Digital Zombies - the Reanimation of our Digital Selves

Tabea Tietz
FIZ Karlsruhe - Leibniz Institute for Information Infrastructure
Karlsruhe Institute of Technology
tabea.tietz@fiz-karlsruhe.de

Francesca Pichierri
FIZ Karlsruhe - Leibniz Institute for Information Infrastructure
francesca.pichierri@fiz-karlsruhe.de

Maria Koutraki
FIZ Karlsruhe - Leibniz Institute for Information Infrastructure
Karlsruhe Institute of Technology

Dara Hallinan
FIZ Karlsruhe - Leibniz Institute for Information Infrastructure
dara.hallinan@fiz-karlsruhe.de

Franziska Boehm
FIZ Karlsruhe - Leibniz Institute for Information Infrastructure
franziska.boehm@fiz-karlsruhe.de

Harald Sack
FIZ Karlsruhe - Leibniz Institute for Information Infrastructure
Karlsruhe Institute of Technology
harald.sack@fiz-karlsruhe.de

ABSTRACT

What happens to our social media profiles when we die? The episode "Be Right Back" as part of Netflix's series "Black Mirror"[1] provides a possible scenario. A digital avatar is created to communicate with close relatives which learns from past social media activities of the deceased user. While the users entrust their social media content to one or more companies, even after their death, it may be reasonable to ask: What will the company really do with a deceased user's data: sell it to manipulate users or create advertisements? In this paper we tackle the issues of ownership, ethics, and transparency of post mortem user data.

CCS CONCEPTS

• **Security and privacy → Social aspects of security and privacy**; • **Information systems →** *World Wide Web*; *Document topic models*; *Information extraction*; *Sentiment analysis*; • **Human-centered computing →** *Social media*; • **Social and professional topics →** *Codes of ethics*; *Intellectual property*;

KEYWORDS

Social Media; Artificial Intelligence; Privacy; Black Mirror; Law; Ethics; Transparency

ACM Reference Format:
Tabea Tietz, Francesca Pichierri, Maria Koutraki, Dara Hallinan, Franziska Boehm, and Harald Sack. 2018. Digital Zombies - the Reanimation of our Digital Selves. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France,* Pinelopi Troullinou, Mathieu d'Aquin, and Ilaria Tiddi (Eds.). ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3184558.3191606

1 INTRODUCTION

Social media is omnipresent, it affects our daily lives in every aspect, and the next story, video, image, or tweet is just one tap away. We post what we eat, how we exercise, what music we listen to, which politicians we support and which career we pursue. Around 98% of digital users between the age of 16 and 64 are social media users and on average, each user has eight accounts on various social platforms [1]. However, questions which have been raised more often recently are: What will happen to my data when I die? Who will be able to access, share and alter my data? How can my data be protected in the future and who will own these data? Of course, the social media profiles of deceased users could be simply frozen or deleted – a way how Facebook is dealing with it at the moment[2]. But, let's go into a different direction and think about the possibilities that come with all the generated original content. What if my close relatives would be able to view and share our shared (online) memories or interact with my digital avatar, created on the foundation of my previous social media content?

It's not hard to imagine that these *what if's* could be desired by users to cope with grief and void after the loss of loved ones. Taking into account the rapid development of Web technologies in the past decade, it's further not hard to imagine that services which provide these features will be available soon. A few are already on the way. One approach by Eternime[3] "preserves your most important thoughts, stories and memories for eternity". You are able to create a digital avatar which close relatives can interact with post mortem. While Eternime makes use of the content created by the user and plays it back to close relatives, we attempt to take the idea one step further to actually interacting with a deceased user. In this paper, we closely follow the scenario raised in the episode *Be Right Back* of Netflix's series Black Mirror. The episode deals with the issues of grief, memory, data, and ethics. In it, the protagonist and active social media user Ash is suddenly killed. His girlfriend Martha starts using a service which allows relatives to stay in touch with their deceased loved ones. Through learning from past social media posts by Ash, the service is able to generate new content impersonating him.

[1] Be Right Back: Season 2 Episode 1 overview on Wikipedia https://en.wikipedia.org/wiki/Be_Right_Back

[2] Facebook overview of reporting a deceased person: https://www.facebook.com/help/408583372511972/

[3] http://eterni.me/

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191606

While the described Black Mirror episode provides (apart from its ethical problems) a working technological example of the scenario, it is also imaginable that the company that is entrusted with the data of the deceased may exploit it for commercial use or for influencing opinions of the living. In this paper, a scenario describing a possible use of post mortem social media profiles will be given, building on the mentioned Black Mirror episode. In relation to the scenario, ethical and legal considerations will be raised. It will be shown that beyond simply developing more powerful and intelligent (Web) technologies, there are still numerous unanswered questions regarding the rights and wrongs of new technologies, such as the development of artificial intelligence (AI). In this regard, we propose that there is a need for making the use of new technologies on personal data more transparent and it will further be discussed, how technology could play a role in an attempt to detect unexpected use of such deceased digital profiles.

