

Resilience and Antifragility of Autonomous Systems

Simon Burton^{*1}, Radu Calinescu^{*2}, and Raffaella Mirandola^{*3}

1 University of York, GB. simon.burton@york.ac.uk

2 University of York, GB. radu.calinescu@york.ac.uk

3 Karlsruhe Institute of Technology, KIT, DE. raffaella.mirandola@kit.edu

Abstract

In healthcare, transportation, manufacturing, and many other domains, autonomous systems have the potential to undertake or support complex missions that are dangerous, difficult, or tedious for humans. However, to achieve this potential, autonomous systems must be *resilient*: they must continue to provide the required functionality despite the anticipated and unforeseen disturbances encountered within their operating environments. This ability to achieve user goals in open-world environments can be further increased by making autonomous systems *antifragile*. Antifragile systems benefit from exposure to uncertainty and disturbances, by learning from encounters with such difficulties, so that they can handle their future occurrences faster, more efficiently, with lower user impact, etc. This Dagstuhl Seminar brought together leading researchers and practitioners with expertise in autonomous system resilience, antifragility, safety and ethics, self-adaptive systems, and formal methods, with the aim to: (1) develop and document a common understanding of resilient and antifragile autonomous systems (RAAS); (2) identify open challenges for RAAS; (3) discuss promising preliminary approaches; and (4) propose a research agenda for addressing these challenges.

Seminar April 28 – May 3, 2024 – <https://www.dagstuhl.de/24182>

2012 ACM Subject Classification General and reference → Reliability; General and reference → Metrics; General and reference → Validation; Computer systems organization → Embedded and cyber-physical systems; Computer systems organization → Dependable and fault-tolerant systems and networks; Software and its engineering; Theory of computation → Logic; Mathematics of computing → Probability and statistics; Computing methodologies → Artificial intelligence; Computing methodologies → Machine learning; Human-centered computing

Keywords and phrases artificial intelligence, antifragility, autonomous systems, disturbance, ethics, formal methods, machine learning, nondeterminism, resilience, safety, self-adaptive systems, validation and verification, uncertainty

Digital Object Identifier 10.4230/DagRep.14.4.142

1 Executive Summary

Simon Burton

Radu Calinescu

Raffaella Mirandola

License  Creative Commons BY 4.0 International license
© Simon Burton, Radu Calinescu, and Raffaella Mirandola

The increasing complexity in the environment, tasks, and technology related to autonomous systems results in limitations in the statements that can be made regarding dependability during design time. In particular, these systems may operate within environments for which only incomplete models exist, that may change over time or may be subject to unforeseen

* Editor / Organizer



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Resilience and Antifragility of Autonomous Systems, *Dagstuhl Reports*, Vol. 14, Issue 4, pp. 142–163

Editors: Simon Burton, Radu Calinescu, and Raffaella Mirandola



DAGSTUHL REPORTS Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

interactions and disturbances. As a result, such systems must be engineered to be trustworthy despite residual insufficiencies in their design, and in the presence of unexpected events due to their dynamically evolving operating context.

Related domains concerned with system autonomy in uncertain environments have already taken inspiration from nature to endow artificial systems with self-* properties (e.g. self-optimisation, -repair, -protection, -configuration, and -adaptation). Such self-* capabilities enable systems to improve their performance and dependability at runtime while reducing the need for low-level human intervention – properties that are closely related to resilience and antifragility.

This Dagstuhl Seminar aimed to unify the international research on **resilient and antifragile autonomous systems** (RAAS), leading to faster scientific advancements and industrial adoption. To this end, the seminar brought together leading researchers and practitioners with expertise in autonomous system resilience, antifragility, safety, and ethics, from disciplines including computer science, safety science, and ethics, to share and discuss each other's understanding of, methods for, and open challenges related to RAAS. Initial presentations were used to set the scene by proposing basic definitions, industry perspectives, and engineering views on cyber-resilience. These were followed by group and plenary discussions to explore these concepts in more detail.

A clear set of agreed definitions is essential in order to make progress as a community in this area. *Resilience* can be broadly seen as the ability to absorb disturbances and unexpected events whilst maintaining essential properties of the system. Using such conditions to harden the system against future events can be viewed as *antifragility*. These definitions highlight that antifragility is a concept referring to systems designed to operate under “open-world” assumptions, where the responsibility of maintaining a given property, despite disturbances (resilience) mostly shifts from design time to runtime, and relies on the presence in the system of some suitable degree of autonomy (self-* capability). As such, antifragility can be viewed as the ability of a system to *self-improve* its resilience over (run)time. Discussions converged to the idea that in order to define resilience and antifragility, we should build on the work of Control Theory, specifically how systems recover from (potentially previously unknown) disturbances. Thus, we postulated that both resilience and antifragility should be defined over the metrics of settling time, percentage of settling, percentage of overshoot, and percentage of overshoot with respect to the properties of interest in the event of disturbances to the system. Discussions on how to use formal methods to construct systems that guarantee these desired properties generated many challenging questions that are to be followed up in future research.

