From Dataveillance to Datapulation: The Dark Side of Targeted Persuasive Technologies

Claude Castelluccia

From Dataveillance to Datapulation

The Dark Side of Targeted Persuasive Technologies

Claude Castelluccia

INRIA, France
claude.castelluccia@inria.fr

1 Motivations

Online services, devices or secret services are constantly collecting data and meta-data from their users. This data collection is mostly used to target users, customized their services or monitor them. This surveillance by the data, sometimes referred to as *Dataveillance*, is omnipresent and generates a lot of attention [6].

However nowadays, data and technologies are not only used to monitor people, they are also used to motivate, influence or shape their opinions or decisions online. The better understanding of users' behaviors combined with the capacity of building accurate psychological profiles create the opportunities to develop techniques to influence users online, by shaping their behavior. These technologies can encourage positive norms, such as fighting terrorist or racists propaganda online, or can be used to motivate users to drive more safely or economically, to eat healthier or to exercise more[1]. In this case, they are often referred to as "Persuasive technologies or profiling" by psychologists, designers or behavioral economists [11,3].

However, these "persuasive technologies" have also a dark side. They can constitute efficient and targeted informational weapons to deceive or manipulate users' opinions or behaviors maliciously, via fakes news, information disorder, psychological or media manipulation techniques [27]. We define, in the paper, the concept of *Datapulation*, manipulation by the data. Datapulation consists of "mediated" personalized manipulation techniques, based on information, created primarily to change the attitudes and behaviors of users, for malicious intends or intends that go against users' own interests. Datapulation can be used by commercial companies to increase profit [7,4] or by political parties to influence elections [15,8].

We argue that Datapulation can be dangerous for privacy, human rights and democracy, and deserves more attention by policy makers and researchers. The main goal of this paper is to define the concept of Datapulation, by formalizing how data can be used to manipulate our decisions. We believe this is an important step in order to address it properly.

2 Modeling Cognitive Systems

In order to tackle the problem of datapulation it is important to understand how humans learn and make decisions. *Cognitive Psychology* is the scientific discipline that tries to answer this fundamental question. It is a branch of psychology concerned with mental processes (as attention, language use, memory, perception, problem solving, creativity, and thinking) especially with respect to the internal events occurring between sensory stimulation and the overt expression of behavior.

2.1 A Simplistic Model inspired from Machine Learning

Since Machine Learning, and more specifically Deep Learning, mimics in some ways what goes into the human brain, we decided, in this report, to use the same high-level abstract model. We are aware that, despite their biological inspiration and performance achievements, these systems differ from human intelligence in crucial ways [17]. However, we believe it provides a good enough model to discuss cognitive attacks and security. Fig.2 shows a very simplistic model of how humans learn and make decisions. It is is composed of two main interleaved phases:

[1] "The new rules of persuasion", http://captology.stanford.edu/resources/article-new-rules-of-persuasion.html

– *Learning Phase:* During the learning phase, the human's brain, which implement the learning "algorithm", processes various types of information to generate a cognitive model. As a result, the generated cognitive model depends on human internal brain, but also his various experiences and past history.
– *Decision Phase:* During the decision phase, the cognitive model is solicited in order to make a decision and to take some actions.