2 SCENARIO

Laura is an active social media user and has accounts on multiple platforms, which she uses daily. She is made aware that there is a company that helps her to make use of all of the content she has created after her death – to communicate with her loved ones with the help of a digital avatar. Laura loves the idea that her friends and family will be able to stay "in touch" as if she was still there and proceeds to tick a box in her social platforms' security and privacy settings, to have her profile's content analysed by the third party company after her death. The company provides a service using artificial intelligence (AI) to learn from Laura's original content and to create a digital avatar. The avatar is smart, it is able to interact with close relatives and friends through a chat bot sharing thoughts and feelings as if she had never been away. It shares past memories and adventures and it publishes completely new content generated by an AI on the bases of the original postings by Laura, following her footsteps closely.

Under Laura's consent to all of these features, however, the company has also assumed to have the power to manipulate and use Laura's profile for their own ends. For example, the company begins to:

1. Exploit Laura's profile and personal data in order to manipulate the conversations between the avatar Laura and her family, playing with their emotional content, using sorrow and grief as leverage for its own economic advantage (e.g. pushing people to use the service more frequently, creating a sort of addiction that holds people back from moving on).
2. Disseminate specific messages (e.g. on political topics) which Laura never would have agreed to.
3. Share private information (e.g. from chats) which were not meant to be shared.
4. Publicly share Laura's data.
5. Sell Laura's profile to advertising companies to create advertisements based on her and her friend's profiles.
6. Use Laura's data for analytics and marketing purposes.

While these cases of further use are solely examples, there are many more to think of. The general questions that can be raised now are: How can Laura's data be protected after she deceased? How can Laura be confident that the company she entrusted her data with after her death will not exploit the simple fact that she cannot react to content newly created by her avatar? The following section 3 attempts to discuss legal and ethical considerations on how to deal with the content of deceased users.

3 ETHICAL AND LEGAL PERSPECTIVES

Ethical considerations. As Laura gave her consent to have her profile's content analysed in order to create a digital version of herself after her death, the company is, in fact, able to manage Laura's profile, potentially engaging with, and effecting, the emotions of the living – for example, Laura's relatives and friends. In turn, it also takes over as a manager and manufacturer of Laura's memory. On an individual level, considering the emotional connotations of the deceased, this is power indeed. On a social level, considering the meanings attached to death, and the deceased – for example the finality associated with the concept – the company exercises an ability to alter a construct which has been central to human society for millennia. Unsurprisingly then, a number of fascinating ethical questions emerge. For example:

- Are there any ethical templates which could be extrapolated to provide guidance in relation to deceased's digital profiles?
- Are there any clear wrongs which might be associated with the use of deceased digital profiles?

Legal considerations. Ordinarily, one might look to law to provide a framework – or at least general rules – outlining the boundaries of legitimate conduct in any given scenario. In this case, however, areas of law which one might expect to regulate such an issue appear not to be relevant. A prime example is data protection law. Ostensibly, this looks highly relevant – it is, after all, the area of law designed to protect rights in relation to the processing of individual's data. Yet, the scope of data protection law in the EU tends to exclude the deceased. Accordingly, the whole area of law cannot apply to deceased digital profiles. The new General Data Protection Regulation, GDPR, specifically states in Recital 27 that it does not apply to the personal data of deceased persons, following the path of the Directive 95/46/EC (cf. [5, 8, 9]). As with many technological phenomena, perhaps the questions posed simply have not needed to be subject to the cold steel of the lawmaker's pen. This raises a number of legal questions. For example:

- How far can consent go? Can consent be given "forever"?
- Which legitimate rights and interests might be recognized in relation to deceased digital profiles – the deceased person's personality rights, relatives' rights to a memory?
- Are there any areas of law already applicable and if so, what principles do they outline – property law?
- Given the novelty of the scenario, what kind of regulation – if any at all – would be beneficial?