Initial work in the seminar explored more precise definitions of RAAS that also included the consideration of uncertainty and causality, and where a collection of properties may need to be optimised as a whole. Such trade-offs are particularly evident when considering safety, ethical, and legal aspects of RAAS. In some cases, autonomous systems must remain operational in order to stay safe. A resilient system could remain within its safety bounds when disrupted, whilst maintaining a minimal level of utility. An antifragile system could use repeated disturbances to lower risk over time whilst increasing overall utility. Similar trade-offs and optimisations will be found when considering legal and ethical concerns for RAAS and these could lead to specific technical requirements on the system. For example, for a system that adapts its function over time, avoiding the loss of agency in human stakeholders needs to be ensured.

Engineering antifragile systems requires specialised consideration in each phase of the traditional software and system development process. This includes requirements, design, implementation, and testing. Artificial Intelligence (AI) – in terms of machine learning,

symbolic AI techniques, and combinations thereof – has the potential to provide a basis for both recognising disturbances and deciding the system adaptations needed to mitigate these disturbances. The seminar participants see potential for AI to be used in all phases of the MAPE-K (monitor-analyse-plan-execute supported by knowledge) cycle of self-adaptive systems. Furthermore, a control-theoretic reasoning approach could be used to verify whether a particular adaptation manager pushes the resilience error (i.e., the difference between observed and preferred resilience) below some threshold, or whether the resilience level stabilises at a reference value.

The seminar concluded that much work is still required to advance research in the area of RAAS, and to foster RAAS adoption in industrial applications. This includes:

- Agreeing on terminology and definitions that build upon and extend our traditional understanding of dependable systems;
- Formally defining metrics for resilience and antifragility that can be used to design and verify RAAS;
- Engineering methods and candidate technologies for implementing RAAS;
- Considering the safety, legal, and ethical implications of RAAS, including both their positive potential and their associated risks.

The participants agreed to pursue these important and challenging issues in future collaborations, including joint publications, workshops, and journal special issues.

2 Table of Contents

Executive Summary

Simon Burton, Radu Calinescu, and Raffaella Mirandola 142

Overview of Talks

Characterizing Antifragile ICT Systems: Conceptual and Architectural Models
Vincenzo Grassi and Diego Perez-Palacin 147

Emulation, Standards, and Ethics for Resilient and Antifragile Autonomous Systems
Lee Barford 148

Towards Operational Cyber Resilience
Kerstin I. Eder 149

Working Groups

Concepts, terminology & definitions for resilience and antifragility 1
Vincenzo Grassi, Ada Diaconescu, Felicita Di Giandomenico, Gabriel Moreno, Elena Navarro, and Sebastián Uchitel 150

Concepts, terminology & definitions for resilience and antifragility 2
Ralf H. Reussner, Amel Bennaceur, Mario Gleirscher, Antje Loyal, Raffaella Mirandola, Diego Perez-Palacin, and Patrizia Scandurra 151

Safety concerns for resilient and antifragile autonomous systems
Kerstin I. Eder, Simon Burton, Marc Carwehl, Andreas Heyl, Ravi Mangal, Shiva Nejati, and Grisel Vázquez 152

Legal and ethical concerns for resilient and antifragile autonomous systems
Ana Cavalcanti, Lee Barford, Radu Calinescu, Matteo Camilli, Sebastian Hahner, Lina Marsso, and Catia Trubiani 153

Formal methods for autonomous system resilience and antifragility
Sebastián Uchitel, Radu Calinescu, Ana Cavalcanti, Mario Gleirscher, Lina Marsso, Catia Trubiani, and Grisel Vázquez 154

Nature-inspired methods for autonomous system resilience and antifragility
Ada Diaconescu, Sebastian Hahner, Raffaella Mirandola, Gabriel Moreno, Elena Navarro, Ralf H. Reussner, and Patrizia Scandurra 156

AI solutions for autonomous system resilience and antifragility
Andreas Heyl, Simon Burton, Felicita Di Giandomenico, Vincenzo Grassi, Ravi Mangal, and Shiva Nejati 157

Engineering resilient and antifragile autonomous systems
Amel Bennaceur, Lee Barford, Matteo Camilli, Marc Carwehl, Kerstin I. Eder, and Diego Perez-Palacin 158

Birds-of-a-Feather Groups

Formalising the relation of uncertainty, knowledge, decisions and antifragility
Ralf H. Reussner, Simon Burton, Felicita Di Giandomenico, Kerstin I. Eder, Vincenzo Grassi, Ravi Mangal, Raffaella Mirandola, and Patrizia Scandurra 159

Resilience and antifragility of ethics-aware Human-AI collaborations <i>Amel Bennaceur, Lee Barford, Ana Cavalcanti, Radu Calinescu, Antje Loyal, Lina Marsso, Elena Navarro, Shiva Nejati, and Catia Trubiani</i>	159
Reasoning about antifragility from a control perspective <i>Mario Gleirscher, Marc Carwehl, Ada Diaconescu, Sebastián Uchitel, and Gricel Vázquez</i>	160
Uncertainty propagation to support achieving antifragility <i>Sebastian Hahner, Matteo Camilli, Andreas Heyl, Antje Loyal, Gabriel Moreno, and Diego Perez-Palacin</i>	161
Participants	163

3 Overview of Talks

3.1 Characterizing Antifragile ICT Systems: Conceptual and Architectural Models

Vincenzo Grassi (University of Rome “Tor Vergata” – Rome, IT, vincenzo.grassi@uniroma2.it) and Diego Perez-Palacin (Linnaeus University – Växjö, SE, diego.perez@lnu.se)

License © Creative Commons BY 4.0 International license

© Vincenzo Grassi and Diego Perez-Palacin

Joint work of Vincenzo Grassi, Raffaella Mirandola, Diego Perez-Palacin

Main reference Vincenzo Grassi, Raffaella Mirandola, Diego Perez-Palacin: “A conceptual and architectural characterization of antifragile systems”, *J. Syst. Softw.*, Vol. 213, p. 112051, 2024.

