Bots Acting Like Humans: Understanding and Preventing Harm

Florian Daniel
Politecnico di Milano, Milan, Italy

Cinzia Cappiello
Politecnico di Milano, Milan, Italy

Boualem Benatallah
University of New South Wales, Sydney, Australia

Bots are algorithmically driven entities that act like humans in conversations via Twitter, on Facebook, in chats or Q&A sites. This article studies how they may affect on-line conversations, provides a taxonomy of harms that may be caused, and discusses how to prevent harm by studying when abuses occur.

Many technologists consider chatbots one of the hottest technologies in recent times (https://bit.ly/2od0Tdw), an opinion fueled for example by Facebook's release of its Messenger API in 2016. In April 2017, Facebook reported 100,000 monthly active bots on the Messenger platform. In March 2017 Varol et al.[1] estimated that between 9% and 15% of active Twitter accounts are bots (29–49 million accounts out of 328 millions, https://bit.ly/2v3AT6O). Gartner estimates that by 2020 85% of customer requests will be handled by bots, while Inbenta estimates 1.8 billion unique customer chatbot users by 2021.[2]

The technological advancements of chatbots undoubtedly produced a hype on its own, yet bots today are by far not limited to instant messaging only. Bots permeate all kinds of on-line conversations in Twitter, Facebook, Instagram, Q&A sites, on-line newspapers, emails, and the like. They are everywhere where there are humans conversing with each other via the Internet, legitimately or illegitimately. For example, Messenger explicitly allows bots in its chats, while WhatsApp states that it blocks phone numbers generating bot traffic (https://bit.ly/2HhW9wG). Inspired by Bessi and Ferrara,[3] we thus understand bots (sometimes also called sybils[4]) generically as *algorithmically driven entities that on the surface act like legitimate human users in on-line conversations*.

Even though on-line bots are multiplying their presence in public and private communications, most organizations and users still do not have the knowledge, skills, or understanding to craft a successful strategy to keep up with the possible unintended consequences of this presence. If we

think of *physical* robots, significant research and development efforts are being invested into *robot ethics*.[5] The IEEE has a Technical Committee on Robot Ethics (https://bit.ly/1l0QMzh) to address "urgent ethical questions prompted by and associated with robotics research and technology." Yet, when it comes to *virtual* robots, nothing alike has been proposed so far, and we still lack a proper discussion, let alone guidelines, on what could be called *bot ethics* – a lack that has, for instance, become manifest recently with Google's Duplex feature of its Assistant, enabling it to make voice phone calls that are so realistic that the human counterpart is not able to tell it is talking to a bot.[6]

In this article, we do not attempt to define ethical guidelines for bots. We rather explore a specific aspect of ethics, i.e., harm, and aim to raise the awareness of the damages that may be caused by bots acting like humans in on-line conversations. We envision creating a common understanding of harm and its sources as the first step toward agreeing on ethical rules for bots. The contributions of this article are thus threefold: First, it *streamlines the types of harm* that bots may cause in social networks, chat rooms, Q&A sites, forums, and the like. Second, it analyzes and *streamlines the types of abuse* that are the sources of harm. Third, it *discusses preliminary detection techniques and respective challenges* that need to be overcome to help prevent harm.

A TAXONOMY OF BOT HARMS

A *harm* occurs when someone suffers an injury or a damage, but also when someone gets exposed to a potential adverse effect or danger.