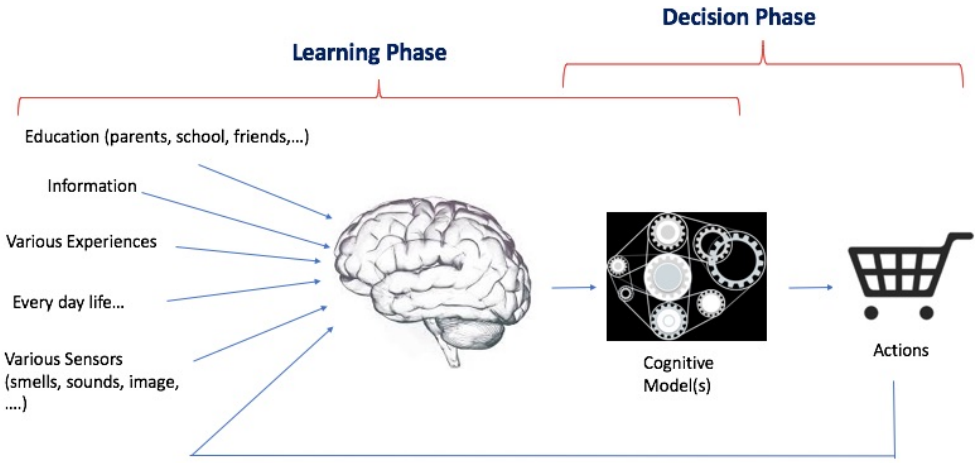

Fig. 1. A simplistic Model of Human learning and decision making

This model involve a feedback loop, because often the response change the environment and hence create new information to be processed. This is also an important phase of the learning process.

We consider, in this paper, the cognitive model as a "black box". However, we believe that it is useful to consider the Information Processing metaphor, commonly used by cognitive psychologists.

2.2 The Information Processing (IP) Metaphor

The advent of computers offered psychologists a metaphor between how computers and our minds work. This *computational metaphor* that is dominant in cognitive psychology considers that the mind is governed by programs or set of rules analogous to those which govern computers. Just like a computer, the human mind takes in information, processes, organizes and stores it to be retrieved at a later time[2]. Researchers that study the human information processing system often consider the generic and simple functional model shown in Fig. 2 [29] to model human cognitive systems. This model is composed of 3 main phases: the *perceptual encoding*, *cognitive* and *responding* stages.

1. The *perceptual encoding:* Whenever a human interacts with his environment, different types of sensory information is received by his various receptors (sight, hearing, smell, taste and feeling). This information is then stored in the sensory register for a short period of time. It is then processed by the *perception* stage that adds meaning to it by comparing it to information stored in the long term memory.
2. *The Cognitive Phase:* The resulting, augmented, information is then processed by the "central processing" unit that either generates an immediate response, selected from a broader array of possibilities, or inputs it into the "decision making" module. This module further processes the information, solves problems, and generates a response. The information is also use to update the decision making process and the memories: the process of *cognition*.

[2] "The validity of the IP metaphor in today's world is generally assumed without question. But the IP metaphor is, after all, just another metaphor - a story we tell to make sense of something we don't actually understand." `https://aeon.co/essays/your-brain-does-not-process-information-and-it-is-not-a-computer`.

It is interesting to notice that these 2 types of cognitive processing is similar to the Kahneman's Dual Systems model. In fact, Daniel Kahneman, argues that our cognitive processes are divided into two systems [13]. System 1 produces the fast, intuitive reactions and instantaneous decisions that govern most of our lives. It uses many of our cognitive biases. System 2 is the deliberate, rational type of thinking involved in focus, deliberation, reasoning or analysis, such as calculating a complex math problem, exercising self-control, or performing a demanding physical task.

3. *The Action Phase:* Finally, once a response is selected, the brain sends motor signals for actions.

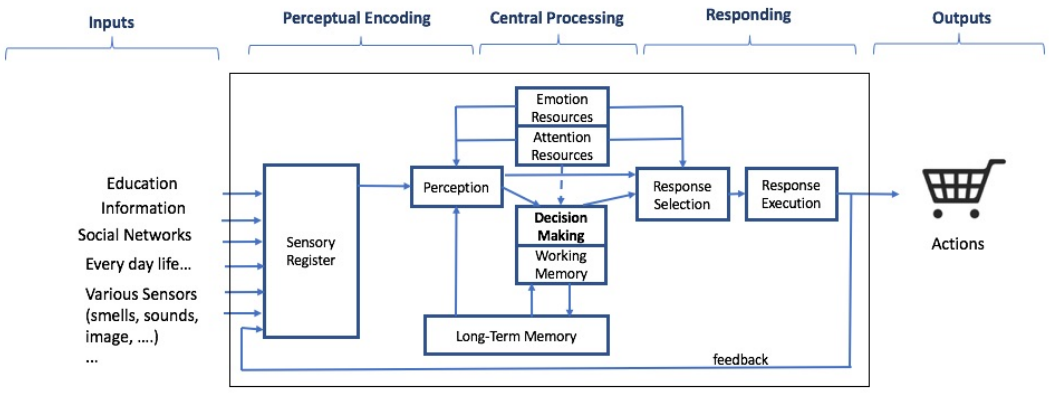

Fig. 2. A Generic Model of Information Processing, "An Introduction to Human Factor Engineering"

Attention is also vital component for much of the information processing. It plays an important role in the perception phase by filtering the important information, and provides the necessary mental resources to the various stages of information processing.

