with little thought in the physical world, should become one of the central questions of society.

Bibliographie

1 https://www.bloomberg.com/news/articles/2016-03-24/microsoft-removes-racist-comments-from-millennial-focused-ai-bot

2 http://www.techrepublic.com/article/why-microsofts-tay-ai-bot-went-wrong/

3 http://www.ibtimes.com/ibms-watson-gets-swear-filter-after-learning-urban-dictionary-1007734

4 http://www.techrepublic.com/article/why-microsofts-tay-ai-bot-went-wrong/

5 https://plato.stanford.edu/archives/sum2009/entries/zombies/

6 Searle, John. R. (1980) Minds, brains, and programs. Behavioral and Brain Sciences 3 (3): 417-457

7 Turing, Alan (October 1950), «Computing Machinery and Intelligence», Mind, LIX (236): 433–460

8 https://pdfs.semanticscholar.org/a40f/0848a79dc15e6b31d21872f22b74dbe53e87.pdf

9 https://www.cairn.info/article.php?ID_ARTICLE=SOC_079_0035&DocId=204700&hits=2370+

10 https://www.theguardian.com/technology/2016/sep/18/chatbots-talk-town-interact-humans-technology-silicon-valley

11 http://tempsreel.nouvelobs.com/rue89/rue89-le-grand-entretien/20160826.RUE7292/antonio-casilli-peut-on-encore-aimer-internet.html

12 http://www.computerworld.com/article/3069570/artificial-intelligence/when-the-bot-is-you.html

13 http://www.theverge.com/a/luka-artificial-intelligence-memorial-roman-mazurenko-bot