Tay is you. The attribution of responsibility in the algorithmic culture.

Sara Suárez-Gonzalo*, Lluís Mas-Manchón*, Frederic Guerrero-Solé*

*Universitat Pompeu Fabra, Spain

Abstract

Social media have changed the communication practices by creating an acute need for continuous interaction. The use of social chatbots is growing as an effective way to communicate with publics. Bots have become social actors and then, someone must account for their actions. Since responsibility is bounded to agency and rationality, it cannot be directly attributed to bots. Who should be held responsible for non-human beings' actions, particularly when the consequences of these actions are negative?

We address this controversy from both theoretical and empirical perspectives. Firstly, we discuss the adequacy of the notions of moral responsibility and accountability regarding non-human artificial agents, as they are ruled by complex, intentionally opaque and unpredictable interactions and processes. We do it from the two approaches currently predominant: context-dependent and structuralist. Secondly, we draw on the assumption that the failure of a computer system is an opportunity to gain knowledge about the interested powers behind its design and functioning. Then, taking the concept of media frame as an implicit way of spotting the agent of the story, we perform an exploratory analysis on how responsibility was attributed by the media in the paradigmatic case of the transformation of Tay, a chatbot launched by Microsoft in 2016, turned into a racist, Nazi and homophobic hate speaker.

Our results illustrate the difficulties media experienced in consistently attributing the responsibility for the chatbots' malfunction. Results show the discourse is, in general, simplistic, non-critical and misleading, and tends to depict a reality that favors business's interests. We conclude that, while all the actors interacting with the chatbot share the responsibility of its actions, it is only Microsoft who must account for these actions, both retrospectively and prospectively.

Keywords: Chatbots, big social data, artificial intelligence, Twitter, algorithmic culture, accountability, attribution of responsibility, Tay, hate speech.

Submitted: 22nd November 2018

Accepted: 31st March 2019

How to quote this article:

Suárez-Gonzalo, S., Mas-Manchón, L, Guerrero-Solé, F. (2019). Tay is you. The attribution of responsibility in the algorithmic culture. *Observatorio*, 13(2), 1-14.

Introduction

Social media have challenged the traditional one-to-many communication paradigm (Castells, 2009; van Dijk, 2013) and have lead institutions and organizations to adapt to a new environment in which personalized interactions with thousands, even millions of people, are required (Neff & Nagy, 2016). Since human actors cannot perform such a large amount of interactions, companies have developed the so-called chatbots, technologies based on big data analytics and machine learning that conversationally interact with users, mainly on social networks (Ferrara, Varol, Davis, Menczer & Flammini, 2016). Social network sites (SNS) are particularly suitable platforms for the collection of great amounts of social data (Tufekci, 2014; boyd & Ellison, 2007) that can be processed in real-time by machine learning technologies and become a source for machine-generated content (Nichols, 2010). This content is the basis for personalized interactions carried out by chatbots that emerge as relevant non-human social actors in SNS.

Scholars have begun to scrutinize the potential negative impact of bots on society. From a critical perspective, these technological artifacts are characterized by the opacity and black-boxed nature of the algorithms that rule their actions (Pasquale, 2015). Due to the enormous complexity of such algorithms, scholars have put the focus on cases of unsatisfactory functioning as a way to gain knowledge about their inner characteristics and effects (Karppi & Crawford, 2016). This knowledge is essential when dealing with highly problematic concepts such as agency, accountability or the attribution of responsibility in the context of the algorithmic culture (Hallinan & Striphas, 2015).

The aim of this work is to contribute to the debate about chatbots' acts responsibility and accountability from both theoretical and empirical perspectives. On the one hand, we examine the adequacy of traditional notions of responsibility and accountability in the context of the algorithmic culture. On the other, this research explores the discourse of media outlets in such a case in which computational systems fail. We perform an exploratory frame analysis (Entman, 1993; Semetko & Valkenburg, 2000) of the news stories on Tay's failure, a Twitter chatbot launched by Microsoft Corporation in 2016. Since after a 24 hours period of interaction with Twitter users the bot's messages turned into racist, homophobic and sexist hate speech, we pay a special attention to how media attributed responsibility of this misbehavior. Finally, we discuss whether the algorithmic culture is influencing the construction of media discourse about chatbots' agency and responsibility.

From user to machine generated content

The rapid development of information technologies at the end of the 20th century made the volume, variety and velocity of data grow dramatically (Laney, 2001). In 1997 Michael Lesk predicted that in the year 2000 there would be enough space of storage to register almost any expression of human activity. He also warned that such a large amount of data could not be inspected by humans in the future, and a continuous automatic evaluation would be a requisite to decide what portions of information should get "the precious resource of human attention" (Lesk, 1997: 9). Two decades later, Lesk's predictions have been widely confirmed, and nowadays we live in a big data ecosystem (boyd & Crawford, 2011), being the *datafication* of everyday life an increasingly pervasive trend (Baruh & Popescu, 2015; Suárez-Gonzalo, 2017). In this sense, the size and complexity of datasets undermines humans' capacity to deal with data and to make sense of them (Baldi, 2017; Shalev-Shwartz & Ben-David, 2014). According to Hallinan and Striphas (2016: 119), a new algorithmic culture has emerged, which lies in "the use of computational processes to sort, classify, and hierarchize people, places, objects, and ideas, and also the habits of thought, conduct, and expression that arise in relationship to those processes'.