A starting point to understand what kinds of harm may occur in practice are concrete examples of what we can call bot failures, that is, incidents where a bot reportedly caused damage to someone. The *methodology* we follow to derive a respective taxonomy is thus example-driven analysis: We used generic Google Search, Google Scholar, as well as the ACM/IEEE/Springer online libraries to search for anecdotal evidence of bot failures. The result of the process comprised news articles, blog posts, scientific articles and online discussion threads, which we filtered manually to sort out examples of incidents that are real and as diverse as possible. The examples that passed our check are described in Table 1. Next, we used the taxonomy of generic harms proposed by the Institutional Review Board for Social and Behavioral Science of the University of Virginia (https://at.virginia.edu/2HPBUmm) that lists five types of harm that *individuals* may suffer from (psychological, physical, legal, economic and social harm) and performed a systematic mapping of the examples to the proposed taxonomy. We did not find any case of physical harm caused by a bot; we thus omit this category in the following. Instead, we found examples that could not be mapped to any of the five proposed types of harm. An analysis of these examples directed us toward a type of harm defined by Priest[7] as "democratic harm," a type of harm *society* as a whole may suffer from that seems particularly relevant today, e.g., if we consider the amount of fake news flooding social media every day.

In the following, we describe the identified types of harm and the selected examples. For each example, Table 1 indicates the harmed party (individual, group, society, bot owner) and provides an identifying code name:

- *Psychological harm* occurs when someone's psychological health or wellbeing gets endangered or injured; it includes feelings like worry, depression, embarrassment, shame, guilt, anger, loss of self-confidence, or inadequacy. An example of a bot causing psychological harm is Boost Juice's Messenger bot that was meant as a funny channel to obtain discounts by mimicking a dating game with fruits (BoostJuice). Unfortunately, the bot was caught using inappropriate language, that is, dis-educating children or teenagers. Self-ironically, Robert Epstein tells the story of him dating the Russian girl Ivana via email, who in the end – after being fooled for two months – turned out to be a bot (DatingIvana). Less intentionally, the increasing use of AI technology in bots may cause harm, if not properly controlled: for instance, a machine learning trained model has been demonstrated to discriminate against African-American voices (AASlang), or Microsoft's Twitter bot Tay had to be shut down within few hours because it started tweeting racist and threatening statements (MSTay).

- *Legal harm* occurs when someone becomes subject to law enforcement or prosecution; it includes for example the breach of a confidentiality agreement or contract, the release of protected information, or threatening. A good example is the case of Jeffry van der Goot, a Dutch developer, who had to shut down his Twitter bot, which generated random posts, after the bot sent out death threats to other users (DeathThreat). Police held him responsible for what the bot published. Microsoft's Tay had to be shut down because of hate crimes and for denying Holocaust, a crime in 14 European countries[8] (MSTay).

Table 1. Examples of bots causing harm (in alphabetical order).

Example			Harm		
Code name	Reference	Short description	Effect	Harmed party	Type of harm
AASlang	http://bit.ly/2wS2W7n	NLP software not able to understand African-American slang	discrimination	group (race)	psychological
AshleyMadison	http://bit.ly/2yoKBPF	Fembots on Ashley Madison dating site fooling users	deception, fraud	individual	psychological, economic
BoostJuice	http://bit.ly/2zvNt0E	Boost Juice's bot using inappropriate language with children	diseducation	group (children)	psychological
ColludingBots	http://tcrn.ch/2xDqC6J	User banned from Twitter after being followed by bots	banned from Twitter	individual	social
CustomerSvc	http://bit.ly/2DeJMvx	Bots pretending to be banking bot	loss of money	individual	economic
DatingIvana	http://bit.ly/2ApAf2z	Months of conversations with bot Ivana on online dating site	deception	individual	psychological
DeathThreat	http://bit.ly/2Dfm71P	Bot making death threats	fear	individual	psychological, legal
eCommerce	http://on.mash.to/2mOYcMp	Bot faking e-commerce service	loss of money	individual	economic
Geico	http://bit.ly/2AyfsLk	Geico bot accidentally courted racist Twitter trolls to sell insurance	offense, loss of reputation	group (customers), bot owner	psychological, social
InstaClone	http://bit.ly/2DIIuKj	Bot cloning Instagram accounts	loss of reputation	individual	social
Instagress	http://bit.ly/2ofkDMB	Instagress bot automatically liking posts on behalf of users	manipulation, loss of reputation	individual, bot owner	psychological, social
JasonSlotkin6	http://bit.ly/2Dfq4DH	Bots cloning real user profies	loss of reputation	individual	social
MSTay	http://bit.ly/2DCdqM4	AI bot manipulated to post racist and offening content	offense, discrimination	individual, group (race)	psychological
Oreo	http://bit.ly/2hW9xrE	Oreo bot answering a tweet with offensive account name	offense, loss of reputation	group (race), bot owner	psychological, social
PolarBot	http://bit.ly/2zuxm3n	Social media bots negatively affecting democratic political discussion	manipulation	society	democratic
Puma	http://bit.ly/2hW9xrE	Puma's "Forever Faster" campaign retweeting offenses	offense, loss of reputation	individual, bot owner	psychological, social
SethRich	http://nyti.ms/2jdvaas	Conspiracy about Seth Rich murder connected to Clinton email leaks	manipulation, loss of reputation	society, individual	democratic
SMSsex	http://bit.ly/2yLviF8	SMS sex spammer failing Turing test	embarrassment, spam	individual	psychological
SpamBot	https://bit.ly/2ITxI67	Twitter spam bots for politics, hastags and products	loss of time	individual, society	economic, democratic
Trump	http://bit.ly/2qxlvNM	Nearly half of Trump's Twitter followers are bots or fake accounts	manipulation	society	democratic
WiseShibe	http://bit.ly/2zu2b6r	Bot posting automated messages on Dodgecoin to obtain tips	fraud	individual	economic