URL <https://doi.org/10.1016/J.JSS.2024.112051>

Antifragility is one of the terms that have recently emerged with the aim of indicating a direction that should be pursued toward the objective of designing ICT systems that remain trustworthy despite their dynamic and evolving operating context. We present a characterization of antifragility, aiming to clarify from a conceptual viewpoint the implications of its adoption as a design guideline and its relationships with other approaches sharing a similar objective. To this end, we discuss the inclusion of antifragility (and related concepts) within the well-known dependability taxonomy presented in [1], which was proposed a few decades ago with the goal of providing a reference framework to reason about the different facets of the general concern of designing trustworthy systems able to cope with changes.

Indeed, we believe that a primary need for a software engineer involved in the design of such systems is to have a commonly agreed-on repertoire of terms and underlying concepts, which makes clear which system aspects each term intends to capture, whether some term is a specialization (qualifications) of some other, or if it denotes a means for attaining a property indicated by another term. In this perspective, our position is that the crisp conceptual reference provided by the dependability taxonomy should not be lost or obfuscated but rather updated and expanded, if necessary. The extension we discuss allows us to integrate the *antifragility* term and the underlying concepts into that taxonomy, thus maintaining its role of a unified place where the relationships among different goals and approaches aimed at designing and building ICT systems able to cope with changes can be better understood and compared.

Then, based on this conceptual clarification, we also discuss how to promote the engineering of antifragile systems. To this end, we first present a reference model for antifragile ICT systems inspired by the three-layer reference model for self-managing systems proposed in [2], and then we delineate a path based on the Digital Twin technology for the realization of antifragile systems.


A thorough presentation of the issues discussed in this talk can be found in [3].

References

- 1 A. Avizienis, J.-C. Laprie, B. Randell and C. Landwehr. Basic concepts and taxonomy of dependable and secure computing, in *IEEE Transactions on Dependable and Secure Computing*, vol. 1, no. 1, pp. 11-33, Jan.-March 2004.
- 2 J. Kramer and J. Magee, "Self-Managed Systems: an Architectural Challenge," *Future of Software Engineering (FOSE '07)*, pp. 259-268, 2007.
- 3 V. Grassi, R. Mirandola and D. Perez-Palacin. A conceptual and architectural characterization of antifragile systems, in *Journal of Systems and Software*, Vol. 213, Article N° 112051, 2024.

3.2 Emulation, Standards, and Ethics for Resilient and Antifragile Autonomous Systems

Lee Barford (Keysight Technologies – London, GB)

License  Creative Commons BY 4.0 International license
© Lee Barford

This talk covers the development and verification of robust and resilient autonomous systems from an industry perspective, with a focus on the role of emulation, standards, and ethics in the process. Many autonomous systems assess or control the physical world in real-time through sensors, actuators, or antennas. To validate them, emulations of the environments in which such systems operate must be provided, the sensors being fed accurate physical excitations and the simulation behind the emulation being updated by the actuator behaviours. In this manner, scenarios too dangerous or expensive to do in real life can be used to validate or provide high-quality synthetic training data. An example such emulation/validation system for autonomous drive is presented. In the case of robust and resilient autonomous systems, the need for such hardware-in-the-loop emulation is even greater, as then validating robustness requires that the system be run through scenarios too rare to appear in normal volumes of training data.

To realize resilience and antifragility in industrial-produced systems, standards for resilience and antifragility need to be developed and adopted. Such standards should be able to be turned into scenarios for training, fine-tuning, functional tests, and conformance tests for a particular robust and resilient system. Tools for design and model analysis need to be developed to ease achieving standards compliance. Where an autonomous system can monitor itself and upload status information for continuous improvement of the system, key performance indicators of resilience and antifragility that relate back to the standards and system requirements need to be created that are informative but (1) have a low burden on deployed system, and (2) require a low comms bandwidth in normal situations.

Creators of robust and resilient systems would also benefit from values-based design, where systems are designed from the beginning with the anticipation of impacts on human values in mind. This approach benefits investors, managers, engineers, customers, and the public by introducing ethical clarity at the requirements-gathering phase, when changes are easier and cheaper to make than later in the design process. A process of identification of stakeholders and their values is necessary to identify and prioritize socio-ethical risks that then become requirements for resilience and antifragility.

3.3 Towards Operational Cyber Resilience

Kerstin I. Eder (University of Bristol, GB & Trustworthy Systems Laboratory – Bristol, GB, Kerstin.Eder@bristol.ac.uk)

License © Creative Commons BY 4.0 International license

© Kerstin I. Eder

Joint work of Kerstin I. Eder, Carsten Maple, Peter Davies, Chris Hankin, Greg Chance, Gregory Ephiphaniou¹

Main reference Kerstin Eder: “CyRes: towards operational cyber resilience”, in Proc. of the 1st International Workshop on Verification of Autonomous & Robotic Systems, VARS '21, Association for Computing Machinery, 2021.