Under the name of artificial intelligence (AI), highly sophisticated computational techniques have been developed to replace humans in relevant activities (Carbonell, Michalski & Mitchell, 1984), such as content creation, information management and distribution and decision-making (Marsland, 2015). Machine learning is the current paradigm in the field of AI (Bostrom, 2015) that trains computers to learn. It allows computers to adapt their actions to the changes in the environment (Shalev-Shwartz & Ben-David, 2014), and to discover complex structures and patterns in high-dimensional datasets (LeCun, Bengio & Hinton, 2015). For Nath and Levinson (2014), machine learning can be understood as algorithms that constantly improve their outcomes by means of available data.

The nature of social chatbots

Social chatter robots (chatbots) are a particular case of this new generation of machine learning technologies that make use of social media data to generate natural language outputs and engage in conversations with human users (Griol, Sanchis de Miguel & Molina, 2014). They are, nowadays, an effective way to communicate with users (Chakrabarti & Luger, 2015). Due to the enormous amount of data spread by users, SNS have become particularly thriving ecosystems for the development of chatbots. In the past few years, chatbots have settled in social networks (Sandvig, Hamilton, Karahalios & Langbort, 2016; Ferrara et al., 2016) and are holding strategic roles in organizations' communication actions (Neff & Nagy, 2016). In this sense, communication technology corporations such as Apple, Samsung, Microsoft or Facebook have already developed their own chatbots. They have also played a relevant role in political events, such as the latest United States presidential election campaign (Kollani, Howard & Wooley, 2016).

However, chatbots can also contribute to amplify old biases in society and are acquiring perilous roles in public life (Caplan & boyd, 2016). Research has shown that can they lead to algorithmic discrimination (boyd, Levy & Marwik, 2014) and are capable of swaying public opinion (Marechal, 2016), perpetuating social damaging stereotypes (Sandvig *et*. al, 2016), destabilizing financial markets (Karppi & Crawford, 2016), or amplify the spreading of misinformation (Ferrara et al, 2016) and hate speech (Marwick & Lewis, 2017), among others. Additionally, corporations and governments foster the opacity of such algorithms through real secrecy, legal secrecy and intentioned obfuscation (Pasquale, 2015: 2).

Thereby, considering the complex and opaque nature of algorithms in chatbots, and the fact that machine-learning technology takes advantage of the contents published by users, a question emerges: who is responsible of chatbots' behavior when they fail?

The responsibility gap: attributing responsibility to artificial beings in a networked society

Responsibility has been traditionally bounded to actions with concrete intentions (Asaro, 2012; Hellström, 2013) and significant consequences (Fisher, 1999). Responsibility has been usually attributed to individuals, groups of individuals or institutions (referred as the "agent") when their actions have an effect on others (referred as the "patient") (Floridi & Sanders, 2004). The attribution of responsibility requires the agent to be rational, as well as to have intention and agency (Mitcham, 2014; Guilbeault, 2016). Consequently, responsibility establishes a link between agents and patients and organizes social relations.

Accountability is the assumption of responsibility by the agent. Bovens defines it as "a social relationship in which an actor feels an obligation to explain and to justify his or her conducts to some significant other" (2005: 184), especially when it comes to actions with negative consequences. Generally, accountability is part of the relationship between society and the state (Caplan & boyd, 2016). According to Rosenblat, Kneese and boyd (2014), it is fundamentally about checks and balances to power. It has a retrospective dimension (being blamed or punished for an action), which is the most commonly accepted, and a prospective one (defining obligations and duties related to that action). Both responsibility and accountability are dependent on identifying the agent of the action and whether her or his actual intention is rationally aligned with the consequences of this action (Groom et al., 2010).

In digital environments such as SNS, in which chatbots become algorithmically controlled actors that learn from others' behaviors, agency and intention are not easily attributed. Questions such as whether a chatbot can be considered as a rational agent or not, or its actions as intended, become problematic. Besides, in modern sociotechnical relations, tasks are distributed between human and non-human entities in a way that unpredictably affects each other. This makes it hard to identify the agent of a certain action.

As noted by Kroll et al. (2017), social bots are peculiar black boxes in which the inner workings are either too complicated or based on randomness, and so the outcomes become difficult to foresee. Following Matthias (2004), while the operators of learning machines are not capable to predict the future behavior of such machines, they cannot be held responsible for their actions. Therefore, intentionality, causality and the agent-patient link become hard to define. As the complexity and autonomy of learning machines keep growing, humans cannot continue to be directly responsible for them. In some way, humans lose control over them, and bestow the decision-making process to the machines themselves. In such cases, society must address what Matthias calls a "responsibility gap". Gotterbarn (2001) and Waelbers (2009) add two pervasive misconceptions about technology and responsibility that complicate the attribution of responsibility in such cases: the alleged ethical neutrality of technological artifacts; and the predominant reductionist understanding of responsibility, which only considers its retrospective conception.

All this complexity has lead several authors to notice that the case of artificial beings requires rethinking the very concepts of responsibility and accountability so to make them applicable to networked environments (boyd, Levy & Marwick, 2014). Several authors have also emphasized the need for algorithmic transparency (Kemper & Kolkman, 2018), external control of algorithmic processes (Pasquale, 2015; boyd, 2016) and to design them according to previously agreed values (boyd, 2016) such as the five defined by Diakopoulos and Friedler (2016): responsibility, explainability, accuracy, auditability and fairness.