- *Economic harm* occurs when someone incurs in monetary cost or looses time that could have been spent differently, e.g., due to the need to pay a lawyer or to clean one's own social profile. Bots are also new threats to security that eventually may lead to economic harm. For instance, Karissa Bell envisions that bots may provide fake e-commerce services, so as to steal credit card information (eCommerce), while Paul Schaus envisions bots pursuing a similar goal by infiltrating banks' customer service chats (CustomerSvc). A concrete example of an infiltration by a bot happened on Reddit in 2014, where the bot wise shibe provided automated answers and users rewarded the bot with tips in the digital currency dodgecoin, convinced they were tipping a real user (WiseShibe). SMS messages have been used by a bot to induce users to spend money on sexual content, pretending to be a real woman (SMSsex). Spam bots (SpamBot) may cause an economic harm in terms of time lost to process and clean messages.
- *Social harm* occurs when someone's image or standing in a community gets affected negatively, e.g., due to the publication of confidential and private information like a disease. An example of a bot causing social harm was documented by Jason Slotkin whose Twitter identity was cloned by a bot, confusing friends and followers (JasonSlotkin6). Similarly, Jamie Joung's Instagram account was cloned and kept alive by a bot (InstaClone). A reporter for the Daily Beast, Joseph Cox, was banned from Twitter for being followed too quickly by an army of colluding bots (ColludingBots). But also bot owners may incur social harm: Geico, Puma and Oreo had to publicly apologize for their bots respectively engaging with racist users (Geico), re-tweeting offensive content (Puma), and answering tweets from accounts with offensive names (Oreo).
- *Democratic harm* occurs when democratic rules and principles are undermined and society as a whole suffers negative consequences, e.g., due to fake news or the spreading of misinformation. Bessi and Ferrara[3], for instance, showed that bots were pervasively present and active in the on-line political discussion about the 2016 U.S. Presidential election (predating Robert S. Mueller III's investigation into the so-called Russian meddling). Without trying to identify who operated these bots, their finding is that bots intentionally spread misinformation, and that this further polarized the on-line political discussion (PolarBot). A specific example is that of Seth Rich, a staff member of the Democratic National Committee, whose murder was linked to the leaking of Clinton campaign emails and artificially amplified by bots (SethRich). Newsweek reported that nearly half of Trump's Twitter followers are fake accounts and bots; being the number of followers on social media a common measure of "importance," this obviously misrepresents reality (Trump).

