ESReDA
European Safety, Reliability & Data Association

Newsletter

www.esreda.org **Contribute | Influence | Change** December 2022 III Issue

ESReDa reflexions: *Manipulating AI but not the Data: New Attacks on Machine Learning can Even Respect Data Integrity*

Stefan Rass
Full Professor, LIT Secure and Correct Systems Lab, Johannes Kepler University Linz, Austria

Machine learning has demonstrated incredible abilities on classification tasks and has significant potential to relieve humans from tedious and repetitive cognitive tasks in many domains. In many cases, reliability of the algorithm is a matter of robustness against manipulated inputs, such as noise, to make AI behave very differently from its designated purpose.

Recently, potential attacks on AI that respect the AI's data integrity, without touching the input data at any time, were identified. We discuss the threat for supervised and unsupervised learning, concluding that "robustness", in the sense of integrity-protection, should extend to configurations of algorithms, in addition to the algorithms themselves.

The first approach for a possible attack was discovered upon the question of whether sensitive training data inside a machine learning model, e.g., a deep net, is confidential, or if it can actually leak out. Imagine that an adversary, whom we shall call Bob, has stolen data from Alice and has trained it into a machine learning model X. Bob, for example, may seek to predict Alice's next travels, future purchases, likely diseases, or similar. Alice has somehow found out about this and accuses Bob of a misuse of her personal information. A recent theorem [1] puts Bob into the position to deny any culpability, based on the following knowledge: under certain conditions, the machine learning model, one can take a trained ML model X, can produce random data Y that looks "plausible" in the current context (e.g., it looks like having come from Alice). Equipped with X and Y, Bob can construct a topological metric, with which he can demonstrate that the model X was trained exactly from the data Y. Hence, Bob can claim to not have used any of Alice's data, since his model arose from completely unrelated information Y (which he simply created, but which Alice cannot prove). This is only possible because Bob can violate the assumption that ML training is done using "default settings", including the training metric in particular. Unless Bob is verifiably committed to using a particular metric, he will be able to deny the use of any particular, especially sensitive, data. In some cases, such a denial may be desirable, however: imagine that Bob runs a company to process data on Alice's behalf and Alice is concerned about her data being well protected. If a hacker steals Bob's ML model and thereby claims to (implicitly) possess Alice's data, Bob can plausibly deny such claims by demonstrating that the stolen model came out of some data Y which has nothing to do with Alice. This demonstration will not reveal any of Alice's data, since Y can come from a randomness generator.

A second approach for a possible attack has been found on clustering algorithms. Whenever AI is intended to "objectivize" decisions based on properties of people, recent work [2] demonstrates that the topological metrics can be changed towards producing any desired result. For example, imagine that three people, Alice, Bob and Charlie, ask for a loan. All three fill in their application forms. Alice and Charlie give similar answers, while Bob answers quite differently. Claiming to treat all people equally, the institute may feed all data into its clustering AI to put Alice, Bob and Charlie into bins of "loan accepted" or "loan declined". Since Alice and Charlie have filed similar applications, strongly different from Bob's details, an equal-treatment principle of all people would let us expect Bob to be treated differently to Alice or Charlie. However, recent research has demonstrated how to change the mathematical metric that defines "similarity" for the algorithm accordingly, to accomplish any desired classification of Alice, Bob, and Charlie to receive or deny their loans. Like with supervised learning, the manipulation hinges on a secret change of the (topological) metric in the interest of the adversary.

Eliminating exposure to these attacks is actually easy and does not even require using new or different algorithms: the typically large size of ML models, together with the "de facto standards" of how ML algorithms are trained, already mitigate the attacks described. The only requirement is transparency and a publicly verifiable commitment of an AI provider to all details of its algorithms and their configurations. This is a standard application of cryptographic commitments; a well-known primitive besides encryption and signatures. This is indeed not in conflict with business secret protection, since commitments do not reveal secret information, but prevent posterior changes thereof, exactly as needed to mitigate the described attacks. Said differently, this is Kerckhoffs' principle, translated to AI and thereby slightly strengthened: "keep only the training data secret, but be publicly open and committed on all other details of the algorithms”, for otherwise, we get “obscure AI”, which is manipulable AI.