There are interesting contributions regarding the intentionality and agency of artificial beings, which can be classified into two main approaches regarding the attribution of responsibility and the accountability of chatbots's actions. One is the context-dependent approach: responsibility corresponds to the environment with which the bot interacts, and so its assumption disperses among all actors. The other is the structuralist approach: assuming the responsibility of bot's actions corresponds to the forces involved in the design and management of the bot.

Regarding the context-dependent approach, Floridi (2014) poses that when an artifact learns from the context in which it performs, intentionality spreads through the different relationships and outer interests involved in the interaction. In this same line, Introna (2014) draws on Foucault to develop an interactional concept of intentionality, defined as the inseparable interaction between technical artefacts ("dispositifs") and humans. According to van Dijk (2013), given that the environment as a whole can be considered as the input of social bots and also that it is based on simple interactions that define each other's identity, agency could be considered also a networked concept. Similarly, Neff and Nagy (2016: 4916) develop the concept of symbiotic agency, defined as: "what users, actors, and tools do when interacting with complex technological systems […] In other words, what people say about bots influences, what people can do with them and what capacities the bots have for social action".

Regarding the structuralist approach, Johnson (2006) argues that although computer technologies are not intentional, they do have intentionality, always related to that of their designers and users. She emphasizes the need to consider the social, political and institutional forces interested in shaping technological developments. In this same line, Ford, Dubois and Puschmann (2016) pose that chatbots' actions should be accounted by a set of different interests participating in the co-creation of these chatbots. Through a quantitative and qualitative study on Youtube, Rieder, Matamoros-Fernández and Coromina (2018) highlight the intricate mesh of mutually constitutive agencies that play a role in algorithm's functioning. Rieder (2018), moreover, examines the relationship between governmentality and computing and notes the importance of dealing with computers as political tools in the hands of interested actors or think tanks. Murthy *et al.* (2016) note that bots are created by social, political and economic systems of power (an idea also supported by Karppi & Crawford, 2016).

@TayandYou, a paradigmatic case of study

On Saturday, March 23rd 2016, Microsoft launched Tay*,* a new chatbot on Twitter. The bot was designed to simulate a young American millennial girl, with the purpose of informally interacting with Twitter users, millennials preferably, and conduct research on conversational understanding. As stated by Microsoft (2016), Tay was built "by mining relevant public data and by using AI and editorial developed by a staff including improvisational comedians". In order to have the most personalized and satisfactory experience, Microsoft warned users that the more they chatted with Tay, the smarter she would get. However, hundreds of users started tweeting with the chatbot by making misogynistic and racist comments. Because of its machine-learning nature, Tay's messages, tone and vocabulary also became racist and misogynistic dramatically. A few hours later and as a result of Tay`s inappropriate behavior, Microsoft removed the chatbot arguing that it suffered a malicious attack (Lee, March 25th, 2016). Three days later, on March 30th, Microsoft launched a renewed version of Tay. However, its behavior soon became even worse than before, and Microsoft definitively removed the chatbot from Twitter.

Media framing of Tay's event

As pointed out by Druckman and Bolsen (2011), public opinion plays a critical role on how people perceive emergent technologies. In this regard, the way media portrays a new technology is a definitive factor for its success. Spicer (2005) pointed out that the way complex digital technologies will be used is shaped during their process of social inclusion by political and economic forces. Stahl (1995), for his part, conducted a major study on *Time*'s framing of the first IBM personal computers. Results show that magical and religious language was commonly used in news media as a plan for legitimizing computers' black boxed condition. Besides, he argues that machines are frequently portrayed (antrophormized) as the active partners in human-computer relationships, making people feel powerless facing technology. Stahl concludes that, since not all social groups are equally able to define new technologies, media tend to stabilize and close the technological business' frame. That is: they promote business' definitions of technology. As noted by Puschmann and Burgess (2014), media discourse on science and technology usually tends to overgeneralize and subjugate the reality to power disputes. Campbell (2010) maintains that risk has also been frequent in media representations of emerging technologies, in particular when these technologies challenge the stability of other sociotechnical discourses.

Media outlets immediately reported on the failure of Tay. Assuming that media discourse depicts a particular understanding about artificial intelligence, machine learning, and chatbots, our main aim is to perform an exploratory analysis of how media framed and attributed the responsibility of the transformation of Tay. As defined by Entman (1993), framing involves selection and salience to prescribe and promote interpretations and evaluations of issues in media. Frames draw attention toward certain aspects of reality while marginalizing others (Lawrence, 2000). While the attribution of responsibility is conceptualized as a process of explicitly spotting the primary agent of certain phenomenon (the ultimate cause), framing is well known theory on how the media shed light into some direction regarding any stories' agents, hence responsibility can only be derived from framing implicitly. Thus frames must just be considered as a sort of premise for the attribution of responsibility. We will not go deeper in this insight.