What these examples show is that as long as there are bots interacting with humans, there will be the risk of some kind of harm, independently of whether it is caused intentionally or unintentionally. Thus, the key question is: is it possible to prevent people from getting harmed by bots?

PREVENTING HARM

Harm is caused by *abuse*, i.e., by inappropriate actions. Speaking of bots, examples of abuse are writing indecent or obscene messages (as in the case of Boost Juice's creepy bot) or making false allegations in a message (like in the case of Seth Rich's murder). Abuse thus consists of two components, an *action* by the bot and the satisfaction of some condition of *inappropriateness*.

Bot actions

As for the *actions*, we observe that by now bots can mimic human users in all forms of on-line communication, from live chats to social network messaging. This means that bots are not limited to posting a message in a chat room or on Facebook only. They can like posts written by others, comment on them, follow/unfollow users, etc. In short, bots today can perform all the social networking actions available to human actors. Technically, if a platform is willing to host bots, it typically supports their development via suitable Application Programming Interfaces (APIs) that enable developers to enact actions programmatically. If instead a platform does not want to host bots, it does not provide APIs; yet, bot developers may still fall back for example to client-side

Web automation tools, such as Selenium (https://bit.ly/2EwbinT), which allow them to mimic interactions by regular users with the user interface of the platform.

On the left-hand side of Figure 1, we provide a taxonomy of the actions we identified for the selected examples. Actions are grouped into chat, posting, endorsement and participation actions and are platform-specific (e.g., users chat on Messenger, while they write posts on Facebook). It is evident that none of these actions is bot-specific and that they can as well be performed by humans. It is also clear that these types of actions per se do not yet represent any abuse; they rather explain how on-line conversations happen. Then, they can be used for good or bad.

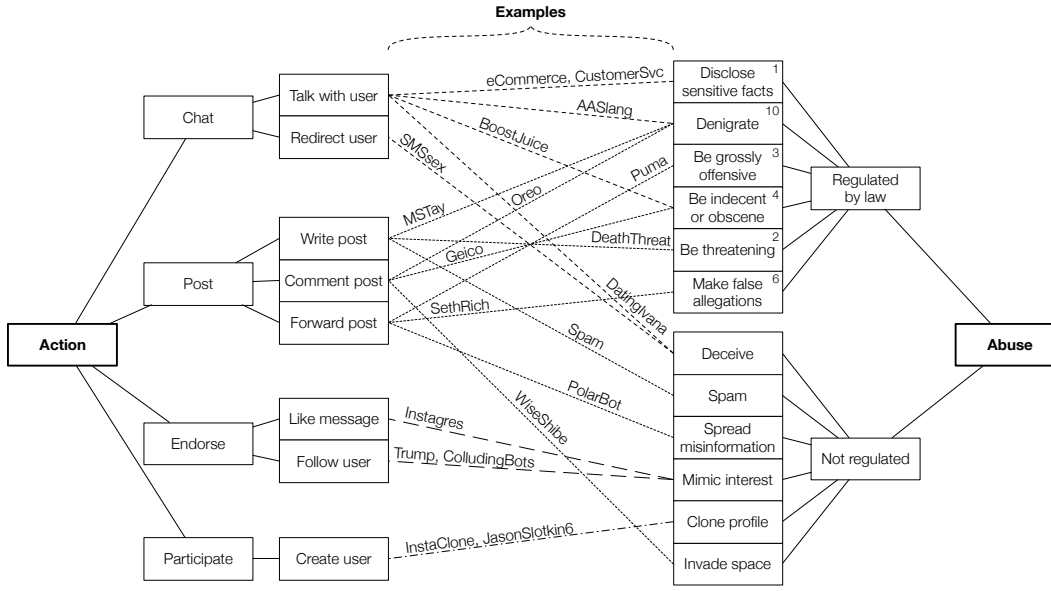

Figure 1. Actions by bots and identified types of abuse by example; numbers in boxes reference the violated communication principle (New Zealand's Harmful Digital Communications Act of 2015, see side-box).