[1] S. Rass, S. König, J. Wachter, M. Egger, und M. Hobisch, „Supervised Machine Learning with Plausible Deniability“, *Computers & Security*, Bd. 112, S. 102506, 2022, doi: https://doi.org/10.1016/j.cose.2021.102506.

[2] S. Rass, S. König, S. Ahmad, und M. Goman, „Metricizing the Euclidean Space towards Desired Distance Relations in Point Clouds“. arXiv, 7. November 2022. doi: 10.48550/arXiv.2211.03674.

Resilience and the birth of a new engineering design approach: regenerative engineering

John Stoop
KINDUNOS Safety Consultancy Ltd
NL

Resilience engineering is a promising new concept in safety thinking that has gained support particularly in socio-organizational network domains, such as health care. However, application of resilience engineering is not widespread in socio-technical systems, where hesitations are ventilated about its broader acceptance and application, such as in transport systems (Zimmerman 2011, ESReDA 2020). Introducing two additional essentials of resilience -initiative and reciprocity- aims at a practical guidance to maintain operational performance at the operational boundaries (Woods 2019). This contribution explores the origins of such hesitations, posing that these two essentials are necessary but not sufficient and leave room for an additional essential, coining the notion of ‘regenerative engineering’.

Based on the mission of the chair of forensic engineering and safety investigations, an analysis was performed of system architecture, technological concepts, functional configuration, and operational context of legacy systems with a high level of technology and complexity. Particularly in the aviation, shipping and railways domain, the issue is addressed how such legacy systems are capable to adapt to changes in optimizing their performance envelope, incorporating short term internal policy & operational constraints and long-term societal values and performance requirements into their system transitions. In the transport industry, the fishing industry, such change capabilities have already manifested themselves respectively as ‘retrofitting’ and ‘regenerating’.

In analysing several system performances, the concept of ‘graceful extensibility’ (Woods 2019), has demonstrated its limitations in gradually expanding its operating envelope. This is particularly the case when integrating externally driven and higher-order changes in performance requirements, requiring transitions beyond their designed system state space boundaries and the predominant solution concepts. The adaptive capability of these systems can be either exhausted due to their complexity, maturity and legacy, or limited by their fundamental design assumptions, modelling restrictions and knowledge deficiencies. As new requirements are added with respect to sustainability and circular economy (e.g. Horizon 2050 and Green Deal demands), new technological challenges arise in the domains of IT, human-machine teaming and mechanical (composite) structures. This plethora of new requirements will for sure drive further derivation of system designs with increased ‘graceful extensibility’, but this iterative design philosophy will reach its limits. Due to the lack of stop rules, resilience engineered systems will eventually be stretched beyond their design space boundaries, becoming ‘systems with a promising past’. Consequently, a distinction should be made between resilient, derivative design evolutions, and a more regenerative, disruptive form of adaptation.

Taking inspiration from Vincent’s theorem of presumptive anomalies (Vincent, 1990):

“Assumptions derived from science may indicate that under future conditions conventional systems may fail or that a radically different system will do a much better job.”

The need for such a radical transition of a system into a different architecture, technology or operational concept should be recognized, identified, and timely analysed in order to prevent an unnoticed performance anomaly that may result in catastrophic failure and loss of values. However, the timely identification of this limit is a challenge in and of itself.

In order to successfully transit to system configurations that can comply with new performance requirements and operating conditions, a selection of sophisticated engineering design methods has been evaluated such as collaborative engineering, knowledge-based design, value chain transitions and cyclic innovation modelling. From a resilience engineering perspective, a key element of this evaluation is distinguishing the need for either a derivative or a disruptive transition. It questions whether the resilience engineering notion can be applied at such a bifurcation point, or resilience engineering should be supplemented with a new, auxiliary notion. Our inquiry into the history of such anomalies in various socio-technical systems indicate the need for critical self-reflection in such evaluations throughout the areas of design, certification, and operations.