Many scholars in the communication field have proposed diverse taxonomies of media frames. In particular, Semetko and Valkenburg (2000) identified the five prevalent frames in previous researches on news media content and systematized their identification in their classic work on European politics. These frames are: F1. *Conflict*: emphasizes the existence of conflicts between individuals, groups, or institutions as a means of capturing audience interest (it can induce public cynicism and mistrust); F2. *Human interest*: brings a human face or an emotional angle to the presentation of an event, issue, or problem: is supposes an effort to personalize, dramatize or "emotionalize" the news in order to capture and retain audience interest; F3. *Economic consequences*: reports an event, problem, or issue in terms of the consequences it will have economically on an individual, group, institution, region, or country; F4. *Morality*: puts the event, problem, or issue in the context of religious tenets or moral prescriptions, often by means of an indirect reference. It may contain moral messages or offer specific social prescriptions about how to behave; and finally, the one which they identified as the predominant one, F.5. *Attribution of responsibility*: presents an issue or a problem in such a way to attribute responsibility for its cause or solution to either an institution, individual or group. It encourages people to offer individual-level explanations for social problems.

Semetko and Valkenburg elaborate on a deductive, rather than an inductive (Gamson, 1992), approach to framing. This deductive approach involves having a clear idea of the types of frames that are likely to appear in the news and, afterwards, quantify them in the sample of news. Unlike the inductive approach, which is arduous to apply as it involves analyzing the news with an open view, the deductive method is easily replicable. Because of that, it has been employed by a multitude of researchers, especially in relation to media news on political issues and crisis communication (Coman & Cmeciu, 2014; An & Gower; 2009).

In view of previous theoretical considerations and the role played by media outlets in shaping public opinion, we tried to answer the question about how news media framed Tay's failure, and how they attributed the responsibility of this failure. For this purpose, we performed an exploratory research by collecting and analysing a sample of news about the Tay event from April to November 2016, when the number of news stories about the case falls significantly. As for the sample selection, we draw on the ranking published by Comscore MMX Multi Platform of the most read digital newspapers in Spain during the period analyzed. Then, we gathered the news published by the seven generalist newspapers of this list that published two or more news fully dedicated to Tay's event during that period. Moreover, we added to the sample the two international online dailies (The Guardian, The New York Times) and the three technology newspapers (The Verge, Wired, ZD Net) of reference in Spain that published the greatest number of news about Tay during the period analyzed. Ten keywords in English and Spanish were employed to find the news in media's search engines: Tay, *Inteligencia Artificial*, Artificial Intelligence, AI, *IA*, Bot, Chat Bot, Microsoft, digital assistant, *asistente digital*. 56 news stories were finally collected from thirteen international digital newspapers. Six of them in Spanish (El País, El Mundo, La Vanguardia, ABC, Eldiario.es, La Razón, and The Huffintong Post – Spain Edition), and seven of them in English (The Huffington Post - UK Edition, The Guardian, The New York Times - International Edition, The Verge, Wired and ZD Net).

From the methodological point of view, our approach is deductive. A first reading of the news shows us that the media was likely to have adopted primarily or exclusively an approach focused on attribution of responsibility and conflict. Moreover, Tay's event is a political issue (it raises concerns on Nazism, racism, homophobia or sexism and involves questions regarding the limits of freedom of expression, or the regulation of big data and artificial intelligence technologies), and it is a case of crisis communication. On this basis, the objective of the analysis is to quantify the presence of the five frames defined by Semetko and Valkenburg (2000) in the media coverage of Tay.

The sample of news stories, then, was categorized by three different coders by means of Semetko and Valkenburg's classification (Krippendorff's alpha = 0.91), focusing on the way media outlets attributed the responsibility of Tay's failure and turning into a misogynist and racist chatbot. We also coded the actors involved in the event, the causes of the failure and its responsible, the consequences and the actors that were affected by these consequences.

Concurring with Semetko and Valkenburg (2000)'s study, the results of the framing analysis revealed that the attribution of responsibility was the main frame (one out of two news stories) used by media in depicting Tay's failure. This frame was complemented by the conflict frame in 8 out of 56 cases (14%). The third most used frame was that of human interest, while none of the news stories analysed used the frames of economic consequences or morality.

Framing and content analysis show that the event was depicted to shape public understanding about who is to be blamed for Tay's malfunctioning. Media outlets tended to represent the event through the following pattern: Twitter users (the agent) maliciously misused and attacked (causal contribution) a feeble and vulnerable chatbot called Tay (the patient) that had to be disconnected by its designer (consequence).

Results show that almost three out of four stories were focused on trying to identify the culprits of Tay's malfunctioning. A third of the news stories described an orchestrated attack from Twitter users, which abused Tay and led it to behave in an inappropriate manner as the cause of the event. Precisely Twitter users were identified in 40% of the news as the actor responsible (agent) of the incident, while only a 17% do it with Microsoft. On the other hand, 18% of the news reported the interaction between humans and Tay's software as the trigger of the fiasco, while 14% of the stories described the malfunctioning as a failure of Tay's machine-learning code.

The consequences reported were the following: in one out of three news stories, the consequence reported was the disconnection of Tay and the apologies given by Microsoft. By doing so, media assume the retrospective approach to attribution of responsibility as the only possible. One out of four stories (25%) reported that the main consequence of the failure was that Tay had become a mirror of the worst of humanity by "learning" how to be racist and misogynistic. Conversely, norms and risks were not relevant in media depictions of Tay's failure. By describing the action as an attack, and clearly identifying responsible and affected actors, media outlets stressed the existence of a conflict between Twitter users and the chatbot. Besides this, news stories are emotionally charged by depicting Tay as a person and making sensitive judgements about it.