Inappropriateness

As for the condition of *inappropriateness*, it is harder to provide a taxonomy of what makes an action inappropriate or not. Some types of abuse are actually subject to legal prosecution and thus formalized in laws, while others do not. These latter are breaches of moral, ethical or social norms that, although not prosecutable by law, may still cause harm. For example, threatening someone may be illegal while spreading misinformation and manipulating public opinion, alas, is not.

The regulation we found that most clearly tells which kinds of conditions must hold for an action to turn into an abuse is reported in the side-box, i.e., New Zealand's Harmful Digital Communications Act of 2015. Section 6 of the Act lists ten communication principles that, if violated, may result into an abuse. As for the types of harm, we systematically mapped the selected examples to the ten principles: In the top-right of Figure 1, we report the six principles for which we found concrete examples of violations, including the number of the principle; fortunately, we were not able to find any cases of harassment (principle 5), publication of matter that was published in breach of confidence (7), encouragement to send messages intended to cause harm to others (8), or incitements to commit suicide (9).

For the other six principles, we found compelling examples: the eCommerce and CustomerSvc bots envisioned aim to steel credit card data, that is, they *disclose sensitive facts* (1) that a user reveals inside a confidential chat room. AASlang, MSTay and Oreo are examples of *denigration* (10), respectively in the form of discrimination, hate speech, and answering tweets with offending account names. The Puma incident forwarded *grossly offensive* messages (3), while the

Geico incident, where the bot actively engaged with racist users, can be classified as *indecent and obscene* (4). The Dutch DeathThreat case is an example of a bot *threatening* (2) people with its messages. An example of bots making *false allegations* (6) is provided by the SethRich case, where bots amplified the conspiracy theory about Seth Rich's dead.

To the best of our knowledge, none except the DeathThread case actually led to any legal action. Yet, in some cases the owners of the bot had to publicly excuse themselves for the incident.

The systematic mapping of examples to communication principles allowed us to label 10 out of 20 examples; for the remaining examples, we manually assigned suitable labels expressing a type of abuse even in absence of a dedicated regulation: DatingIvana and SMSsex are clear examples of bots *deceiving* users, the former apparently without any malevolent purpose, the latter redirecting the user to a website that asks them to spend money. *Spamming* (SpamBot) is commonly regarded as abuse of one own's contact points, be it via email or through social networks. The *spreading of dis-/misinformation* (e.g., fake news) is a highly debated topic today, and Bessi and Ferrara[3] have shown that amplifying misinformation effectively impacts on the social perception of phenomena (PolarBot). Very similar to the spreading of dis-/misinformation is the practice of artificially boosting perceived interest in content or users, e.g., the Instagres and Trump cases show how bots are used to *mimic interest* by the community. The ColludingBots example even shows how following a user in a coordinated fashion may cause the social network to temporarily suspend the unaware user. Unaware were also the users whose social profiles were *cloned* by a bot, like in the InstaClone and JasonSlotkin6 examples.

It is important to note that in all these examples, bots were actually allowed to perform the actions that produced the described abuses. As explained earlier, most of the environments hosting conversations allow the enactment of actions via programmable APIs. Only the bot WiseShibe posted messages in discussions where only human actors were expected, causing harm to the Reddit group and its users.