Engineering design methodology can be described and evaluated to identify and effectuate constraints and transition barriers discriminating between the suitability of either derivative or disruptive design solution strategies in the socio-technical design space.

In this method, the role of the systems architect is clarified to oversee and strategically guide system design at a much higher level of flexibility and adaptivity. The method also elaborates on necessary adaptations to the fundamental steps in the engineering design process, facilitating an explicit design of system dynamics in both the short- and long-term and explores necessary transitions to incorporate new operational demands, policy constraints, new technologies and transitions in societal values.

Such a method enables foresight on future proof system adaptations, combining efforts to enhance derivative evolutions with methods for disruptive system transitions. Instead of pushing the boundaries of resilience under ETTO considerations, the necessity to introduce a transition in engineering design methodology itself is disclosed. This essential perspective enables the identification of bifurcation points, at which the resilient behaviour capacity of a system is exhausted and radically new systems become inevitably superior. A new engineering discipline provides the capability to cross boundaries in the world of socio-technical system transitions. Such a disruptive revitalising of legacy systems under futureproof conditions is introduced as regenerative engineering.

A few words from the President of ESReDA

ESReDA President
Mohamed Eid
EID Consultant / RiskLyse, France

My few words today are coloured by my deep concerns regarding the global situation of Man's society. The year 2022 has started by activating new geopolitical concerns in addition to the existing already running ones: pandemic, global water and energy inaccessibility, global warming, and energy transitions.

Will Man's society be resilient enough to absorb, dissipate, and shortly recover from all these concerns?

Can engineering sciences and technology contribute effectively into the enhancement of the resilience of Man's society and the accessibility of all mankind to welfare and peace.

I strongly believe in the capacity of sciences and technology to find out the adequate answers to all kind of threats including nature-induced and man-induced threats.

ESReDA, like all the other micro-organisations in the sciences and technology universe, will continue to promote its global scientific and technical collaborative actions through the coming year 2023 and beyond, as it has always done.

It is a crucial moment when Man may doubt of the effectiveness of scientific and technological collaborative efforts to face global threats. Scientific and technological communities are placed in the most advanced lines of defence of Man's society. They know that global threats require global answers, as the Man's today-society is only one global society. And global answers can't be worked out without knowledge, sciences, and technology.

For ESReDA the new year 2023 will bring more challenging collaborative projects within a global scientific and technological thinking strategy.

Forthcoming ESReDA SEMINARS

The 62nd ESReDA Seminar

Alberto Martinetti
University of Twente, the Netherlands

The 62nd ESReDA Seminar on Managing the unexpected: designing systems to embrace disorder for increasing asset reliability

April 12th – 13rd, 2023, University of Twente, the Netherlands [62nd ESReDA Seminar]

Dealing with complex systems has certain characteristics that require consideration to be managed successfully. Understanding and dealing with unexpected events and the unknown are major challenge in asset management.

Unexpected drifts from normal working conditions pose several concerns about the decrease in safety levels as well. Despite the enormous changes and developments in the industry in the last decades as 'an unprecedented fusion between and across digital, physical and biological technologies', approaches for guaranteeing comparable safety and reliability improvement do not evolve quickly enough to offer adequate solutions in managing the mentioned complexity.

Complex assets require a different approach to dealing with unpredictable events and disorder. Consequently, it appears necessary, during the design phase of a complex system, to use tools and techniques for both withstanding stress and becoming stronger but without the necessity of predicting every circumstance. Reliability professionals are in need for ‘antifragile’ methods for embracing disruptive situations and unknowns.

The aim of the seminar is thus to discuss the state of the art and ongoing developments in dealing with unexpected events for complex systems (i.e. infrastructures, energy production), presenting new techniques and methodologies and discussing their strength, weakness, and uncertainties in order to improve reliability.