Media tend to present the chatbot as the most affected actor (patient) in the event. A third of the news stories (33%) referred to Tay as a human being and a "baby robot" harmed by the abusers. Additionally, 9% of the stories pointed at AI as the one affected by the failure, and 5% of them at Microsoft. To sum up, the chatbot, its technology, and even its designers were presented as those ill affected by Tay's malfunctioning. Surprisingly, only 12% of the news stories considered ethnic and religious vulnerable social groups (such as black people or the Jewish community) and women offended by Tay's messages as those harmed by the incident. Finally, 9% of the stories points at Twitter users and humans in general as the injured party. There was no explicit reference to hate speech and its consequences over people, nor to legal issues.

Tay's event: the media deconstruction of reality

Media representation of Tay's event depicts a biased and misleading reality that concurs with the traditional mainstream media discourse on new technologies defined by Stahl (1995): it tends to stabilize and close the discourse of Microsoft. By presenting the event as an isolated phenomenon from any context, media do not contribute to people's media and technology literacy, nor to their social empowerment. Content and framing analysis of news showed a contradictory discourse: on the one hand, media personalized Tay and treated it as something capable to feel and suffer. On the other hand, Twitter users are dehumanized and found guilty on Tay's turn into a misogynistic and racist being. Media discourse, then, reinforces the idea that Tay failed because of Twitter users. Media referred to a retrospective accountability action performed by Microsoft (apologize) and, by assuming company's discourse, depicted a reality that favored AI business' —and particularly Microsoft's— interests. They give voice and credibility to the company, which, far from being affected, gains visibility and come out reinforced by positioning its discourse in the public sphere. In that regard, it should be noted that Tay's event coincided with the celebration of Build 2016, the annual congress of Microsoft Corporation. Consequently, 21% of news stories replicate literal ideas pronounced by Satya Nadella (CEO of Microsoft) in his opening speech at the Build 2016, about the future of artificial intelligence and his company's plans on chatbots. The most repeated one is the following: "We want to build technology that gets the best of humanity and not the worst". Likewise, the content analysis revealed that the most of the news explain the cause of the event as an orchestrated attack, an idea exposed by Peter Lee (March 25th, 2016) (Microsoft Healthcare's Corporate Vice President) in an official statement: "Unfortunately, in the first 24 hours of coming online, a coordinated attack by a subset of people exploited a vulnerability in Tay." On doing so, media repeats a discourse that goes in the best interests of Microsoft: to dwell on a conventional-retrospective approach to the responsibility of Tay's case, instead of a prospective one, while they blame Twitter users for it and present the fiasco as an isolated event.

There is a huge academic controversy among scholars in relation to the attribution of responsibility in algorithmically-controlled environments. Designers, users or both (when interacting) have been proposed as the presumed responsible for the punishable crimes committed by algorithms. However, media outlets were inclined to blame just one of the actors involved: the users. Users can be considered as instigators of

Tay's wrongdoings. They drove the chatbot to deal with specific issues, vocabulary and tone. In this sense, users persuaded Tay to be misogynistic and racist. However, this cannot be considered as the unique cause of Tay's failure. The issue stems from the fact that the algorithm was not properly designed to handle such a situation (Sandvig et al., 2016), although it seemed to be designed to simply replicate its interlocutors tone and vocabulary. As a non-rationale machine, Tay did not have the capacity of understanding what is right or wrong, but designers should have been aware of the potential harms of such a design. As noted by Diakopoulos and Friedler (2016), some a priori values should have been programmed in order to prevent the fiasco.

A remark has to be added in relation to hate speech and the attribution of responsibility. Hate speech increases social inequality, violate sensitivities and impose the domination of social groups over stigmatised others, but it could even drive the victims to fatal consequences (Zollo & Loos, 2017). Research has shown that hate speech and extremist ideologies are flourishing on the digital space because of the far-right media manipulation and spread of disinformation (Marwick & Lewis 2017, Matamoros-Fernández, 2017). Hate groups, bots, trolls and the dynamics of social networks are some of the main contributors to this phenomenon. Due to the seriousness of hate speech's social implications, it is everyone's responsibility to help eradicating it. Consequently, it makes no difference whether the messages that Twitter users actually addressed to Tay had a purpose (attack, persuade, play) or not. Users were responsible for their own messages and they should account for them, although not for those of Tay.

As pointed out by Karppi and Crawford's (2016) the failure of a computational system is an opportunity to gain knowledge about it and its social consequences. In this sense, Tay's malfunctioning must lead society to reflect on the potential harms of automated bots' behaviors, and to make clear the different responsibilities that have to be assumed by social actors, including users, designers and the owners of the platforms in which robots perform their actions. We must be specially concerned about the rise of hate speech and other unacceptable attitudes and behaviors, and force designers to prevent their algorithms from turning into Nazi, misogynistic and racist abusers. Finally, media depictions of algorithms' failures should include the complexities of algorithmically-controlled environments and foster the public debate about who is responsible for what.

Conclusions

Digital environments and, in particular, social networks have driven technology companies to design algorithms that make use of social big data and machine learning strategies to interact with a myriad of users (Neff & Nagy, 2016; Ferrara et al., 2016). However, algorithms such as those used by chatbots are playing a role in public life, so the responsibility for their actions must be taken by someone. Chatbots' very nature challenges traditional notions of responsibility and accountability. On the one hand, they lack of intentionality and agency, which are conceived as human capacities. On the other hand, randomness, complexity and opacity are well-known characteristics of the algorithms that rule chatbots. These characteristics pose a difficulty in the identification of causes and consequences of chatbots' behavior and in the attribution of responsibility for their actions in case of failure.