3 Hacking the brain

3.1 Attack Surface and goals

Attacks surface. This paper only considers "online" attacks, where the adversary (ADV) does not have physical contacts with the target. We consider a "black-box" model where the adversary has very limited knowledge about how the system is structured and has no (direct) access to the internal modules of the target cognitive model. For example, for obvious reasons, the adversary cannot have physical access nor modify the perception, attention or decision making modules of the system. We instead assume that the ADV can "only" interact with the inputs and outputs of the target cognitive system (see Fig3).

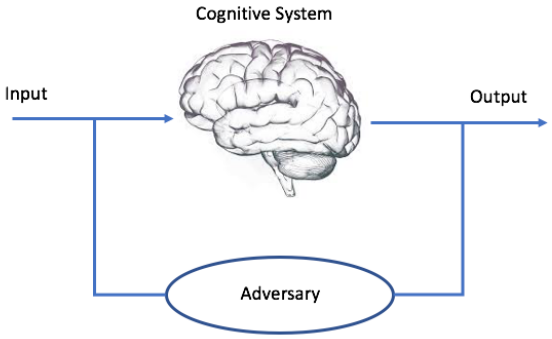

Fig. 3. Attack Surface

More specifically, we assume that the ADV can potentially[3]:

1. monitor, modify/block *inputs* and inject data into the system. For example, the ADV can inject fake information or block access to some content (censorship).
2. monitor the target's *outputs* (response selection and execution), but can not modify, nor block them . For example, a service, such as Facebook, can access and potentially modify the information (fake news) that the target receives. It can also monitor some of the target's outputs, such as posts, but we assume that it does not modify them (although technically some could do it).
3. monitor the target's "physical" behavior, such as movements or reactions, remotely (via camera, sensors)[4]. However, the adversary does not have physical access to the victim. For example, the adversary can not physically threaten the subject or inject drugs.

Depending on his strenghts, an ADV can (a) have access to some pre-defined {input; output} couples of the model[5], and/or (b) test the cognitive model with his own inputs and access the generated outputs[6]. In this case, the ADV is very powerful since she can modify/inject data, look at the generated outputs and adapt her attack strategy dynamically.

Attacks Goals. An important part of threat modeling is to define the adversary goals. Why is the adversary attacking the cognitive system? What is he trying to achieve? The CIA triad, commonly used in information security, defines 3 specific goals: Confidentiality, Integrity and Availability.

- *Confidentiality:* Attacks on confidentiality attempt to extract "private" information about the target's system. In the context of a cognitive system presented by Fig.2, the adversary might be interested in *retrieving* information about the target's perception, attention, memory or decision making modules The main goals are often to predict the target's decisions, behaviors or actions or to personalize attacks. These confidentiality attacks affect the privacy and the autonomy of the victim.
- *Integrity:* The main goal of the integrity attacks is to influence the target's decisions or actions. For example, the goal might be to influence a victim to buy a product or service that he does not really need, or influence his vote for a specific candidate. In other words, the goal of the ADV is to add some biases into the target's decisions.
- *Availability:* The main goal of the availability attacks is to confuse the target, by "noising" his cognitive system. It is to prevent the system from functioning "normally", or as "expected"Availability attacks attempt to reduce the quality or performance of a system. Propaganda attacks, such as used by Russians, are examples of such attacks. In fact, as noted by Paul and Matthews, the main goals of Russian propaganda are to entertain, confuse and overwhelm the audience [23].

[3] The attacking capabilities will, of course, depend on the power of the ADV. A government will, for example, be more powerful than a simple hacker.

[4] This opens, for example, the use of Eulerian video magnification to identify stress level of target (`https://cacm.acm.org/magazines/2017/1/211095-eulerian-video-magnification-and-analysis/fulltext`

[5] Similarly to known plaintext attacks in cryptography

[6] Similarly to chosen plaintext attack in cryptography

Note that these attacks might be targeted (the targets are predefined) or indiscriminate (the targets are random).