The above ethical and legal considerations show the great uncertainty associated with the company's use of the deceased digital profile. Nevertheless, the obvious novel power the company has obtained, the emotions involved on the part of living individuals and the centrality of the concept of death to the social fabric mean that certain assertions can be un-problematically made: 1) this is an issue which would be of considerable public interest; 2) there is the potential for great individual and social harm. In the light of these two assertions, whilst it may be premature to declare specific normative positions in relation to the use of deceased digital profiles, it is not premature to consider means by which such positions may be arrived at – both individually and socially. In this regard, a first step may be to consider how to make the phenomenon as transparent as possible. For example, by devising means whereby those engaging with such profiles might know they relate to the deceased.

4 WEB TECHNOLOGIES AND DATA MISUSE

While the scenario developed in the Black Mirror episode describes a future perspective, the technologies needed to create a digital avatar which automatically generates new content from learned original content are already there. Machine learning technologies enable text prediction, even to the scope of (semi-automatically) generating completely new chapters from books like Harry Potter[4]. Natural Language Generation (or text generation) is a task in NLP in which natural language text is automatically generated. There is a large corpus of research works on these topics with some being [14, 18]. Thus, for the purpose of this work, it is assumed that the company controlling the profile of the user is able to automatically generate new content to be published in the profile. However, the goal of this section is not to analyse how the Black Mirror scenario can be enabled, but how to tackle the issue of data misuse. In the context of the proposed scenario, data 'misuse' can not be clearly defined (as the previous chapter discussed) from the legal and ethical perspective. Also on the technological level it is by far not trivial to clearly define where a useful feature ends and where data exploitation starts. It can further be asked who decides what a violation of the original poster's data actually is, since the poster herself cannot take action on future generated chats and posts by their avatar. One way to tackle this problem might be to compare the original postings to the new AI generated content. It could be asked: Does the bot post in the same way the original poster would? Are there changes on the emotional level? Are there changes in the frequency, structure or content of postings? Of course, a clear line cannot be drawn here either. However, answering these questions may provide indicators which help to increase transparency in the new content to protect the original poster in retrospect. Following, a selection of current technologies and works will be discussed which may enable to find indicators of data misuse in the described context.

Topic Detection. A technology which can be proven rather helpful in detecting misuses in a social media profile, as was described in the scenario of Sec 2, is *Topic Detection* or *Topic Modeling*. A topic model is a type of statistical model for discovering topics occurring in a text. One of the most famous topic modeling algorithms is Latent Dirichlet allocation (LDA) [2]. LDA works as following: it represents a collection of documents as a mixture of topics and a word in a document being attributed to a topic. Many works have already proposed for topic modeling (detection) in social media [10, 21, 23], the one focusing on Twitter[5] are surveyed in [11]. The intuition here is that by applying topic detection first on content that was originally provided by the user herself, before her death, one can extract the topics that she was interested in. For example, if she used to have several posts similar to *"I wish that you all enjoy the sunshine today!"* or to *"The last Star Wars movie was my favorite."* Hence it is being detected that she was mostly posting about *Weather* and *Movies*. Later on, when the company manages the profile, topic detection can be applied again. In that case if the new detected topics are completely different to the previous ones, e.g. there are many posts about *Politics*, that can be a hint of misuse.

Emotion/Sentiment Detection. Emotions and sentiments are key in human to human communication, and therefore emotion and sentiment detection play a major role in the advancement of AI technologies [4][6]. In current research the application fields of emotion detection and sentiment analysis is incredibly large and includes numerous fields including market analysis (e.g. product reviews) or psychology (e.g. suicide prevention), the prediction of political events (e.g. elections) and many more. Especially the social media platform Twitter proves to be a fruitful source of emotion detection and sentiment analysis with around 300 million users and approximately 500 million tweets per day. Giahanou and Crestani [6] discuss a number of current approaches of sentiment analysis and emotion detection including machine learning techniques, lexicon-based methods, hybrid methods and graph-based techniques. This ongoing field of research and especially the efforts incorporating microblogging platforms such as Twitter might also be beneficial in the detection of the misuse of digital avatars. With the analysis of emotion and sentiment, it could be systematically investigated whether the digital avatar is suddenly beginning to use sarcasm extensively (e.g. [7, 13, 17]), react to political events (e.g. terror attacks) with different emotions than the original poster (e.g. [3]), or the digital avatar suddenly posts hate speech (e.g. [19]) or suddenly shows different emotions towards certain celebrities or products (e.g. [16]). When comparing both, the original postings and the postings by the digital avatar, and sudden emotional changes take place regarding specific topics, it may be an indicator that the AI is programmed to manipulate the people it interacts with, deliver advertisement, or create political opinions the original poster would not agree to.