From a theoretical perspective, we have observed that two basic approaches to the concepts of attribution of responsibility and accountability stand out: the structuralist (Johnson; 2006) and the context-dependent

(Kroes & Verbeek, 2014). However, our exploratory research has shown that media discourse on Tay's failure was, in general, simplistic, non-critical and misleading. Although the main frame used by media to depict the event was this of the attribution of responsibility (Semetko & Valkenburg, 2000), the responsibility of Tay's conversion into a Nazi and misogynistic chatbot was attributed to Twitter users, who were described as abusers that took advantage of the machine-learning algorithms leaging its actions. Far from a structuralist or contextual approach, Tay was treated as a human entity, while users were, in general, vilified and dehumanized. This fact leads us to conclude that media depiction of Tay's event was highly biased, and that it reproduced the dominant discourse about technology, algorithms and chatbots. Media adopted Microsoft discourse by stressing that it was Tay and the company itself the ones affected by users' unappropriate interaction with the chatbot. In sum, media contribute to the construction of a friendly and neutral image of AI technology, with no responsibility to be held.

Finally, media and other social institutions should put pressure on tech companies and denounce the undesirable consequences of the opacity of their algorithms as well as to push them to be accountable for the actions performed by their artifacts. As both our theoretical review and frame analysis reveal, there is a huge controversy about who is to be blamed by machine-learning algorithms misfunctioning. Media should contribute to the debate by publishing critical approaches and explaining to their audiences how complex attributing responsibility is in an algorithmically controlled environment. In addition, the sociotechnical system where bots function and interact should also be made comprehensible.

Considering the complexity of social bots' outputs formation, there are, at least, two main ideas related to responsibility and accountability that should be transmitted by the media in cases such as Tay's. Firstly, while the responsibility for the bot's behavior belongs to the whole environment involved in bots' development and functioning, the accountability belongs only to those involved in its development and design. That is, on the part of responsibility: developers, designers (including those who decide bots' type of learning and the environment in which it is inserted); those interacting with the bot, the interaction and the environment itself. On the part of accountability, developers, designers and those who lead the process of insertion of the bot. Secondly, it seems necessary to stress the social role of responsibility as a way to balance powers and not only to blame culprits. This line implies to stress not only on the retrospective notion of accountability and responsibility, but on the prospective one, as a path for creating suitable conditions for the development of new technologies and preventing undesirable future outputs.

References

An, S-K. and Gower, K.K. (2009). How do the news media frame crises? A content analysis of crisis news coverage, *Public Relations Review,* 35, 107-112. doi: doi:10.1016/j.pubrev.2009.01.010.

Asaro, P. M. (2012). A body to kick, but still no soul to damn: Legal perspectives on robotics. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics*. Cambridge: MIT Press.

Baldi, V. (2017). Beyond the algorithmic and automated society. Towards a critical reappropriation of digital culture. *Observatorio (OBS),* 11(3), 186-198. doi: 1646-5954/ERC123483/2017.

Baruh, L. and Popescu, M. (2015). Big data analytics and the limits of privacy self-management. *New Media & Society,* 19(4), 1-18. doi: 10.1177/1461444815614001.

Bostrom, N. (2015). *What happens when our computers get smarter than we are?* Conference at TED Talks. https://www.ted.com/talks/nick_bostrom_what_happens_when_our_computers_get_smarter_than_we_are?language=en.

Bovens, M. (2005). Public accountability. In E. Ferlie, L. E. Lynn, & C. Pollitt (Eds.), *Oxford handbook of public management*, pp. 182–208. New York: Oxford University Press.

Boyd, D. and Crawford, K. (2011). Six provocations for big data, *Paper presented at: A decade in internet time: Symposium on the dynamics of the internet and society*. doi: http://dx.doi.org/10.2139/ssrn.1926431.

Boyd, D., and Ellison, N. B. (2007). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer Mediated Communication*, 13(1), 210–230. doi: 10.1111/j.1083-6101.2007.00393.x

Boyd, D., Levy, K. and Marwick, A. (2014). The networked nature of algorithmic discrimination. In S. Gangadharan (Ed.) *Data and discrimination: Collected essays*, 53-57. Washington, DC: Open Technology Institute – New America Foundation.

Campbell, P. (2010). Boundaries and risk: Media framing of assisted reproductive technologies and older mothers. *Social Science & Medicine*, 72, 265-272. doi:10.1016/j.socscimed.2010.10.028.

Caplan, R. and Boyd, D. (2016). Who controls the public sphere in an era of algorithms? Mediation, automation, power, *Contemporary Issues and Concerns Primer, Data & Society*. Retrieved from: https://datasociety.net/events/who-controls-public-sphere.

Carbonell, J. G., Michalski, R. S. and Mitchell, T. M. (1984). An overview of Machine Learning. An Artificial Intelligence Approach. Berlin: Springer-Verlag.

Castells, M. (2009). The Rise of the Network Society: The Information Age: Economy, Society, and Culture. Cambridge: Blackwell Publishers.

Chakrabarti, C. and Luger, G. F. (2015). Artificial conversations for customer service chatter bots: Architecture, algorithms, and evaluation metrics. *Expert Systems with Applications,* 42, 6878-6897. doi: 10.1016/j.eswa.2015.04.067.