Topics
Unexpected events
Reliability
Resilience Engineering
Antifragility Engineering
Resilience of infrastructures and equipment
Emergency and crises management models & tools

Domains (among others):
Power generation & supply
Process industry
Gas & Oil production, storage & transport
ICT networks, data storage & servers
Medical & health care
Transport: rail, road, air and maritime
Supply chain process
Water supply and water works

Registration form and the practical information package will be made available soon on the ESReDA website. **62nd ESReDA Seminar**

Past ESReDA SEMINARS

The 61st ESReDA Seminar

Micaela Demichela
Politecnico di Torino, Italy

The 61st ESReDA Seminar on Advances in Modelling to Improve Network Resilience

22-23 September 2022, Torino, Italy. **61st ESReDA Seminar**

The 61st ESReDA seminar was a two-day event that took place on September 22-23, 2022 at the Lingotto Building of Politecnico di Torino in Italy and on-line. The seminar focused on the topic of technological disruptions triggered by natural events, also known as NaTech events. Thirteen speakers shared their experience in several domains: energy, process, critical infrastructures, with a glance to a larger domain, the territory. Three keynote lecture were presented: Marcelo Masera, former Head of Unit "Energy Security, Distribution and Markets", Joint Research Center, The Netherlands discussed about the “Resilience as capacity: a proposed approach”; Prof. Valerio Cozzani, Università di Bologna, about the ”Advances in the management of NaTech events”; Prof. George Boustras, University of Cyprus, showed the “Impact of Natech's to SE Med Critical Infrastructure”. The presentation of the other participants integrated the Keynote lectures in a multidisciplinary and enriching environment, giving the opportunity for extensive and open discussions. A thanks to the PhDs student of the SAfeR research group of DISAT, Politecnico di Torino, that supported the organization of the Seminar.

The 60th ESReDA Seminar

Rasa Remenyte-Prescott
Kate Sanderson
John Andrews
Univ. of Nottingham, UK

Christophe Berenguer,
Sylvie Perrier,
Jean-Marc Tacnet
Julien Baroth
Univ.Grenoble Alpes, FR

The 60th ESReDA Seminar on Advances in Modelling to Improve Network Resilience, 4-5/5/2022, Grenoble, France

The seminar has been organized by the University Grenoble Alpes under the Risk@UGA Idex project framework and hosted by Grenoble INP ENSE3. It has been a forum for exploring issues related to engineering resilience against different threats, such as failures of aging infrastructure, natural disasters and climate change, intentional attacks (cyber-security and terrorism), and emerging threats, met by different industries, critical infrastructures and urban settlements. This seminar closed a 3 years project group "Resilience Engineering and Modelling of Networked Infrastructure", managed by the University of Nottingham, particularly J. Andrews and R. Remenyte-Prescott (in the center of the group picture). Contributions have covered a wide range of topics concerning several stakeholders, from practitioners to researchers (industrialists, regulators, safety boards, universities, R&D organisations, engineering contractors and consultants, training specialists) who presented their work in sessions about resilience of Electrical Networks, transport networks and Smart Cities, Infrastructure Networks… Theories, concepts, and experiences of methods for improved network resilience have been discussed. Authors have been invited to present their research and experience and discuss challenges in enhancing resilience through modelling. Papers have been published soon in JRC Technical Notes. The programme and presentations are available on the ESReDA website.

The proceedings can be downloaded here.