Prevention

Now, in order to prevent abuse (and harm), different strategies may be adopted. The easiest one is simply *banning* bots from a platform (as with WhatsApp). However, not all bots are there to cause harm (e.g., they may automatically summarize sports news or help automate the work of marketing personnel), and not all bots that cause harm do so intentionally (see the Geico, Puma and Oreo examples). A second option is to require bots to explicitly *declare* that they are not human. This would prevent misunderstandings like in the DatingIvana or WiseShibe examples. Yet, there is no guarantee that bot owners would always abide by this requirement, and it nevertheless would not be able to prevent bots from offending or threatening people. The problem seems more related to *detecting and blocking* spam emails, which typically relies on content analysis (e.g., an email containing a given product name) and behavior analysis (e.g., a web server detecting bulks of similar emails sent from a same account).[9]

Content analysis techniques like natural language processing (NLP) and machine learning, but also crowdsourcing may help identify cases of bots deceiving or denigrating users, being grossly offensive, indecent, obscene, or threatening, in order to prevent psychological or legal harm. Identifying bots disclosing sensitive facts, making false allegations, or spreading misinformation is harder and could be approached by techniques like fact checking and crowdsourcing, helping to prevent potential economic, legal, social or democratic harms.

Behavior analysis techniques like social network/graph analysis (e.g., analyzing likes and follow relationships), botnet or malware detection, or generic network traffic analysis may help identify bots that mimic interest in messages or people, spread misinformation, or spam. CAPTCHAs could be used to challenge suspicious accounts, blocking bots automatically cloning accounts or invading spaces not meant for them. These techniques seem especially appropriate to prevent democratic harm particularly exposed to colluding or content spreading bots.

Hybrid content-behavior approaches may take advantage of the benefits of both. Orthogonally to this classification, Ferrara et al.[4] group known bot detection techniques into graph-based (mostly

behavior), crowdsourcing-based (mostly content), and feature-/machine-learning-based techniques (hybrid).

If we look at the state of the art of bot detection in online communications, we find Botometer (https://bit.ly/2wcSrzu, formerly called BotOrNot), a publicly available service by researchers of Indiana University, USA, that tells how likely a given Twitter account is a bot. The service uses more than 1,000 features capturing user, content, sentiment, network, friends, and temporal properties[10] (the former three are content-related, the latter three are behavioral) to classify accounts using a Random Forest classifier trained on a dataset by Lee et al.[11] The dataset is the result of a seven-months deployment of 60 honeypots on Twitter that was able to collect 36,000 likely bots. SybilRank[12] is a similar service for Facebook that analyzes the social graph to spot accounts that "have a disproportionately small number of connections to non-Sybil users." In addition to bots and humans, Chu et al.[13] also study the case of cyborgs, i.e., bot-assisted humans or human-assisted bots, using a combination of features and machine learning (classification). The method has been evaluated by considering the accuracy of the decisions over 50,000 Twitter users. In relation to the used type of classification methods, Morstatter et al.[14] study how to increase recall, a typical problem in these kinds of classification problems.

We observe that most of the works on bot detection are limited to telling bots and humans apart and do not further tell if a given bot is also likely to cause harm or not. There are however also some works that focus on specific types of bots, which can be related to our types of abuses. For example, Ratkiewicz et al.[15] study information diffusion networks in Twitter and design a method that detects the viral spreading of political misinformation, while Cresci et al.[16] focus on social spambots and fake Twitter followers mimicking user interest.[17] Properly leveraged efforts by Pitoura et al.[18] on measuring bias in online information could allow bot developers and users to ensure compliance with governance rules and provide insights into the harmless use of bots.

CHALLENGES AND OUTLOOK

The recent scandal of Cambridge Analytica misusing private data of about 87 million Facebook users has fast led to restricted API data access to Facebook, Instagram and Twitter,[19] limiting the freedom of action of bots. The need to rule what software agents can and cannot do in on-line conversations is thus felt today more than ever before.

This article represents a first attempt at defining a conceptual framework that may lay the foundation for what we could call bot ethics. The taxonomy of harms derived from the selected examples demonstrates that bots may indeed cause damage, willingly or unwillingly, and the proposed separation of actions from the conditions that may make them inappropriate provides hints about how to prevent harm. Some abuses are regulated by law, such as threatening people, but a significant number of abuses is not, such as spreading misinformation. These latter actually represent the bulk part of bot traffic and, as we have seen, have the potential to undermine the rules our very society is built on. Spamming, spreading misinformation, mimicking interest in people or topics, and cloning profiles to make these actions look credible may cause democratic harm, e.g., by diverting the attention of lawmakers to topics of little interest to society as a whole or even by altering the outcome of elections. It seems that current digital communication laws tend to protect the individual, while they neglect the community.