Coman, C.& Cmeciu, C. (2014). Framing Chevron Protests in National and International Press. *Procedia - Social and Behavioral Sciences,* 149, 228 – 232.

Diakopoulos, N., and Friedler, S. (2016). How to Hold Algorithms Accountable, *MIT Technology Review, November 2016*. Retrieved from: https://www.technologyreview.com/s/602933/how-to-hold-algorithms-accountable/?utm_content=buffer19bc5andutm_medium=socialandutm_source=twitter.c%E2%80%A6.

Druckman, J.N. & Bolsen, T. (2011). Framing, Motivated Reasoning, and Opinions about Emergent Technologies. *International Journal of Communication, 61*: 659-688. doi:10.1111/j.1460-2466.2011.01562.x.

Entman, R. M. (1993). Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication 43*(4): 58-67. doi: https://doi.org/10.1111/j.1460-2466.1993.tb01304.x.

Ferrara, E. Varol, O. Davis, C. Menczer, F. & Flammini, A. (2016). The Rise of Social Bots. *Communications of the ACM,* 59(7), pp. 96-104. doi: 10.1145/2818717.

Fisher, J.M. (1999). Recent work on moral responsibility. *Ethics,* 110(1), 93–139.

Floridi, L. (2014). Artificial Agents and Their Moral Nature. In Kroes, P. & Verbeek, P. (Eds.), *The Moral Status of Technical Artefacts, 17*: 185-212. Springer Dordrecht Heidelberg, New York, London.

Floridi, L. and Sanders, J. (2004). On the morality of artificial agents, *Minds and Machines, (14)*3, pp: 349. doi: https://doi-org.sare.upf.edu/10.1023/B:MIND.0000035461.63578.9d.

Ford, H.R, Dubois, E. & Puschmann, C. (2016). "Keeping Ottawa Honest—One Tweet at a Time? Politicians, Journalists, Wikipedians, and Their Twitter Bots". *International Journal of Communication, 10*: 4891-4914. doi: 1932–8036/20160005.

Gamson, W. A. (1992). *Talking politics*. New York: Cambridge University Press.

Gotterbarn D. (2001). "Informatics and professional responsibility,". *Science and Engineering Ethics, 7*(2): 221–230.

Griol, D., Sanchis de Miguel, A. and Molina, J. M. (2014). 'Giving Voice to the Internet by Means of Conversational Agents'. In Corchado E., Lozano J.A., Quintián H., Yin H. (Eds). *Intelligent Data Engineering and Automated Learning*. doi: 10.1007/978-3-319-10840-7_53.

Groom, V., Chen, J., Johnson, T., Kara, F. A. and Nass, C. (2010). Critic, Compatriot, or Chump?: Responses to Robot Blame Attribution, *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction , March 02-05, 2010, Osaka, Japan*. doi: 978-1-4244-4893-7.

Guilbeault, D. (2016). "Growing Bot Security: An Ecological View of Bot Agency". *International Journal of Communication, 10*: 5003-5021. doi: 1932–8036/20160005.

Hallinan, B. and Striphas, T. (2016). Recommended for you: The Netflix Prize and the production of algorithmic culture. *New Media and Society, 18*(1): 117-137. doi: 10.1177/1461444814538646.

Hellström, T. (2013). On the moral responsibility of military robots. *Ethics and Information Technology, 12*(2): 99-107.

Introna, L. D. (2014). 'Towards a Post-human Intra-actional Account of Sociomaterial Agency (and Morality)'. In Peter Kroes & Peter-Paul Verbeek (Eds.), *The Moral Status of Technical Artefacts:* 31-53. New York, London: Springer.

Johnson, D. G. (2006). Computer Systems: Moral Entities but not Moral Agents. *Ethics and Information Technology, 8*, pp: 195–204.

Karppi, T., and Crawford, K. (2016). Social Media, Financial Algorithms and the Hack Crash. *Theory, Culture and Society, 33*(1), 73-92. doi: 10.1177/0263276415583139.

Kemper, J. and Kolkman, D. (2018). Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society,* doi: 10.1080/1369118X.2018.1477967.

Kollani, B., Howard, P. and Wooley, S. C. (2016). Bots and Automation over Twitter during the Third U.S. Presidential Debate, *Data Memo 2016.3. Oxford, UK: Project on Computational Propaganda*. https://www.oii.ox.ac.uk/blog/bots-and-automation-over-twitter-during-the-third-u-s-presidential-debate/.

Kroes, P., and Verbeek, P-P. (2014). *The Moral Status of Technical Artefacts*. New York, London: Springer.

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., and Yu, H. (2017). Accountable algorithms, *University of Pennsylvania Law Review*, 165: 633. Retrieved from: https://heinonline.org/HOL/LandingPage?handle=hein.journals/pnlr165&div=20&id=&page=&t=1558886997.

Laney, D. (2001). 3D data management: Controlling data, volume, velocity and variety, *Application delivery strategies*, *File 949.* Meta Group Research Note. Retrieved from: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Lawrence, R. G. (2000). Game-Framing the Issues: Tracking the Strategy Frame in Public Policy News, *Political Communication*. doi: 10.1080/105846000198422.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Review. Deep Learning. *Nature, 521*: 436-444. doi: 10.1038/nature14539.