Interaction Analysis of Content Consumers. Apart from the analysis of changes in the postings and chats of the digital avatar itself, the perspective of the consumers of this content could also be taken. It could be analysed, how the living who have access to the digital avatar change their reactions to the posts [22]. Are these consumers sharing or 're-tweeting' the content frequently? Are they 'liking' the generated content or do they ignore it after all? Of course, one could argue that the reactions friends and acquaintances show towards the content of the living might be quite different to the reactions they show regarding the generated content of a digital avatar. However, it could still be an indicator of how well the digital

[4] The Botnik project provides a semi-automated method for text prediction to be used by lay users http://botnik.org/

[5] https://twitter.com/

[6] Both terms – emotion and sentiment – are often used interchangeably in computer science. According to Munezero et al. however, emotions are referred to as "brief episodes of brain, autonomic, and behavioral changes", while sentiments are emotional dispositions developed over a longer time period [15]. Both concepts play a major role in analysing social media profiles and can be applied to textual content as well as multimedia analysis.

avatar's content is received. For instance, a frequent sharing and liking of products or political messages may be an indicator for user manipulation while reporting or even blocking the digital avatar's content may be an indicator for inappropriate or controversial content. Out of people's reactions to posts and chats, a sentiment analysis could be performed. This may include an analysis of the used Facebooks reaction emoji's [20] or an in-depth analysis of re-tweets on twitter [12].

It is clear that the provided examples only partially cover the possibilities of making changes in social media posting behavior transparent. For instance, the field of linguistic text analysis provides a broad tool set with a variety of approaches to examine a text's grammatical and morphological characteristics. Further analysis could also revolve around an investigation of the frequency of postings or the media formats used (video, photo, text). However, it was shown that it is possible to monitor the interaction of digital avatars with the world of the living and provide at least transparency towards changes in their posting behavior to increase the protection of a social media user.

5 DISCUSSION AND CONCLUSION

In this paper a future scenario on the use of social media profiles of deceased users is presented, following Netflix's Black Mirror episode "Be Right Back". The scenario describes that a digital avatar of a user's original posts is created which is programmed to generate new content in the form of posts and chats.

In relation to it, we then made some ethical and legal considerations. Superficially, the scenario seems to depict a dark plot indeed. As we think more carefully about the legal and ethical issues at stake, however, uncertainty creeps in. Do we know that the company's use of deceased digital profiles will be negative? As one begins to think about the ethical issues the scenario raises, we are immediately faced by the novelty of the situation and the difficulty it brings in asserting any specific ethical framework – with its own sets of normative certainties. Ordinarily, we might look to law to provide – at least procedural – certainty. Yet here we meet the same lack of clarity. The areas of law one might think most relevant – such as data protection - seem not to apply. Despite the current lack of ethical and legal guidance, we are not without a way forward.

The idea that people may have the chance to interact with their deceased relatives is likely to have considerable impact. It has the potential to evoke considerable emotional response, even harm, on the part of the individuals concerned. It seems highly likely to provoke clear responses from groups with certain moral positions relating to the dead – religious groups, for example. What does one do when faced by a phenomenon in relation to which there is moral and legal uncertainty, yet which seems likely to have a strong impact and be potentially problematic? The answer is to make it, wherever possible, transparent. To put in place the mechanisms which allow the gathering, on an individual and social level, of the empirical data which allow clearer moral – and perhaps eventually legal – judgments to be made. Here, it seems likely that technology can play an important role. Current technologies enable to make changes in social media postings transparent. Changes may occur on the emotional level, on the topics used or the interaction with followers, friends, and relatives. While these technologies cannot clearly state when a user's data has been misused, they help to find indicators to be brought into context.

The scenario itself raises interesting questions about the way in which technology can force questions about the physical world in which we live. At its essence, the technology permits a form of reanimation of the dead. This poses fundamental questions as to how final death really is, and as to how close digital versions of individuals can ever approximate the real thing. In turn, the scenario forces us to reconsider the social overlays of morality and law we currently apply to the living and the dead. What a convenient and clear distinction this has proven to be, and what a difficult situation we find ourselves in when the distinction is no longer clear. It raises questions such as whether the deceased can have rights, such as privacy rights, and how - if at all - these might be effectively protected in law given they are no longer there to protect themselves.