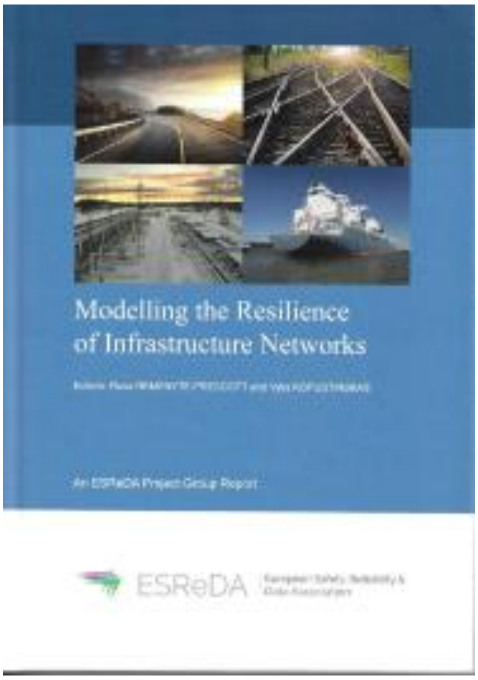
Latest News from the Project Groups

Rasa Remenyte-Prescott
University of Nottingham, UK

John Andrews
University of Nottingham, UK

Project group on Resilience Engineering and Modelling of Networked Infrastructure

Joint Project group Leaders:
– Dr Rasa Remenyte-Prescott, University of Nottingham,
– Professor John Andrews, University of Nottingham.
Project group Secretary – Kate Sanderson, University of Nottingham.


Modelling the Resilience of Infrastructure Networks
An ESReDA Project Group Report
ESReDA

Findings from the project group have been published in a book entitled "Modelling the Resilience of Infrastructure Networks", edited by Rasa Remenyte-Prescott and Vytis Kopustinskas.

This book is a selection of contributions written by members of the Project Group and concentrates on the themes of transportation and utilities. The papers intend to provide an insight into the state of the art of resilience modelling with a focus on Networked systems. The book is aimed at both an industrial and academic readership with interests in the resilience of engineering systems.

We would like to thank the authors for their contributions to this publication, and our colleagues at DNV for their practical support with printing and distribution.

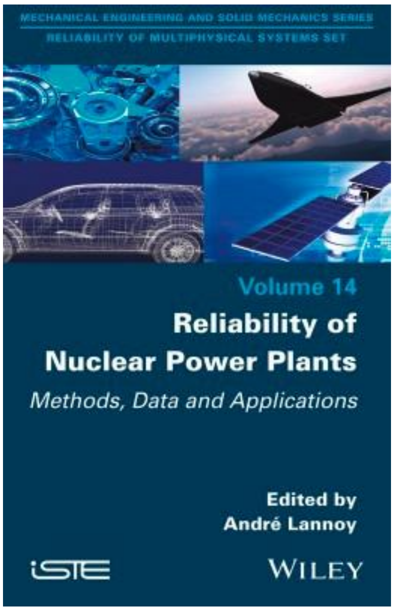
For information on how to purchase a copy please contact ajguillen@us.es ESReDA General Secretary, Antonio J. Guillén (Ingeman, Spain).

ESReDA community recommended books

ESReDA
Honoray President
Jean-François Raffoux

Reliability of Nuclear Power Plants Methods, Data and Applications

Abdelkhalak El Hami


MECHANICAL ENGINEERING AND SOLID MECHANICS SERIES
RELIABILITY OF MULTIPHYSICAL SYSTEMS SET
Volume 14
Reliability of Nuclear Power Plants
Methods, Data and Applications
Edited by André Lannoy
iSTE
WILEY

Since the 1970s, the field of industrial reliability has evolved significantly, in part due to the design and early operation of the first generation nuclear power plants. Indeed, the needs of this sector have led to the development of specific and innovative reliability methods, which have since been taken up and adapted by other industrial sectors, leading to the development of the management of uncertainties and Health and Usage Monitoring Systems.

In this industry, reliability assessment approaches have matured. There are now methods, data and tools available that can be used with confidence for many industrial applications. The purpose of this book is to present and illustrate them with real study cases. The book addresses the evolution of reliability methods, experience feedback and expertise (as data is essential for estimating reliability), the reliability of socio-technical systems and probabilistic safety assessments, the structural reliability and probabilistic models in mechanics, the reliability of equipment and the impact of maintenance on their behavior, human and organizational factors and the impact of big data on reliability. Finally, some R&D perspectives that can be developed in the future are presented.