Lee, P. (March, 25th 2016). *Learning from Tay's introduction*, Statement published in The Official Microsoft Blog. Retrieved from: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/.

Lesk, M. (1997). *How Much Information Is There In the World?* Retrieved from: http://www.lesk.com/mlesk/ksg97/ksg.html.

Marechal, N. (2016). "When Bots Tweet: Toward a Normative Framework for Bots on Social Networking Sites". *International Journal of Communication, 10*: 5022-5031. doi: 1932–8036/2016FEA0002.

Marsland, S. (2015*). Machine Learning: An Algorithmic Perspective.* Boca Raton: CRC Press.

Marwick, A. and Lewis, R. (2017). *Media Manipulation and Disinformation Online*, New York: Data & Society Research Institute. Retrieved from: https://datasociety.net/output/media-manipulation-and-disinfo-online/.

Matamoros-Fernández, A. (2017). Platformed Racism: The Mediation and Circulation of an Australian Race-Based Controversy on Twitter, Facebook and YouTube. *Information Communication and Society*, 20 (6), 930–46. doi:10.1080/1369118X.2017.1293130.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3): 175-183.

Microsoft Corporation. (2016). *Tay.ai Official Webpage* [now disabled]. Retrieved from: www.tay.ai.

Mitcham, C. (2014). 'Agency in humans and in artifacts: A contested discourse'. In Kroes, P. and Verbeek, P_P. (Eds.), *The moral status of technical artifacts*: 11-29. Dordrecht, The Netherlands: Springer Science and Business Media. doi: 10.1007/978-94-007-7914-3_2.

Murthy, D., Powell, A. B., Tinati, R., Anstead, N., Carr, L., Halford, S.J., and Weal, Mark (2016). Bots and Political Influence: A Sociotechnical Investigation of Social Network Capital*. International Journal of Communication*, 10: 4952-4971. doi: 1932–8036/20160005.

Nath, V. and Levinson, S.E. (2014). *Autonomous robotics and Deep Learning.* New York, London: Springer.

Neff, G. and Nagy, P. (2016). Talking to Bots: Symbiotic Agency and the Case of Tay. *International Journal of Communication, 10*, 4915-4931. doi: 1932–8036/20160005.

Nichols, N. (2010). Machine-Generated Content: Creating Compelling New Content from Existing Online Sources. Ph.D. Dissertation. Northwestern University, Evanston, IL, USA.

Olsher, D. (2014). Semantically-based priors and nuanced knowledge core for Big Data, Social AI, and language understanding. *Neural Networks, 58:* 131-147. doi: 10.1016/j.neunet.2014.05.022.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge: Harvard University Press.

Puschmann, C. and Burgess, J. (2014). Metaphors of Big Data. *International Journal of Communication*, 8: 1690–1709. http://ijoc.org/index.php/ijoc/article/view/2169.

Rieder, B. (2018). Beyond Surveillance: How Do Markets and Algorithms 'Think'? *Le foucaldien*, 3(1) 8, pp. 1–20, DOI: https://doi.org/10.16995/lefou.30.

Rieder, B., Matamoros-Fernández, A., and Coromina, Ò. (2018). From ranking algorithms to 'ranking cultures': Investigating the modulation of visibility in YouTube search results. *Convergence*, 24(1), 50–68. http://doi.org/10.1177/1354856517736982.

Rosenblat, A., Kneese, T. and boyd, d. (2014). Algorithmic Accountability, *The Social, Cultural & Ethical Dimensions of "Big Data"*. doi: http://dx.doi.org/10.2139/ssrn.2535540.

Sandvig, Ch., Hamilton, K., Karahalios, K., and Langbort, C. (2016). When the Algorithm Itself Is a Racist: Diagnosing Ethical Harm in the Basic Components of Software. *International Journal of Communication, 10*: 4972-4990. doi: 1932–8036/20160005.

Semetko, H. a., and Valkenburg, P. M. (2000). Framing European Politics. A Content Analysis of Press and Television News. *Journal of Communication*, *50:* 93-1009. doi: 10.1111/j.1460-2466.2000.tb02843.x.

Shalev-Shwartz, S. and Ben-David, S. (2014*). Understanding Machine Learning: From theory to Algorithms*. New York: Cambridge University Press.

Spicer, A. (2005). The political process of inscribing a new technology. *Human Relations, 58*(7): 867-890. doi: 10.1177/0018726705057809

Stahl, W.A. (1995). Venerating the Black Box: Magic in Media Discourse on Technology. *Science, Technology and Human Values, 20*(2): 234-258. http://www.jstor.org/stable/689992*.*

Suárez-Gonzalo, Sara (2017). Big Social Data: some limitations of Notice and Choice for privacy protection. *El profesional de la información, 26*(2), pp. 283-292. doi: 10.3145/epi.2017.mar.15

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *Procedings of the 8th Intl AAAI Conferece on weblogs and social media.* https://arxiv.org/abs/1403.7400.

van Dijk, J. (2013). *The cutlure of connectivity. A critical history of social media*. New York: Oxford University Press.

Waelbers, K. (2009). Technological Delegation: Responsibility for the Unintended. *Science & Engineering Ethics, 15*(1): 51–68.

Zollo, S. A. and Loos, E. (2017). No Hate Speech Movement: evolving genres and discourses in the European online campaign to fight discrimination and racism. *Observatorio, 11*(2): 91-107. doi: 1646-5954/ERC123483/2017.