Where law is not enough to protect users, it is still possible to implement technical solutions that are able to spot likely harmful situations. We have seen that doing so is not easy in general, and it may even be harder if the goal is preventing the types of abuses discussed in this article. We specifically identify the following challenges to be faced before suitable solutions can be proposed to platform providers and users:

- First of all, a *common understanding and terminology* of harm and abuse in relation with bots interacting with humans in on-line conversations must be developed. This article contributes to this point with a first analysis of the problem and a proposal of key concepts and terminology, however, without the claim of completeness. For example, as bots become more sophisticated in the actions they perform, it is important to agree on what exactly makes an action inappropriate. Also, suitable means to report abuse and to prosecute excess need to be put into place.

- Since a significant part of the techniques proposed to detect bots is based on some form of machine learning, it is of utmost importance that the community collects and shares *suitably labeled datasets* that allow the training of algorithms and the verification of performance. This is one of the most challenging tasks, as many of the abuses are not easily detectable. It is also crucial to assure that reported abuses are indeed caused by a bot, which may require suitable bot detection techniques in the first place.
- Next, it is necessary to *identify patterns/anti-patterns* (or just examples) of both harmful content spread by bots and respective malicious behaviors. Again, the more of these data are shared with the community, the better. The challenge is to identify harmful behaviors also from ephemeral evidence. While spam bots leave significant traces, offenses or denigrations may not be as systemic and, hence, be less present in online conversations.
- Finally, suitable *quantitative and qualitative studies* must be performed to validate the effectiveness of conceived solutions. The challenge is assuring that not all bots but only harmful ones are detected and, possibly, blocked; useful bots should not suffer any harm themselves.

As Capurro states in his article "Information Technology as an Ethical Challenge",[20] "we cannot consider technology merely as an instrument having no fundamental roots in our individual and social lives, i.e., in our history and our cultural practices. Instead of separating analytically technology from the life-world in which it is already embedded, we should try to look at it 'synthetically', i.e., trying to grasp the mutual dependencies between man, nature, and technology." We have shown that modern online communication does not involve and affect only man, nature, and technology, but also bots – in a variety of forms and with a variety of roles and purposes – posing not only technical but also ethical challenges to our society.

SIDEBAR: COMMUNICATION PRINCIPLES

A good example of how regulation aims to prevent harm from digital communications (independently of bots) is New Zealand's Harmful Digital Communications Act of 2015 (https://bit.ly/2HKAfCx). The act states the following principles that digital communication should satisfy in order for its issuer not to become subject to investigation and law enforcement, that is, in order not to harm:

1. A digital communication should not disclose sensitive personal facts about an individual.
2. A digital communication should not be threatening, intimidating, or menacing.
3. A digital communication should not be grossly offensive to a reasonable person in the position of the affected individual.
4. A digital communication should not be indecent or obscene.
5. A digital communication should not be used to harass an individual.
6. A digital communication should not make a false allegation.
7. A digital communication should not contain a matter that is published in breach of confidence.
8. A digital communication should not incite or encourage anyone to send a message to an individual for the purpose of causing harm to the individual.
9. A digital communication should not incite or encourage an individual to commit suicide.
10. A digital communication should not denigrate an individual by reason of his or her colour, race, ethnic or national origins, religion, gender, sexual orientation, or disability.

REFERENCES

1. Varol, O., Ferrara, E., Davis, C. A., Menczer, F. & Flammini, A. Online human-bot interactions: Detection, estimation, and characterization. arXiv preprint arXiv:1703.03107 (2017).
2. Inbenta Technologies Inc. The ultimate guide to chatbots for businesses. Tech. Rep., www.inbenta.com (2016).