3.2 Confidentiality Attacks

Attacks on confidentiality attempt to extract "private" information about the target's system. An adversary can learn a lot about a person from his writing, speech, actions or behaviors. For example, text, speech, tweets and any posts of a person can be used to analyze his sentiments, views or emotions [18], mobility activities/traces can reveal depression [22], search queries can be used to detect neurodegenerative disorders [28] or "skin" monitoring can reveal stress level [26].

Human's behaviors, actions or posts can reveal a lot of information about their personality. For example, Chittarajan et al. showed that smartphone usage features, such as average number of calls or SMS, could be predictive of personality traits [5]. Oliveira et al. also investigated the efficacy of using call detail records (CDR) and Social Network Analysis (SNA) of the call graph to infer personality factors [9].

Humans have individual differences in their susceptibility to decision biases such as overconfidence, temporal discounting, and bias blind spot[7]. Such differences depend on the psychometrics profiles of humans. For example, research has shown that neurotic people have a cognitive bias toward home ownership instead of renting [8]. Psychometrics is a field of psychology that is devoted to measuring personality traits, aptitudes, and abilities. Psychological researchers often use a five-factor model to evaluate what are believed to be five core aspects, or traits, of an individual's personality. This framework is a hierarchical model of personality traits that represent personality at the broadest level of abstraction. It consists of five factors, namely extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience [21].

These factors summarize several more specific traits and are believed to capture most of the individual differences in human personality. They do not provide completely exhaustive explanations of personality, but encompass a large portion of personality-related terms. For example, the factor agreeableness includes, on the positive side, terms like generosity, amiability, and warmth and, on the negative side, aggressiveness and temper.

Using a questionnaire-based testing, composed of questions such as "Do you believe in the importance of art?" or "Do you seldom feel blue?", psychologists measure the degree to which each of these traits is individually expressed[9]. Some recent research results showed that people's psychological characteristics can be accurately predicted from simple digital footprints, such as their temporal patterns of smartphone and personal computer activity or their Facebook Likes [12,30]. It was established that computers' judgments of people's psychological characteristics based on their digital footprints are more accurate and valid than psychological characteristics defined by their close others or acquaintances (friends, family, spouse, colleagues, etc.). These findings demonstrate that people's personalities can be predicted automatically, at a large scale, with little (meta)data and without involving human social-cognitive skills.

This online and mass profiling creates new opportunities, but significant privacy risks. For example, some insurance companies are already using social network data in order to determine drivers or risk profiles[10]. Users then get profiled on what they could do, and not, as usual, on what they have done. This could lead to possible discriminatory decisions.

3.3 Integrity/Availability attacks

Although the goals of integrity and availability attacks are different, the security attacks to achieve them are very similar. That's why we address them together in the rest of this section. The 4 types of integrity attacks, namely pollution, physical, cognitive model exploitation and execution attacks, are presented on Fig. 4 and detailed in the rest of this section. Note that the Physical attacks consist of attack that physically modify the learning algorithm (i.e. the brain here). Since we are only considering online attacks in this paper, this category will not be developed in this paper.

[7] https://en.wikipedia.org/wiki/List_of_cognitive_biases

[8] http://www.science20.com/news_articles/cognitive_bias_why_neurotic_people_make_this_real_estate_choice_more-150409

[9] See https://sapa-project.org

[10] https://medium.com/privacy-international/social-media-intelligence-and-profiling-in-the-insurance-industry-4958f