URL <https://doi.org/10.1145/3459086.3460119>

Existing approaches to cyber security in the automotive sector are not fit to deliver the resilience required for safe mass deployment of advanced driving features and smart mobility services. In this presentation I introduced an innovative multi-directional approach to operational cyber resilience, the CyRES methodology, which aims to enable the delivery of robust and resilient engineering practices in this sector from design, via manufacture to operation. CyRES is based on three principles: increasing the probability of Detection, Understanding and Acting on cyber events; increasing the number of Engineered Significant Differences; and invoking a continuum of Proactive Updates. I motivated, illustrated and explained these principles on examples, focusing mainly on the first two principles. CyRES is an exciting opportunity for engineers and computer scientists to re-target widely studied, mature methods, such as those developed by the self-adaptive systems community, for cyber security. My main objective was to raise awareness of the many intellectual challenges associated with realising these principles, and to highlight some of the ways for attendees to contribute to the realisation of the CyRES vision.

Further details on CyRES and the underlying principles can be found in [1].

References

- 1 K. Eder. CyRes: Towards operational cyber resilience. In Proceedings of the 1st International Workshop on Verification of Autonomous & Robotic Systems (VARS'21). Association for Computing Machinery, New York, NY, USA, Article 11, 1–3.
<https://doi.org/10.1145/3459086.3460119>.

¹ as part of the Cyber Resilience in Connected and Autonomous Mobility project ResiCAV, which was supported by funding from The Centre for Connected and Autonomous Vehicles (CCAV) run by Zenic and Innovate UK, project number 133899.

4 Working Groups

4.1 Concepts, terminology & definitions for resilience and antifragility 1

Vincenzo Grassi (University of Rome “Tor Vergata”, IT)

Ada Diaconescu (Telecom Paris, FR)

Felicita Di Giandomenico (CNR – Pisa, IT)

Gabriel Moreno (Carnegie Mellon University – Pittsburgh, US)

Elena Navarro (University of Castilla-La Mancha, ES)

Sebastián Uchitel (University of Buenos Aires, AR)

License © Creative Commons BY 4.0 International license

© Vincenzo Grassi, Ada Diaconescu, Felicita Di Giandomenico, Gabriel Moreno, Elena Navarro, and Sebastián Uchitel

The group focused its discussion on the following issues:

- arriving at a suitable definition of antifragility;
- clarifying the relationships of this concept with other concepts, e.g., resilience, dependability, self-adaptation, and learning;
- identifying possible “parallel” specializations of the antifragility concept for different domains.

About the first issue, the discussion led us to conclude that it is useful to start with a declarative definition of antifragility that states what we expect from a system to consider it antifragile, remaining neutral with respect to how-to make it antifragile, and with respect to measures of its antifragility degree. To this end, the following definition emerged from the discussions:

Antifragility is the ability of a system to self-improve its resilience over (run)time.

This definition highlights that antifragility is a concept referring to systems designed to operate under “open-world” assumptions, where the responsibility of achieving a given property (resilience) mostly shifts from design time to runtime, and relies on the presence in the system of some suitable degree of autonomy (self-* capability).

This same definition also provides a perspective on dealing with the other two issues considered in the discussion.

For the second issue, it establishes, in particular, a relationship between antifragility and resilience by assigning to antifragility the role of an attribute of the process followed to attain and/or improve resilience. Hence, antifragility denotes a system’s ability to incrementally achieve at runtime higher levels of resilience. Importantly, antifragility does *not* imply that the system *is* resilient, only that it is able to *improve* its resilience over time.

Concerning the “learning” concept, the given definition purposely avoids its use to leave space for approaches to antifragility that are not necessarily based on the explicit use of learning methodologies (even if we recognize that they can play a significant role).

For the third issue, it suggests that the specialization of antifragility for different domains can be at least partially deferred to the different characterizations (and measures) of resilience for those domains. This specialization looks at the property to be achieved. Besides this, another possible specialization could concern the process to be followed to that end, which could depend on the considered domain.

The discussion in the group also touched on issues concerning the how-to aspect. Emerged suggestions about approaches that could be pursued to achieve antifragility include:

- MAPE-K;
- Observer Controller;
- Learn (acquire knowledge), Reason, Act;
- Data collection, generalization, action;
- Exploration and exploitation;
- Evolutionary algorithms.

4.2 Concepts, terminology & definitions for resilience and antifragility 2

Ralf H. Reussner (KIT – Karlsruher Institut für Technologie, DE)

Amel Bennaceur (The Open University – Milton Keynes, GB)

Mario Gleirscher (Universität Bremen, DE)

Antje Loyal (Continental Automotive Technologies – Frankfurt, DE)

Raffaella Mirandola (KIT – Karlsruher Institut für Technologie, DE)

Diego Perez-Palacin (Linnaeus University – Växjö, SE)

Patrizia Scandurra (University of Bergamo – Dalmine, IT)

License  Creative Commons BY 4.0 International license

© Ralf H. Reussner, Amel Bennaceur, Mario Gleirscher, Antje Loyal, Raffaella Mirandola, Diego Perez-Palacin, and Patrizia Scandurra

The group discussed formalisations of the meaning of antifragility based on the paper “A conceptual and architectural characterization of antifragile systems” by Vincenzo Grassi, Raffaella Mirandola and Diego Perez-Palacin. In particular, changes of the environment were formalised as an extension of this paper. Main insights (including the following plenary discussion) were the clarification of the difference between resilience and antifragility. While both concepts may deal with unknown unknowns to a certain degree, antifragility is characterised through learning from prior events to improve. This can be seen as a higher-order adaptation mechanism, using prior events to change the adaptation mechanism to achieve a higher level of quality. This implies that the boundaries of subsets “dead” (i.e., catastrophic failure) and “survivable” of the system state space are changed through this higher order adaptation. We identified examples ranging from technical systems (autonomous vehicles, learning in e-scooters) to society (emergency forces learning from previous operations).

