Title: On the assessment of Machine Learning Algorithms for Fairness

Supervisor: Mohamad Gharib (mohamad dot gharib at ut dot ee)
Co-supervisor: Modar Sulaiman (modar dot sulaiman at ut dot ee)

Artificial intelligence (AI)/Machine learning (ML) can be described as the art and science of letting computers learn to perform complex tasks without being explicitly programmed to [1]. This has led to a dramatic increase in AI/ML adoption in almost all the main domains of our lives. One main advantage of using AI/ML systems is making or assisting in making [critical] decisions. Unlike humans, who might have various biases that can influence their objective decisions, AI/ML systems were expected to make precise and objective decisions [2]. However, AI/ML systems have been proven to suffer from bias and discriminative behavior just like humans [3]. Examples of such biased behavior cover many AI/ML applications [4][5], and may have serious consequences when they occur in sensitive domains, where AI/ML decisions may influence essential human rights (e.g., the right to equality). That is why assuring AI/ML fairness has emerged as an important area for research within the ML community [6].

This has led to a growing interest among AI/ML researchers on the issue of fairness metrics, and vast number of metrics have been developed to quantify AI/ML. However, many recent works have identified limitations, inadequacies, and insufficiencies in almost all existing fairness metrics [7], given that there is no universal means to measure fairness, i.e., there are no clear criteria to assess which measure is the "best".

The aim of this thesis is to: (1) critically review available AI/ML fairness literature; (2) identify the strength and weaknesses of the best current approaches to measure fairness in AI/ML; (3) specify the requirements for developing new metric(s) that address inadequacies/insufficiencies in existing fairness metrics; and (4) implementing and testing adequate fairness metric(s) that satisfy the aforementioned requirements.

Note: for a comprehensive survey of fairness in machine learning, you can refer to [8].

References

[1] M. Gharib, P. Lollini, M. Botta, E. Amparore, S. Donatelli, A. Bondavalli, On the Safety of Automotive Systems Incorporating Machine Learning Based Components: A Position Paper, in: Proc. - 48th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Networks Work. DSN-W 2018, 2018: pp. 271–274. https://doi.org/10.1109/DSN-W.2018.00074.
[2] G. Sheshasaayee, Ananthi and Thailambal, Comparison of classification algorithms in text mining, Int. J. Pure Appl. Math. 116 (2017) 425–433.
[3] P. Molnar, L. Gill, Bots at the Gate: a human rights analysis of automated decision-making in Canada's immigration and refugee system, 2018.
[4] L. Sweeney, Discrimination in online Ad delivery, Commun. ACM. 56 (2013) 44–54. https://doi.org/10.1145/2447976.2447990.
[5] S.L. Blodgett, B. O'Connor, Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English, in: Fairness, Accountability, Transpar. Mach. Learn., 2017.
[6] A. Agarwal, A. Beygelzimer, M. Dudfk, J. Langford, W. Hanna, A reductions approach to fair classification, in: 35th Int. Conf. Mach. Learn. ICML 2018, 2018: pp. 102–119.
[7] Yao, Sirui, and Bert Huang. "New fairness metrics for recommendation that embrace differences." arXiv preprint arXiv:1706.09838 (2017).
[8] Caton, Simon, and Christian Haas. "Fairness in machine learning: A survey." arXiv preprint arXiv:2010.04053 (2020).

Title: From User Stories to Privacy-aware user stories

Supervisor: Mohamad Gharib (mohamad dot gharib at ut dot ee)
Co-supervisor: Kuldar Taveter (kuldar dot taveter at ut dot ee)

Requirements elicitation is defined as the process of uncovering, acquiring, and elaborating requirements for computer-based systems [1]. There is a general agreement in the Requirements Engineering (RE) community that requirements elicitation is one of the most critical activities in the RE process (e.g., [2]), since getting the right requirements is considered a vital success factor for software development projects [3]. Although there are several requirements elicitation approaches and techniques that have been proposed in the literature, including but not limited to: interviews, questionnaires, task analysis, workshops, prototyping, etc., user stories [4] become almost the standard method for eliciting requirements in the industry [5]. A user story is a short description of high-level stakeholders' requirements that is represented using a simple template such as "As a <role>, I want <goal>, so that <benefit>". User stories have been successfully used for eliciting functional requirements, yet they are still being criticized for appropriately eliciting non-functional requirements (NFRs) such as privacy, safety, reliability, etc., where the satisfaction of NFRs is essential for successful software projects.

Privacy has emerged as a key concern since such companies need to protect the privacy of personal information to comply with various privacy laws and regulations (e.g., GDPR in the EU) that many governments have enacted for privacy protection. Accordingly, dealing with privacy concerns is a must these days [6]. Like other NFRs, there is neither standard nor agreed upon user stories approach for eliciting privacy requirements. To this end, the main objective of this thesis is to develop, verify, and validate a privacy-aware user stories approach.