More abstractly, working on the scenario in the framework of the workshop raises questions as to the sense in using 'Black Mirror' scenarios as templates for consideration of ways to deal with the problems of technology at all. Identifying Black Mirror scenarios sets a tone in which certain technologies are, from the outset, painted as bad. Naturally, the response is then to consider means for prevention. Yet, the reality of technology is seldom straightforward, it is seldom possible to place it in one clear box marked 'bad, do not open'. In our case, for example, perhaps the use of deceased digital profiles could in fact be of great assistance in dealing with grief and loss? In turn, perhaps such scenarios are in fact necessary in themselves? It is through negative events – often visceral negative events – that social conscience is awoken and social debate and decision making mechanisms are triggered. Take these away through advance prevention, and this may opens the way for much more insidious harms. Who is to say that these will not be considerably worse?

REFERENCES

[1] 2017. GlobalWebIndex's flagship report on the latest trends in social media. (2017).

[2] David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3 (2003), 2003.

[3] Pete Burnap, Matthew L Williams, Luke Sloan, Omer Rana, William Housley, Adam Edwards, Vincent Knight, Rob Procter, and Alex Voss. 2014. Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining* 4, 1 (2014), 206.

[4] Erik Cambria. 2016. Affective Computing and Sentiment Analysis. *IEEE Intelligent Systems* 31, 2 (2016), 102–107.

[5] Lilian Edwards and Edina Harbina. 2013. Protecting post-mortem privacy: Reconsidering the privacy interests of the deceased in a digital world. *Cardozo Arts & Ent. LJ* 32 (2013), 83.

[6] Anastasia Giahanou and Fabio Crestani. 2016. Like It or Not: A Survey of Twitter Sentiment Analysis Methods. *ACM Comput. Surv.* 49 (2016), 28:1–28:41.

[7] Roberto González-Ibánez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in Twitter: a closer look. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 581–586.

[8] Edina Harbinja. 2013. Does the EU data protection regime protect post-mortem privacy and what could be the potential alternatives. *SCRIPTed* 10 (2013), 19.

[9] Edina Harbinja. 2017. Post-mortem privacy 2.0: theory, law, and technology. *Int. Review of Law, Computers & Technology* 31, 1 (2017), 26–42.

[10] Liangjie Hong and Brian D. Davison. 2010. Empirical study of topic modeling in Twitter. In *Proc. of the 3rd Workshop on Social Network Mining and Analysis, SNAKDD 2009*. 80–88. https://doi.org/10.1145/1964858.1964870

[11] Rania Ibrahim, Ahmed Elbagoury, Mohamed S Kamel, and Fakhri Karray. 2017. Tools and approaches for topic detection from Twitter streams: survey. *Knowledge and Information Systems* (2017), 1–29.

[12] Intzar Ali Lashari and Uffe Kock Wiil. 2016. Monitoring Public Opinion by Measuring the Sentiment of Retweets on Twitter. In *3rd European Conf. on Social Media Research*. 153.

[13] Diana Maynard and Mark A Greenwood. 2014. Who cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis.. In *Lrec*. 4238–4243.

[14] Kathleen McKeown. 1992. *Text generation*. Cambridge University Press.

[15] Myriam D Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. 2014. Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE trans. on affective computing* 5, 2 (2014), 101–111.

[16] Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proc. of the 10th Int. Workshop on Semantic Evaluation (SemEval-2016)*. 1–18.

[17] Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A Behavioral Modeling Approach. In *Proc. of the 8th ACM Int. Conf. on Web Search and Data Mining*. ACM, 97–106.

[18] Ehud Reiter and Robert Dale. 2000. *Building natural language generation systems*. Cambridge university press.

[19] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proc. of the 5th Int. Workshop on Natural Language Processing for Social Media*. 1–10.

[20] Sarah Turnbull and Simon Jenkins. 2016. Why Facebook Reactions are good news for evaluating social media campaigns. *Journal of Direct, Data and Digital Marketing Practice* 17, 3 (2016), 156–158.

[21] Yu Wang, Eugene Agichtein, and Michele Benzi. 2012. TM-LDA: efficient online modeling of latent topic transitions in social media. In *The 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, KDD '12*. 123–131. https://doi.org/10.1145/2339530.2339552

[22] Velissarios Zamparas, Andreas Kanavos, and Christos Makris. 2015. Real time analytics for measuring user influence on twitter. In *Tools with Artificial Intelligence (ICTAI), 2015 IEEE 27th International Conference on*. IEEE, 591–597.

[23] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing Twitter and Traditional Media Using Topic Models. In *Advances in Information Retrieval - 33rd European Conf. on IR Research, ECIR 2011*. 338–349. https://doi.org/10.1007/978-3-642-20161-5_34