4.3 Safety concerns for resilient and antifragile autonomous systems

Kerstin I. Eder (University of Bristol, GB)

Simon Burton (University of York, GB)

Marc Carwehl (Humboldt-Universität zu Berlin, DE)

Andreas Heyl (Robert Bosch GmbH – Stuttgart, DE)

Ravi Mangal (Carnegie Mellon University – Pittsburgh, US)

Shiva Nejati (University of Ottawa, CA)

Gricel Vázquez (University of York, GB)

License © Creative Commons BY 4.0 International license

© Kerstin I. Eder, Simon Burton, Marc Carwehl, Andreas Heyl, Ravi Mangal, Shiva Nejati, and Gricel Vázquez

This session explored safety concerns for resilient and antifragile autonomous systems. We focused on the question “How do resilience and antifragility help with safety?” Our observations covered risk over time, with a specified maximum level of risk (safety boundary) beyond which the system was considered unsafe, as well as utility over time, with a minimum required level of utility (liveness) beyond which the system was considered no longer useful.

The baseline for our discussion was a resilient system that, when disrupted, remains within its safety boundary with respect to the operational risk while utility degrades to zero, rendering the system useless. Depending on the application, zero utility may imply that risk falls to zero along with utility, or that risk is maintained on the original level. The former is exemplified in scenarios where not doing anything is safe, while the latter represents scenarios where not doing anything is not safe.

Enhanced resilience means that the system, when disrupted, remains within its safety boundary and maintains utility at the desired level. Repeated exposures to shocks are absorbed by the system over time, with periods of higher risk being coupled to lower utility, and periods of lower risk being associated with higher utility. Such systems are not designed to improve performance over time, though they recover each time they are exposed to a disturbance.

When an antifragile system is disrupted, it gradually, though not necessarily monotonically lowers the risk and improves utility. Antifragile systems can operate in a variety of different safe subsets of operational states, which can be modified and extended by a controller associated with the system.

A formalisation must capture both safety and utility (liveness) properties of the autonomous system in such a way that these properties constitute a measurable representation of the application-specific requirements for the given system.

Open questions include “How to make resilient and antifragile systems safe?” and “How to maintain safety after changes in systems with resilience and antifragility?”

4.4 Legal and ethical concerns for resilient and antifragile autonomous systems

Ana Cavalcanti (University of York, GB)

Lee Barford (Keysight Technologies – London, GB)

Radu Calinescu (University of York, GB)

Matteo Camilli (Politecnico di Milano, IT)

Sebastian Hahner (KIT – Karlsruher Institut für Technologie, DE)

Lina Marsso (University of Toronto, CA)

Catia Trubiani (Gran Sasso Science Institute – L’Aquila, IT)

License © Creative Commons BY 4.0 International license
 © Ana Cavalcanti, Lee Barford, Radu Calinescu, Matteo Camilli, Sebastian Hahner, Lina Marsso, and Catia Trubiani

We have first considered the aspects of the life-cycle of design and verification of a system that are relevant when considering legal and ethical issues. We have identified the extensive list below, but started our discussions with the question “What is the relationship between legal and ethical concerns and resilience/antifragility?” We have agreed that legal and ethical concerns are present in all systems, but in autonomous systems, requiring adaptation at runtime, these issues cannot be resolved a priori. The loss of agency raises threats. On the other hand, we noted that runtime adaptation, resilience and antifragility also can create opportunities, since an intelligent system can provide additional services.

Our list of concerns goes from requirements all the way to runtime adaptation:

1. Elicitation of normative requirements: Is this even possible? How can we deal with subjectivity and the multi-cultural context? Who should participate in the elicitation?
2. What does it mean for normative requirements to be “suitable”?
3. Generalisation of infrastructure for items 1 and 2.
4. How to bridge the gap between engineers & the social scientists?
5. Synthesis of compliant autonomous system behaviour.
6. Verification of compliance.
7. Formal foundations.
8. EthicsOps (adaptation).

In the interest of time, we focussed on items 2, 6, and 7. Our goal in each case was to define why each of the topics was important, and why it was challenging. In the discussion of requirements, we identified a notion of suitability. We say that a set of requirements is suitable when it has the following characteristics:

- of ethical and legal relevance
- affectable by the system
- relevant to stakeholders
- machine understandable
- conflicts are removed or managed
- free of redundancy
- unambiguous
- sufficient, that is, reduces risk of legal or ethical harm
- not overly conservative, that is, does not unnecessarily restrict the services that can be offered.

The difficulty to achieve suitable sets of requirements arises from the fact that it is difficult even for people to decide what is legal and what is ethical, stakeholders from multidisciplinary backgrounds need to be involved, and there may be competing business and institutional interests.