Note: the privacy ontology provided in [7] can be used to facilitate understanding and dealing with privacy requirements in the proposed approach.

References

[1] Didar Zowghi and Chad Coulin, "Requirements elicitation: A survey of techniques, approaches, and tools," in Engineering and managing software requirements.: Springer, 2005, pp. 19-46.
[2] Ian Sommerville, Software engineering 8: pearson Education limitd, 2007.
[3] Jones Carpers, "Applied Software Measurement: Assuring Productivity and Quality," McGraw-Hill, New York, vol. 17, no. 1, p. 2.
[4] Cohn, M.: User stories applied: for agile software development. Addison Wesley (2004)
[5] Lucassen, Garm, et al. "The use and effectiveness of user stories in practice." International working conference on requirements engineering: Foundation for software quality. Springer, Cham, 2016.
[6] Gharib, Mohamad, John Mylopoulos, and Paolo Giorgini. "COPri-a core ontology for privacy requirements engineering." International Conference on Research Challenges in Information Science. Springer, Cham, 2020.
[7] Gharib, Mohamad, Paolo Giorgini, and John Mylopoulos. "COPri v. 2—A core ontology for privacy requirements." Data & Knowledge Engineering 133 (2021): 101888.

A safety-aware architecture for Safety-Critical Systems incorporating Machine Learning components

Supervisor: Mohamad Gharib (mohamad dot gharib at ut dot ee)

Machine learning (ML) components are increasingly adopted in many automated systems. Their ability to learn and work with novel input/incomplete knowledge and their generalization capabilities make them highly desirable solutions for complex problems [1]. This has motivated many system manufacturers to adopt ML components in their products in many industrial domains (e.g., medical, automotive), performing complex tasks such as pattern recognition, image recognition, and even control [2]. However, some of these systems can be classified as safety-critical systems (SCS), where their failure may cause death or injuries to humans [3]. Accordingly, the performance of such ML components must be assessed and guaranteed to be compliant with the safety requirements of incorporating SCS. Although the area of system safety is well-established, and there exist various methods to identify potential components faults/failures along with countermeasures to eliminate or at least limit the consequences of such faults/failures. Most of these methods do not apply to ML components as they do not properly address the special characteristics of ML components such as non-determinism, non-transparency, and instability to mention a few [4].

The objective of this thesis is to propose general-purpose fail-controlled [5] software architecture for incorporating ML components into SCS. The architecture will adopt state-of-art system and safety engineering principles, and adapt them to address the special characteristics of ML components. The architecture should be able to identify when an ML component may fail to behave as expected and tackle hazardous situations resulting from such failure by implementing countermeasure mechanisms appropriate for the type of failure. The architecture will be validated by applying it to a real/realistic case study/scenario concerning an SCS.

Note: Section 3 in [6] provides a short description of fail-controlled software architecture.

References

[1] Z. Kurd, T. Kelly, and J. Austin, “Developing artificial neural networks for safety-critical systems,” Neural Computing and Applications, vol. 16, no. 1, pp. 11–19, oct 2007.
[2] J. Schumann, P. Gupta, and Y. Liu, “Applications of Neural Networks in High Assurance Systems,” in Neural Networks, 2010, vol. 268, pp. 1–19.
[3] M. Bozzano and A. Villafiorita, Design and safety assessment of critical systems. Auerbach Publications, 2011.
[4] Gharib, Mohamad, et al. "On the safety of automotive systems incorporating machine learning-based components: a position paper." 2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W). IEEE, 2018.
[5] Avizienis, A., Laprie, J. C., Randell, B., Landwehr, C., 2004. Basic Concepts And Taxonomy Of Dependable and Secure Computing. IEEE Transactions On Dependable And Secure Computing 1 (1), Pp. 11–33.
[6] Gharib, Mohamad, Tommaso Zoppi, and Andrea Bondavalli. "On the Properness of Incorporating Binary Classification Machine Learning Algorithms into Safety-Critical Systems." IEEE Transactions on Emerging Topics in Computing (2022).

Title: Towards an information type lexicon and taxonomy to improve informational self-determination

Supervisor: Mohamad Gharib (mohamad dot gharib at ut dot ee)

The monetary value of information, and especially Personal Information (PI), is large and growing, and many organizations have already started profiting from this trend. Accordingly, breaches and misuse of PI have increased [1]. For example, privacy merchants shadow Internet users to create very detailed profiles concerning their online behavior and activities. Then, sell these profiles to whoever pays the demanded price [2]. In response to that and other potential misuses of PI, many governments around the world have enacted laws and regulations for privacy/PI protection (e.g., the GDPR in the EU). However, these laws and regulations rely heavily on the concept of informational self-determination that is, usually, implemented through the notice and consent/choice model. A notice (e.g., privacy policy) is supposed to inform Data Subjects (DSs) about how their PI will be processed and shared, and a consent/choice is supposed to acquire a signifying acceptance at the DSs' side concerning the offered notice. Although notifying DSs about data practices is supposed to enable them to make informed privacy decisions, current mechanisms for presenting the notice and obtaining the consent are deeply flawed as indicated by many researchers. More specifically, most notices are long, complex, hard to comprehend, change frequently, do not, usually, precisely specify potential future use of PI, and most importantly they either do not specify what type of information/PI is subject to this notice or use very high abstract terms. To improve the understandability of notices (privacy policies) on DSs side, and allow future automated analysis of such notices, a well-defined taxonomy of information/PI types should be provided.