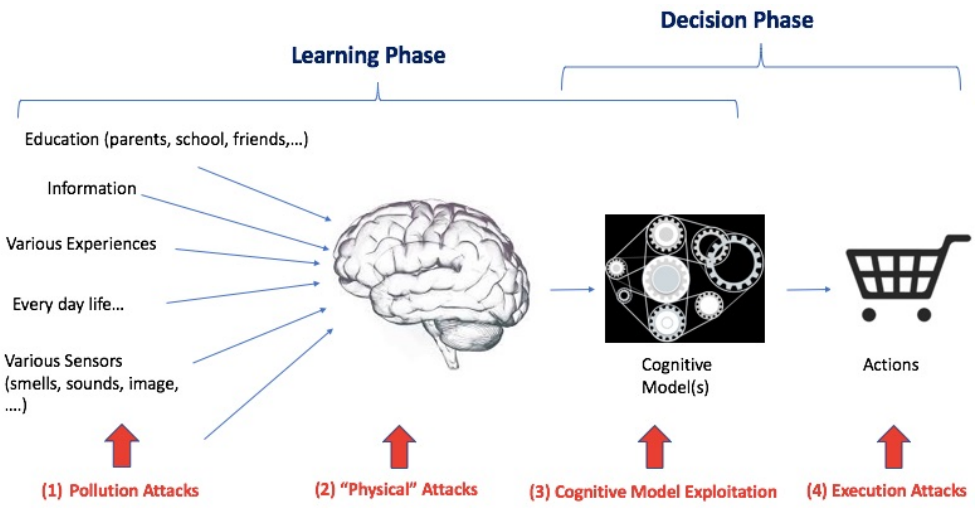

Fig. 4. Integrity/Availability Attacks

Information Disorder Attacks. Information disorder attacks aim at modifying the target cognitive systems, but manipulating the "training" data, i.e. the input data that is used to generate the cognitive model. In fact, as shown previously, the cognitive model depends on the learning "algorithm" and the input information (see Fig.1).

In a recent report, the council of Europe defines 3 types of information disorder[11]: Mis-information, dis-information and mal-information.

Mis-information: consists of information or content that is misleading, possibly false, often created by negligence, mistake or ignorance, but was not meant to harm. The misconception that the earth was flat by ancient cultures is an example of Mis-information that spread out because of the lack of knowledge of people.

Dis-Information/Fake news consists of information that is intentionally created and shared to cause harm. This is what is typically referred to as "fake news". There are different types of dis-information, such as imposter content, fabricated content, false connection, false context or manipulated content. An example of dis-information, or fake new, is the information that circulated just the presidential election is France in 2017 that Saudia Arabia was financing Macron's campaign. This news was published on a bogus website resembling the site of the Belgian newspaper "Le Soir". This information, that was proved to be incorrect, was clearly created to harm the French election.

Mal-Information is information that is genuine, based on reality, but that is shared to cause harm, often by moving information designed to stay private into the public sphere. For example, the leak of stolen or private emails, during a electoral campaign, is a good example of mal-information. Information manipulation techniques that consists in blocking, filtering/censoring, re-ordering or suppressing some specific information, fall also into this category, since, technically, no incorrect and fake information is propagated. For example some form of propaganda used by some states, that overwhelms the population with positive and correct news, sometimes called "cheer-leading" content, in order to eclipse bad news by diverting attention, is also another form of mal-information[12].

Of course, in reality, information disorder can fall into several of these categories, such as attacks that insert fake news into legit news articles, or that insert fake emails into legit stolen emails.

Exploitation Attacks. Exploitation attacks, as opposed to information disorder attacks, do not aim at modifying the target's cognitive system but rather to exploit some of its weaknesses.

[11] https://shorensteincenter.org/information-disorder-framework-for-research-and-policymaking/
[12] https://www.vox.com/world/2017/8/2/16019562/china-russia-internet-propaganda-media

As shown by Fig.5, a cognitive system can be exploited by manipulating the perception [1], emotion [16], attention[13] or memory modules [20,19].