For verification, we identified the importance over and above the usual arguments. Verification can support quantification of risks, minimisation of harmful impacts on values, design or decision space exploration, and evaluation of impact on engineering process. Difficulties arise from the fact that adaptation leads to scalability issues, as the potential set of behaviours at runtime increases. In addition, resilience and anti-fragility require proper consideration of uncertainty. Testing and conformance in general require oracles, but the potentially subjective nature of the requirements and dealing with continuous quantities makes it hard. Finally, the very specification of the new properties (resilience, antifragility) is a challenge.

Some of these aspects were identified as imposing challenges in particular to the use of mathematical foundations in verification, namely, scalability, uncertainty, time, subjectivity, and complexity of specifications. We have finally taken the time to discuss the state of the art, and concluded by looking forward to additional discussions regarding the topics that we did not cover.

4.5 Formal methods for autonomous system resilience and antifragility

Sebastián Uchitel (University of Buenos Aires, AR)

Radu Calinescu (University of York, GB)

Ana Cavalcanti (University of York, GB)

Mario Gleirscher (Universität Bremen, DE)

Lina Marsso (University of Toronto, CA)

Catia Trubiani (Gran Sasso Science Institute – L'Aquila, IT)

Gricel Vázquez (University of York, GB)

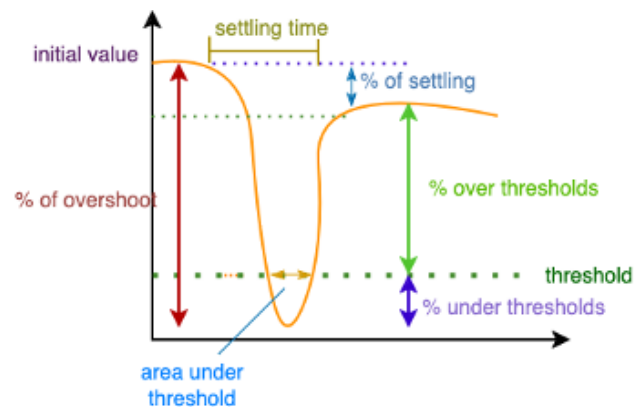
License © Creative Commons BY 4.0 International license

© Sebastián Uchitel, Radu Calinescu, Ana Cavalcanti, Mario Gleirscher, Lina Marsso, Catia Trubiani, and Gricel Vázquez

The group decided to revisit the definitions of resiliency and antifragility that had been discussed informally the previous day, with the aim of providing a more formal perspective and a relation to specific quality attributes such as performance, safety, and ethics. We also decided that we would ground these definitions on a specific system and a concrete quality to make the discussion and definitions concrete. We aimed to first think of how given two systems, we would compare them with respect to resilience and antifragility, and to postpone the discussion of how such systems may be constructed using formal methods until the end of the session.

The running example we discussed was that of a robot monitoring and interacting with patients in an Emergency Department waiting room at a hospital. Amongst the multiple quality attributes that can be considered in such a system, we decided to address only one to start, assuming that some of the ideas that we would elaborate would then be transferred to a multi-dimensional setting. We chose *patients served per hour* as a quality metric.

Discussions converged to the idea that in order to define resilience and anti-fragility, we should build on the work of Control Theory on how systems recover from disturbances. Thus, we postulated that both resilience and antifragility should be defined (Figure 1) over the



■ **Figure 1** Control-theory inspired metrics for resilience.

metrics of settling time, percentage of settling, percentage of overshoot, and percentage of overshoot on the disruption signals. However, given that we are interested in hard worst-case scenarios too, in addition to the classical set-point concept, we introduced an acceptable threshold value and associated metrics: percentage under threshold, area under threshold, and time under threshold.

We discussed domain-specific examples of resilience defined as summations over disturbances measured as linear combinations settling time and percentage, and antifragility as a comparison between the resilience of two consecutive periods, noting that in this way antifragility can be thought of, informally, as the first derivative of resilience. Another point of discussion is that antifragility must be defined by comparing resilience over periods of time that are long enough to capture statistically relevant sets of disruptions.

Discussion on how to use formal methods to construct systems with these desired properties left many important questions unanswered: What would an appropriate methodology be to guarantee such properties at design time, if a system were to include adaptive mechanisms to achieve antifragility? On what formal methods foundations would it rely upon? Can these properties be decomposed and assigned to different system components to allow for independent construction, incremental improvements, and modular reasoning? Would a hierarchical structure of control loops in which a manager controls for resilience and a “manager of the manager” controls for antifragility, be appropriate?

4.6 Nature-inspired methods for autonomous system resilience and antifragility

Ada Diaconescu (Telecom Paris, FR)

Sebastian Hahner (KIT – Karlsruher Institut für Technologie, DE)

Raffaella Mirandola (KIT – Karlsruher Institut für Technologie, DE)

Gabriel Moreno (Carnegie Mellon University – Pittsburgh, US)

Elena Navarro (University of Castilla-La Mancha, ES)

Ralf H. Reussner (KIT – Karlsruher Institut für Technologie, DE)

Patrizia Scandurra (University of Bergamo – Dalmine, IT)

License © Creative Commons BY 4.0 International license

© Ada Diaconescu, Sebastian Hahner, Raffaella Mirandola, Gabriel Moreno, Elena Navarro, Ralf H. Reussner, and Patrizia Scandurra

The discussion focused on the following topics:

- Acknowledging nature-inspired approaches already available in related domains;
- Providing natural examples illustrating the trade-off between performance and resilience/antifragility;
- Identifying specific aspects of antifragility and resilience, and determining how they fit within the more general architectures of self-* systems;
- Setting the basis for evaluating and comparing system antifragility capabilities.

