Abstract

The growth of the Internet and the pervasiveness of Information and Communication Technology (ICT) have led to a radical change in our society, a deep economical, commercial and social impact on our lives. To date, most of our lives takes place online where algorithms shape and guide our behaviour and the governance of our societies.

One of the drawbacks of this change is an increased risk for Internet users about their personal information *privacy*. Indeed an enormous amount of data is being generated and disseminated by people at high pace, often without knowing who is recording what about them. Online browsing, banking, shopping, social network interactions, and any type of online economic, social, personal collaboration and communication could undermine the individuals' privacy due to a variety of factors that include not only the frightening increase of information leakage. Indeed, specific private information can be also inferred/extracted via computational heuristics applied on data (apparently unrelated to such information) users voluntarily disclose on the Internet.

In particular, such privacy leaks can be caused by both *(a)* applications or software users intentionally use unaware of the related risks, and *(b)* malicious (illegal or unfair) practices stealthy perpetrated by "adversaries". Therefore, securing private data, devices and user's privacy in the digital society has become an utmost concern for individuals, business organizations, national governments and researchers.

Given the complexity, inscrutability of the software users engage with and the number of potential attacks and unfair prac-

tices they can incur into, it becomes ever more clear the need for technology-driven safeguards supporting users in the detection and counteract of such threats. In this respect, *machine learning* (ML), with its pattern recognition capability, appears to be a precious ally capable of upholding not only users' privacy but also other rights threatened in the digital society.

In this dissertation, we *focus* on intelligent *privacy* safeguards for users in the digital society. Our *goal* is to explore, both on the theoretical and experimental levels, how ML approaches can be fit to support the protection of privacy and the related rights. The research draws also upon most recent development in the areas of computational law and techno-regulation, two research paradigms emerging, at planetary scale, on the boundaries between computer science and law.

The work is structured as follows. We first depict the theoretical and methodological framework of this work, through a systematic literature review framing the use of ML to protect users' privacy. In doing this, we trace back existing approaches and solutions for the privacy protections to two fundamental categories: *enforcement* (i.e., solutions which *impose constrains* and hamper breaches of norms) and *nudge* solutions (i.e., solutions which inform users and increase their *awareness* to promote privacy-oriented behaviors). We provide a comprehensive taxonomy of main areas, threats, ML methods, type of protection delivered analyzing 143 studies published from January 2017 to October 2020.

Then, we present a series of research activities exploring the applicability of ML-based approaches to the issues arising in the scenarios above described. The activities presented can be ideally split into two parts focusing on privacy protection and other related rights, respectively. In more detail, the first part encompasses two projects tackling the privacy protection in the strict sense.

a) ML for privacy enforcement

We deal with the long-standing issue of third-party tracking on the Web in which users' private data are unfairly stolen for marketing and malicious activities, such as online stalking. We experiment the use of ML to distinguish between trackers and functional resources on the Web, finding that such techniques can be fit for a high classification rate of such threat. The resulting ML-based approach has been implemented into *GuardOne*, a tool to protect users against third-party tracking, which provides *enforcement* solutions to block trackers. *GuardOne* has been evaluated in real-world against similar commercial privacy enforcing solution. The main features of *GuardOne* are: *(i)* a hybrid mechanism based on ML and blacklisting, *(ii)* customization based on the user browsing habits, *(iii)* a very lightweight implementation which does not impact on the users' devices performance compared to commercial solutions, *(iv)* a high effectiveness in detecting and blocking third-party trackers better than the vast majority of commercial solutions.

b) ML for privacy awareness

We deal with the issue of unaware and/or uncontrolled dissemination of personal and private data, in text format, on the Internet. We experiment the use of ML and advanced language processing techniques to support both the classification of the text topic (among the most sensitive ones, e.g., politics and health) and the sensitiveness of the content according to such topic, finding that the performance of our proposal in a simulated environment is comparable with solutions available in literature. Furthermore, we experiment how the ML solutions designed can be fit to learn the user's personal attitudes towards privacy. We then embed such ML-approaches into *Knoxly*, a tool to protect users against the dissemination of personal and/or private information online, which relies on *nudge*-based solutions. Specifically, *Knoxly* aims to raise awareness and promote privacy-oriented behavior by means of alerts/warnings. The main features of *Knoxly* are: *(i)* a Keyword module to detect common sensitive words and personal identifiable information, *(ii)* a Topic module to distinguish the text's topic, *(iii)* a Sensitiveness module to "measure" the sensitiveness of the text content, *(iv)* a Customized module allowing the user to personalize the warnings displayed, *(v)* an intuitive User Interface powered by Visual Analytics techniques, *(vi)* a lightweight implementation which does not impact on the users' devices performance.

The second part encompasses two other projects that exploit methodological and technological solutions and approaches identified in the previous research stage and expand the research scope by applying such insights to other rights, linked to privacy and of great relief in the digital society, i.e., child and consumer protection.

a) ML for child protection
We deal with the challenge of providing online (privacy) protections for a specific category of users, that is children, which, according to the General Data Protection Regulation (GDPR) and UNICEF, need *ad-hoc* safeguards. Within this research project, named *AI4Children*, we experiment several ML-based approaches for users identification which can be seen as the baseline to uphold legal standards for online child protection. In fact, once identified the user, it is possible to trigger the specific safeguards. In more details, the conceived approach (based on data integration techniques) is capable of recognizing the age of a user based on the touch gestures he/her performs on a mobile device with a high accuracy. The main features of *AI4Children* are: *(i)* distinguishing between adults and underages based on commonly performed touch gestures, *(ii)* using a small set of features and touch gesture to perform a high accurate classification, *(iii)* robustness in the classification on different devices.

b) ML for consumer protection
We deal with issue of unlawful clauses in Terms of Service online (ToS), that is clauses which directly threaten users' concrete interests for example regulating how data will be managed and the liabilities on such. We experiment the use of ML and advanced language processing techniques to support the classification of ToS clauses categories and fairness level, finding that our techniques overcome state-of-the-art solutions and can be used to measure the ToS unfariness. The conceived approach has been embedded

into *ToSware*, a tool to raise consumers *awareness* against unlawful practices in online ToS. The main features of *ToSware* are: *(i)* having a mechanism to measure the unfairness of online ToS, *(ii)* making ToS more easy to read thanks to different visualization techniques and visual metaphors evaluated by real users, *(iii)* a lightweight implementation which does not impact on the users' devices performance.

The dissertation ends up with considerations about the challenges for ML research in these specific areas, and the future perspectives unfolded by computational law and techno-regulation.