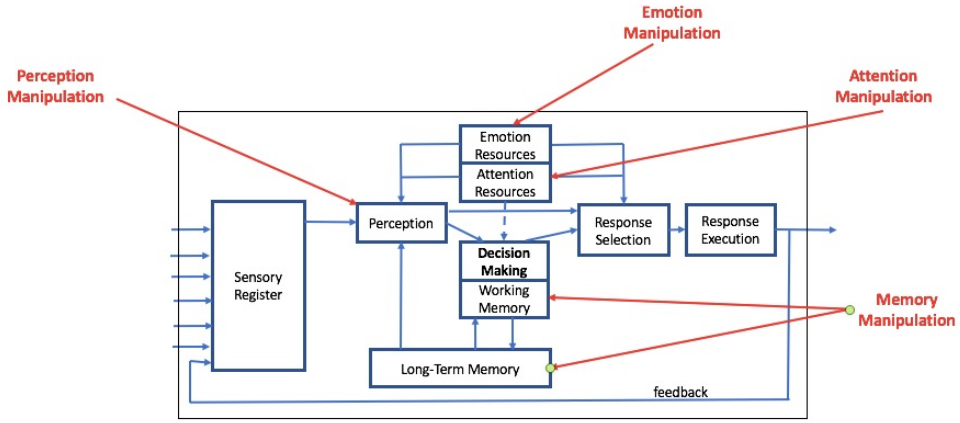

Fig. 5. Cognitive Model Exploitation

A good example is the controversial experiment performed by Facebook that showed that emotion does impact decisions [16]. More specifically, the Facebook study manipulated the News Feeds displayed to more than 600 thousand Facebook users, and show that it could affect the content posted by these users. Negative News Feeds led to more negative messages, whereas more positive News Feeds led to more positive messages. This results are important, and worrisome, because it shows that people's emotion can be manipulated remotely and at scale by social network operators, governments, advertisers or hackers by propagating fakes news/contents or "likes" [14].

Another example, based on the manipulation of the perception module, was demonstrated by Bailenson and al. [1]. This study showed that that voters demonstrate a strong preference for facially similar candidates. This was obtained experimentally by morphing, surreptitiously, the faces of political candidates with the faces of the experiment's participants as shown in Fig. 6.

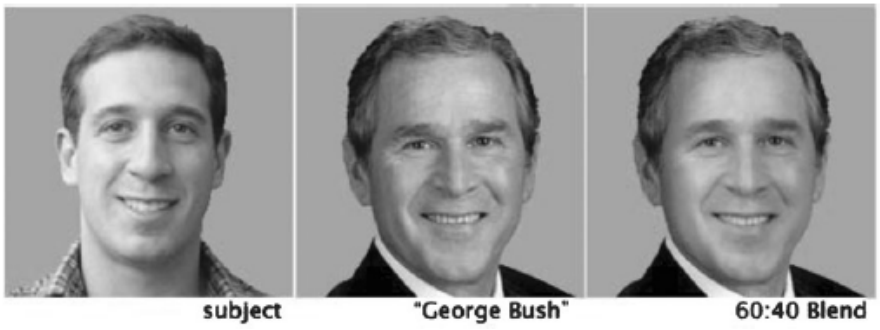

Fig. 6. An Example of A Subject Morphed with Bush (extracted from [1]).

The results showed a significant impact of the votes, in favor of the "morphed" candidates, that could be pivotal in closely contested election. It is noteworthy that these image manipulations are quite trivial

[13] https://www.wsj.com/articles/how-smartphones-hijack-our-minds-1507307811

and inexpensive to perform, and are difficult to detect by the subjects if the morphing ratio is below 40 percent (60 percent of the candidate and 40 percent of the participant). An extensive social science literature demonstrates that people are often drawn to others perceived as similar [2].

Similarly, Epstein and Robertson showed that Internet search rankings, and in general how information is presented to users, have a significant impact on user choices, mainly because users trust and choose higher-ranked results more than lower-ranked results. They demonstrated, experimentally, that Internet search rankings can be manipulated to alter the preferences of undecided voters in democratic elections [10].

These attacks exploit human cognitive biases in order to influence the target's decisions and guide his intends. These attacks are similar to "adversarial examples" in ML, which correspond to fabricated inputs that get misclassified by a given classifier. Micro-targeting, as used by Cambrige-Analytica, is an example of such cognitive attacks. By customizing an input to a victim's psychological profile, and therefore to his specific cognitive biases, an adversary can be able to guide and control his decision.

Execution Attacks Attacks that prevent a victim from achieving his objectives. These attacks do not attempt to modify the victim's intends, but rather to prevent him from achieving them. Examples of such attacks are (1) social engineering or phishing attacks that mislead the victim into doing things they don't want to do or (2) Dark Patterns that deceive users by making some tasks, such as the access to a privacy configuration menu, artificially difficult. Dark patterns can be described as "...features of interface design crafted to trick users into doing things that they might not want to do, but which benefit the business in question."[14]. They, therefore, do not manipulate the target's decisions but make its execution difficult via various tricks such as nagging, obstruction, sneaking, interface interference or foced action (see Fig. 7).