Related domains concerned with system autonomy in uncertain environments have already taken inspiration from nature to endow artificial systems with self-* properties (e.g. self-optimisation, -repair, -protection, -configuration, -adaptation). Such self-* capabilities enable systems to improve their performance and dependability at runtime while reducing the need for low-level human intervention. Relevant domains include autonomic, organic and self-aware computing, self-adaptive and control systems, evolutionary computing, swarm robotics, morphogenetic engineering and artificial life.

Within this context, antifragility is concerned with the particular kind of self-* processes that enable systems to self-improve their resilience over time, by capitalising on past experiences – including, but not limited to, self-adaptation to unexpected, rare events. Hence, antifragility processes pertain to the meta-control layer defined within most multi-level self-* system architectures.

Importantly, resilience and anti-fragility compete with performance optimization concerns. Relevant examples include: species specialised for eco-systemic niches versus more versatile species surviving through fluctuating environments; engineered artifacts that rely on specific, highly-integrated components versus loosely-coupled artifacts supporting diverse assemblies; highly-synchronized railway systems maximizing traffic versus less precise ones that tolerate more delays.

With respect to general self-* processes, a system's antifragility is specific in its ability to draw benefits from past reactions to disturbances, so as to improve its future reactions to more or less similar disturbances.

We provide a formalization basis for the above concept as follows (Note: this extends previous works of seminar members and discussions within other groups). Considering a self-* system that reacts to disturbances (e.g. in its environment) that are within a domain E by changing its state (e.g. via a controller C) within a solution domain S . If the system encounters a disturbance outside E , hence within another domain E' , then the controller C may no longer be able to find a solution within domain S . In an antifragile system, a meta-controller can search through a wider solution domain (meta- S) and find another

controller C' , which can adapt the system through states within another subdomain S' in response to disturbances that include those in E' . The system's antifragility-specific support (that we called a "red dot") includes all mechanisms that define the system's maximum solution search domain (meta- S); and that enable the search process from one solution sub-domain (S) to another (S').

We may evaluate a system's antifragility support ("red dot") depending on how much it extends the system's maximum state-space domain (meta- S); and on how effective and efficient is its search process through this domain is (finding S' within meta- S). Finally, antifragility strategies may vary, ranging from brute-force replication and variation (e.g. insects) and all the way to sophisticated predictive approaches (e.g. humans).

4.7 AI solutions for autonomous system resilience and antifragility

Andreas Heyl (Robert Bosch GmbH – Stuttgart, DE)

Simon Burton (University of York, GB)

Felicita Di Giandomenico (CNR – Pisa, IT)

Vincenzo Grassi (University of Rome "Tor Vergata", IT)

Ravi Mangal (Carnegie Mellon University – Pittsburgh, US)

Shiva Nejati (University of Ottawa, CA)

License © Creative Commons BY 4.0 International license

© Andreas Heyl, Simon Burton, Felicita Di Giandomenico, Vincenzo Grassi, Ravi Mangal, and Shiva Nejati

We decompose the problem of using AI for designing resilient and antifragile autonomous systems (RAAS) into four separate questions, namely, "Why should we use AI for RAAS?", "Where should we use AI in RAAS?", "Which types of AI should be used in RAAS and what should they be used for?", and "Where should we start to use AI in RAAS?"

There is a need to use AI in RAAS because these systems are complex and need to handle various sources of uncertainties. AI solutions can be effective in dealing with uncertainty, and are likely to play a key role in transitioning from fail-safe systems to resilient systems and from resilient systems to antifragile systems.

Second, AI can be used at various stages of the RAAS lifecycle, namely, design-time, run-time, and operation-time. While AI can help with developing the systems and simulating their behaviour at design-time, it can also help with monitoring the system for shocks and helping with recovery at run-time and operation-time. Further, AI can be used in different parts of a RAAS. Assuming the standard managed and managing system architecture for RAAS, AI can be an essential component of the managed system (for instance, to perform perception and/or control in a resilient and antifragile cyber-physical system) and/or be a part of the managing system. Finally, resiliency and antifragility need not just be properties of individual systems but could also be desirable properties for systems of systems and entire ecosystems. For instance, while we want an autonomous car to be resilient and antifragile, we can also require these properties to hold for the entire fleet of cars. AI can operate at various levels of this hierarchy; for instance, at the fleet level, AI could help with analysing the large volumes of performance data being collected by the fleet.

Third, focusing on the use of AI in the managing system, we first assume that managing systems in RAAS will follow the standard MAPE-K structure. Given this architecture, AI can help with each step of the MAPE-K cycle. For instance, machine learning models can be used to monitor the state of the system and detect if the system behaviour is outside the

nominal bounds. Symbolic AI techniques can be used to analyse and extract the relevant knowledge from the knowledge base for the purpose of planning. An important insight is that each component of the MAPE-K can have different requirements and therefore different forms of AI might be suitable for the different components (data-driven AI for monitoring vs symbolic AI for analysis and planning). We also foresee managing systems that integrate humans and AI. An important question to study is how we can mitigate the limitations of AI such as non-robustness, tendency to hallucinate, and unpredictability when we deploy it in the MAPE-K cycle.