Written by several engineers, statisticians and human and organizational factors specialists in the nuclear sector, this book is intended for all those who are faced with a reliability assessment of their installations or equipment: decision-makers, engineers, designers, operation or maintenance engineers, project managers, human and organizational factors specialists, experts and regulatory authority inspectors, teachers, researchers, and doctoral students.

The book can be ordered here.

Forthcoming Conferences & Seminars

Rasa Remenyte-Prescott
University of Nottingham, UK

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

UNIVERSITÉ DE LORRAINE

12th IMA International Conference on Modelling in Industrial Maintenance and Reliability (MIMAR) 4-6 July 2023, Nottingham, UK

The 12th International Conference on Modelling in Industrial Maintenance and Reliability (MIMAR) will take place in Nottingham, UK from 4-6 July 2023. This event is the premier maintenance and reliability modelling conference in the UK and builds upon a very successful series of previous conferences. It is an excellent international forum for disseminating information on the state-of-the-art research, theories and practices in maintenance and reliability modelling and offers a platform for connecting researchers and practitioners from around the world.

The scope of the conference includes:

- Engineering Economy and Cost Analysis
- Life cycle/performance analysis
- Maintenance and Reliability Modelling
- Prognostics and Health Management
- Reliability and Maintenance Engineering
- Safety, Security and Risk Management
- Spare Parts Supply Chain Management
- Warranty Management and Data Analysis
- Machine learning for reliability engineering and maint. optimization

Presentations are encouraged on the theory or application of maintenance and reliability for:

- Autonomous Systems
- Cyber-physical systems
- Data Mining and Machine Learning
- Decision Analysis and Methods
- Expert Elicitation
- Operational Research
- Production Planning and Control
- Quality Control and Management

- Human Factors
- Information Processing and Engineering
- Manufacturing Systems
- Resilience Engineering
- Sustainability
- Smart Technologies
- Systems Modelling and Simulation

Publication

Conference Proceedings: authors are invited to submit their paper for publication in the proceedings. All submissions will be peer reviewed and accepted papers will appear in the conference proceedings. The conference proceedings will be indexed by DOI system. Special Issue: selected authors will be invited to submit an extended version of their papers to a Special Issue in Reliability Engineering and System Safety, guest edited by the conference chairs. For more information, please visit 12th IMA International Conference on Modelling in Industrial Maintenance and Reliability (MIMAR) - IMA

17th Summer Safety & Reliability Seminar – SSARS 2023, 9th – 14th July 2023, Kraków, Poland

The annual Summer Safety and Reliability Seminars are organised to advance the methods for the safety and reliability analysis of complex systems and processes and to disseminate the newest achievements in the field. The subjects of the Seminars, different from year to year, are chosen by the Seminars Board in an effort to dynamically represent the methodological advancements developed to meet the newly arising challenges in the field of safety and reliability. This year the emphasis is addressed but is not limited to the following subjects: Safety and Security Management, Safety and Reliability of Complex Systems and Processes, Risk Assessment, Reduction and Accident Consequences Mitigation of Process Industry and Transport Critical Infrastructures, Cybersecurity, Warning Systems, Food Safety, Product Safety, Safety and Resilience Training. More details are available here.

ESReDA Members, you are kindly invited to contribute to the ESReDA newsletter sharing news, announcement of events, your experiences, ideas, etc. You are supposed to elaborate proposals to create new Project Groups, host ESReDA Seminars or initiate collaborative activities.

ESReDA: European Safety, Reliability & Data Association
Association internationale sans but lucratif, régis par la loi Belge du 27 Juin 1921-Titre III (Registration N°: 0452522618 - Siret:E00005802)
Headquarter: ESReDA, rue Gachard 88 Bte 14, B-1050 Bruxelles, Belgium

Contact ESReDA I info@esreda.org

7