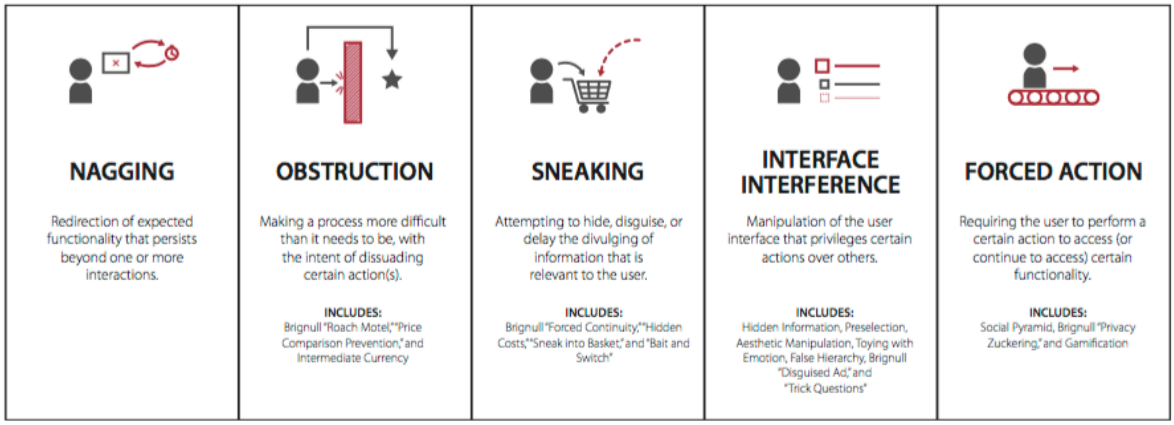

Fig. 7. Dark pattern strategies (from `http://colingray.me/...`)

4 Conclusion

This paper is a first attempt to conceptualize the field of datapulation. Understanding existing cognitive attacks is the first step to address them efficiently. Although the press is mostly talking about fake news, we show, in this paper, that the attack space is much larger. Cognitive attacks affect the confidentiality, integrity and availability of cognitive systems. They are often very powerful because, as shown by the Cambridge Analytica case, they rely on AI and are dynamic, highly targeted, can be performed at scale and are, most of time, fully automated. Furthermore, they don't not require a lot of resource since they are usually performed via existing services and social networks. As a result, they are likely to develop in the coming years.

[14] see "Deceived By Design" `https://fil.forbrukerradet.no/wp-content/uploads/2018/06/2018-06-27-deceived-by-design-final.pdf`).

It becomes urgent for our democracies to address the problem of datapulation. Datapulation is a very complex, and by nature, an interdisciplinary disciplinary field of research. It requires expertise in (at least) psychology, sociology, behavioral economics, political science, computer science, philosophy and law.

The first challenge is to define and formalize the concept of manipulation. As discussed by Sunstein, there exists "Fifty shades of manipulation"[24]. Some of them can be considered ethical, others are clearly not acceptable [25]. It is essential to clearly define, and possibly formalize, the different forms of online manipulation. It is also essential to understand the different ways of manipulating users online and analyse the different attacks and upcoming ones (such as the deepfakes). As shown in this paper, online attacks are comply and rely on misinformation, cognitive bias exploitation or dark patterns to manipulate votes or, more generally, opinions.

We believe that datapulation should be addressed using an holistic approach. Technical solutions should be develop to limit the propagation of misinformation, detect and possibly block manipulation attacks. However, technical solutions won't solve the problem completely and some regulation are also necessary. For example, it is probably necessary to prohibit some online manipulative techniques[15] and regulate online targeting and profiling. This requires to define a whole new research agenda.