This thesis aims to: (1) construct a lexicon of information/PI by analyzing an appropriate number (e.g., 15) of privacy policies; (2) derive a well-defined taxonomy of information/PI from the information/PI lexicon; and (3) verify and validate the information/PI taxonomy by applying it to case studies from different domains and assessing its completeness for classifying information/PI.

Note 1: To get an idea of how an information type lexicon can be constructed, you can refer to [3].
Note 2: The information/PI partial taxonomies provided in [4] and [5], can be used as a reference for the taxonomy to be developed.

References

[1] Gharib, Mohamad, Paolo Giorgini, and John Mylopoulos. "Towards an ontology for privacy requirements via a systematic literature review." International conference on conceptual modeling. Springer, Cham, 2017.
[2] Etzioni, Amitai. "The privacy merchants: What is to be done." U. Pa. J. Const. L. 14 (2011): 929.
[3] Bhatia, Jaspreet, and Travis D. Breaux. "Towards an information type lexicon for privacy policies." 2015 IEEE eighth international workshop on requirements engineering and law (RELAW). IEEE, 2015.
[4] Gharib, Mohamad, Paolo Giorgini, and John Mylopoulos. "COPri v. 2—A core ontology for privacy requirements." Data & Knowledge Engineering 133 (2021): 101888.
[5] Gharib, Mohamad, and John Mylopoulos. "On the Philosophical Foundations of Privacy: Five Theses." IFIP Working Conference on The Practice of Enterprise Modeling. Springer, Cham, 2021.

An integrated approach for analyzing safety and security requirements for Cyber-Physical Systems (CPSs)

Supervisor: Mohamad Gharib (mohamad dot gharib at ut dot ee)

The increased digitization of traditional Physical Systems (PSs) gave birth to the so called Cyber-Physical Systems (CPSs), which integrate sensing, computational, and control capabilities into traditional PSs combined with network connectivity. Consequently, traditional security solutions, although well established and consolidated, might not be effective to protect CPSs against human planned, malicious, complex attacks, which are the typical modern cyber-security attacks. This is quite clear with the increasing number of cyber-security attacks that now can target some of the safety-critical functionalities of CPSs. For instance, modern automotive vehicles have been proven vulnerable to hacking attacks aiming at getting control over the safety-critical functions of the vehicle [1]. An example is the hijacking of the steering and braking units in a Ford Escape [2]. Similarly, hackers were able to remotely hijack a Tesla Model S from a distance of around 12 miles [3]. Chrysler announced a recall for 1.4 million vehicles after a pair of hackers demonstrated that they could remotely hijack a Jeep's digital systems over the Internet [4]. These are just a few examples of how attackers can exploit weaknesses in the design of safety-critical CPSs and use these weaknesses to conduct their attacks. In short, a CPS cannot be safe unless it is secured.

This thesis aims at proposing an approach that can identify potential cyber-security attack(s) that a specific safety-critical functionality of an automotive system might be subject to, and analyze how each identified attack might be performed (e.g., attack method/means, attacker's capabilities), and the potential consequences in case such attack success. Then, identify countermeasures to prevent or at least mitigate/minimize the consequences of the attack.

Note: application domain can be the automotive domain, or any other safety-critical CPS domain such as Industrial Internet of Things (IIoT), Smart Cities, etc.

References

[1] M. Dibaei, X. Zheng, K. Jiang, R. Abbas, S. Liu, Y. Zhang, Y. Xiang, and S. Yu, "Attacks and defences on intelligent connected vehicles: a survey," Digital Communications and Networks, 2020.
[2] A. Greenberg, "Hackers Reveal Nasty New Car Attacks-With Me Behind The Wheel (Video)," p. 1, 2013. https://cutt.ly/4jIQVlX
[3] O. Solon, "Team of hackers take remote control of Tesla Model S from 12 miles away — Technology — The Guardian," 2016. https://cutt.ly/hjIQZ7P
[4] A. Greenberg, "The Jeep Hackers Are Back to Prove Car Hacking Can Get Much Worse," 2016. https://www.wired.com/2016/08/jeep-hackers-return-high-speed-steering-acceleration-hacks/