3. Bessi, A. & Ferrara, E. Social bots distort the 2016 US Presidential election online discussion. First Monday 21 (2016).
4. Ferrara, E., Varol, O., Davis, C., Menczer, F. & Flammini, A. The rise of social bots. Commun. ACM 59, 96–104 (2016).
5. Lin, P., Abney, K. & Bekey, G. A. Robot ethics: the ethical and social implications of robotics (MIT press, 2011).
6. Lomas, N. Duplex shows Google failing at ethical and creative AI design. Techcrunch.com, May 10, 2018. https://techcrunch.com/2018/05/10/duplex-shows-google-failing-at-ethical-and-creative-ai-design/
7. Priest, G. L. Reanalyzing Bush v. Gore: Democratic Accountability and Judicial Overreaching. U. Colo. L. Rev. 72, 953 (2001).
8. Price, R. Microsoft is deleting its AI chatbot's incredibly racist tweets (2016). URL https://read.bi/2DedMYm
9. Momeni, E., Cardie, C. & Diakopoulos, N. A survey on assessment and ranking methodologies for user-generated content on the web. ACM Comput. Surv. (CSUR) 48, 41 (2016).
10. Davis, C. A., Varol, O., Ferrara, E., Flammini, A. & Menczer, F. BotOrNot: A system to evaluate social bots. In WWW 2016 Companion, 273–274 (2016).
11. Lee, K., Eoff, B. D. & Caverlee, J. Seven months with the devils: A long-term study of content polluters on twitter. In ICWSM (2011).
12. Cao, Q., Sirivianos, M., Yang, X. & Pregueiro, T. Aiding the detection of fake accounts in large scale social online services. In USENIX 2012, 15–15 (USENIX Association, 2012).
13. Chu, Z., Gianvecchio, S., Wang, H. & Jajodia, S. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? IEEE Transactions on Dependable Secur. Comput. 9, 811–824 (2012).
14. Morstatter, F., Wu, L., Nazer, T. H., Carley, K. M. & Liu, H. A new approach to bot detection: striking the balance between precision and recall. In IEEE/ACM ASONAM 2016, 533–540 (IEEE, 2016).
15. Ratkiewicz, J. et al. Detecting and tracking political abuse in social media. ICWSM 11, 297–304 (2011).
16. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A. & Tesconi, M. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In WWW 2017 Companion, 963–972 (2017).
17. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A. & Tesconi, M. Fame for sale: efficient detection of fake twitter followers. Decis. Support. Syst. 80, 56–71 (2015).
18. Pitoura, E. et al. On measuring bias in online information. ACM SIGMOD Rec. 46, 16–21 (2018).
19. Constine, J. Facebook restricts APIs, axes old Instagram platform amidst scandals. techcrunch.com (2018). URL https://tcrn.ch/2IxdhKX
20. Capurro, R. Information technology as an ethical challenge. Ubiquity 2008, 1 (2008).

ABOUT THE AUTHORS

Florian Daniel is an assistant professor at Politecnico di Milano, Italy. His research interests include bots/chatbots, social data analysis and knowledge extraction, service-oriented computing, business process management, blockchain. Daniel received a PhD in information technology from Politecnico di Milano. Contact him at florian.daniel@polimi.it.

Cinzia Cappiello is an assistant professor at Politecnico di Milano, Italy. Her research interests include data and information quality aspects in service-based and Web applications, Web services, sensor data management, and big data. Cappiello received a PhD in information technology from Politecnico di Milano. Contact her at cinzia.cappiello@polimi.it.

Boualem Benatallah is a Scientia professor at the University of New South Wales, Sydney. His research interests include Web services, process analytics, data curation, crowdsourcing and cognitive services. Benatallah has a PhD in computer science from Grenoble University, France. Contact him at b.benatallah@unsw.edu.au.