References

1. Bailenson, J. N., Iyengar, S., Yee, N., and Collins, N. A. Facial similarity between voters and candidates causes influence. *Public Opinion Quarterly 72*, 5 (2008), 935–961.
2. Baumeister, R. *Handbook of social psychology.* McGraw-Hill, 1998, ch. The self.
3. Berdichevsky, D., and Neuenschwander, E. Toward an ethics of persuasive technology. *Commun. ACM 42*, 5 (May 1999), 51–58.
4. Calo, R. Digital market manipulation. *George Washington Law Review* (2013).
5. Chittaranjan, G., Blom, J., and Gatica-Perez, D. Mining large-scale smartphone data for personality studies. *Personal Ubiquitous Comput. 17*, 3 (Mar. 2013), 433–450.
6. Clarke, R. Information technology and dataveillance. *Communications of the ACM 31*, 5 (1988), 498–512.
7. Clifford, D. Citizen-consumers in a personalised galaxy: Emotion influenced decision-making, a true path to the dark side? Available at SSRN: `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3037425`, September 2017.
8. Confessore, N., and Hakim, D. Data firm says 'secret sauce' aided trump; many scoff. New York Times, March 2017. `https://www.nytimes.com/2017/03/06/us/politics/cambridge-analytica.html`.
9. de Oliveira, R., Karatzoglou, A., Concejero Cerezo, P., Armenta Lopez de Vicuña, A., and Oliver, N. Towards a psychographic user model from mobile phone usage. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA, 2011), CHI EA '11, ACM, pp. 2191–2196.
10. Epstein, R., and Robertson, R. E. The search engine manipulation effect (seme) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences 112*, 33 (2015), E4512–E4521.
11. Fogg, B. J. The human-computer interaction handbook. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 2003, ch. Motivating, Influencing, and Persuading Users, pp. 358–370.
12. Grover, T., and Mark, G. Digital footprints: Predicting personality from temporal patterns of technology use. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers* (New York, NY, USA, 2017), UbiComp '17, ACM, pp. 41–44.
13. Kahneman, D. *Thinking, fast and slow.* Farrar, Straus and Giroux, 2011.
14. Kates, G. "collusion network" facebook flaw leads to millions of fake "likes". CBS News, September 2017. `https://www.cbsnews.com/news/collusion-network-facebook-flaw-leads-to-millions-of-fake-likes/`.
15. Knott, A. Uses and abuses of AI in election campaigns. `https://ai-and-society.wiki.otago.ac.nz/images/0/0f/Ai-and-elections.pdf`. Retrieved in september 2017.
16. Kramer, A. D. I., Guillory, J. E., and Hancock, J. T. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences 111*, 24 (2014), 8788–8790.
17. Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *CoRR abs/1604.00289* (2016).
18. Liu, B. *Sentiment Analysis and Opinion Mining.* Morgan & Claypool Publishers, 2012.
19. Loftus, E. F. Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory 12*, 4 (2005), 361–366.

[15] `https://www.theguardian.com/media/2019/jul/31/us-could-ban-addictive-autoplay-videos-and-infinite-scrolling-onli`

20. Loftus, E. F., and Pickrell, J. The formation of false memories. *Psychiatric Annals* (1995).
21. McCrae, R. R., and John, O. P. An introduction to the five-factor model and its applications. *Journal of Personality 60*, 2, 175–215.
22. Mehrotra, A., and Musolesi, M. Using autoencoders to automatically extract mobility features for predicting depressive states. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 2*, 3 (Sept. 2018), 127:1–127:20.
23. Paul, C., and Miriam, M. The russian "firehose of falsehood" propaganda model: Why it might work and options to counter it. Tech. rep., RAND Corporation, 2016.
24. Sunstein, C. R. Fifty shades of manipulation. *J. Marketing Behav.*.
25. Thaler, R. H. Nudge, not sludge. *Science 361*, 6401 (2018), 431–431.
26. Wadhwa, N., Wu, H.-Y., Davis, A., Rubinstein, M., Shih, E., Mysore, G. J., Chen, J. G., Buyukozturk, O., Guttag, J. V., Freeman, W. T., and Durand, F. Eulerian video magnification and analysis. *Commun. ACM 60*, 1 (Dec. 2016), 87–95.
27. Waltzman, R. The weaponization of information: The need for cognitive security. Tech. rep., RAND Corporation, 2017.
28. White, R. W., Doraiswamy, P. M., and Horvitz, E. Detecting neurodegenerative disorders from web search signals. *npj Digital Medicine 1*, 1 (2018), 8.
29. Wickens, C. D., and Carswell, C. M. *Information Processing*. 2006, pp. 111–149.
30. Youyou, W., Kosinski, M., and Stillwell, D. Computer-based personality judgments are more accurate than those made by humans. In *Proceedings of the National Academy of Science*.