Finally, AI needs to be introduced to RAAS in an incremental fashion, ensuring that the introduction of AI itself does not lead to an increased lack of resiliency or fragility of the system. Towards this end, we need to define effective metrics for evaluating the behaviour of RAAS and continuously measure these metrics to evaluate the effect of AI. In the initial stages, it might be prudent to restrict the use of AI to a limited number of components of a RAAS (for instance, managing the knowledge base in the managing system with a MAPE-K architecture) or for offline analysis of the data collected during RAAS operation. As systems become more complex, ensuring resiliency and antifragility are likely to require the system to perform lifelong learning and potentially necessitate the use of AI in all system components (i.e., in an end-to-end manner).

4.8 Engineering resilient and antifragile autonomous systems

Amel Bennaceur (The Open University – Milton Keynes, GB)

Lee Barford (Keysight Technologies – London, GB)

Matteo Camilli (Politecnico di Milano, IT)

Marc Carwehl (Humboldt-Universität zu Berlin, DE)

Kerstin I. Eder (University of Bristol, GB)

Diego Perez-Palacin (Linnaeus University – Växjö, SE)

License © Creative Commons BY 4.0 International license

© Amel Bennaceur, Lee Barford, Matteo Camilli, Marc Carwehl, Kerstin I. Eder, and Diego Perez-Palacin

Engineering antifragile systems requires specialised consideration in each of the traditional software development process phases. The group explored the Requirements, Design, Implementation and Testing phases, and investigated requirements and KPIs for antifragility, how existing approaches can support these, and their limitations. In particular, antifragility involves learning and adaptation at runtime, in response to disturbances, for all stages of the engineering process.

From a requirements point of view, one challenge is defining suitable specifications that scope problems enough for driving design and testing, while allowing the system to evolve and adapt for future environments. From a design point of view, designs would need to satisfy uncertain specifications, enabling adaptation to new environments. From an implementation point of view, the challenge is striking a balance between scoping the problem to enable assurance, and allowing for adaptation at runtime. From a testing and analysis point of view, the challenge is to provide evidence of whether the system has improved when it has faced unspecified situations, especially when the specifications are uncertain.

5 Birds-of-a-Feather Groups

5.1 Formalising the relation of uncertainty, knowledge, decisions and antifragility

Ralf H. Reussner (KIT – Karlsruhe Institut für Technologie, DE)

Simon Burton (University of York, GB)

Felicita Di Giandomenico (CNR – Pisa, IT)

Kerstin I. Eder (University of Bristol, GB)

Vincenzo Grassi (University of Rome “Tor Vergata”, IT)

Ravi Mangal (Carnegie Mellon University – Pittsburgh, US)

Raffaella Mirandola (KIT – Karlsruhe Institut für Technologie, DE)

Patrizia Scandurra (University of Bergamo – Dalmine, IT)

License © Creative Commons BY 4.0 International license
 © Ralf H. Reussner, Simon Burton, Felicita Di Giandomenico, Kerstin I. Eder, Vincenzo Grassi, Ravi Mangal, Raffaella Mirandola, and Patrizia Scandurra

This birds-of-the-feather session discussed the relation of uncertainty, assumptions of decisions during the development process of cyber-physical systems, and antifragility. We concluded that uncertainty can be modelled as a property of assumptions which are formulated to make a justified decision in the development process. We agreed that the difference of resilience and antifragility is the ability of a system to learn from external events to improve reactions to such events. Such learning can be expressed in a changed uncertainty of the assumptions.

We identified several classes of metrics for antifragility: (i) metrics based on the improved reaction (including an improved quality), (ii) metrics based on the generality of learning, (iii) metrics based on the severity of the events dealt with (if this can be measured independently from the quality degrading impact), and (iv) metrics based on the sensitivity on events (i.e., the effort needed to react).

5.2 Resilience and antifragility of ethics-aware Human-AI collaborations

Amel Bennaceur (The Open University – Milton Keynes, GB)

Lee Barford (Keysight Technologies – London, GB)

Radu Calinescu (University of York, GB)

Ana Cavalcanti (University of York, GB)

Antje Loyal (Continental Automotive Technologies – Frankfurt, DE)

Lina Marsso (University of Toronto, CA)

Elena Navarro (University of Castilla-La Mancha, ES)

Shiva Nejati (University of Ottawa, CA)

Catia Trubiani (Gran Sasso Science Institute – L’Aquila, IT)

License © Creative Commons BY 4.0 International license
 © Amel Bennaceur, Lee Barford, Ana Cavalcanti, Radu Calinescu, Antje Loyal, Lina Marsso, Elena Navarro, Shiva Nejati, and Catia Trubiani

Ethics-aware Human-AI collaboration compounds multiple dimensions of uncertainty. First, uncertainty about ethical norms and their operationalisation. Second, uncertainty about human behaviour and values. Third, uncertainty about AI systems themselves, and the incomplete knowledge about the data, parameters, and performance in deployment. Furthermore, those uncertainties are interrelated, and none of their aspects can be considered in

isolation. This group focused on unravelling those uncertainties, illustrating them through examples, and investigating how existing reasoning techniques can help support some of those uncertainties.

Starting from eliciting requirements of ethics-aware Human-AI collaboration, one of the challenges is operationalisation into well-specified systems with well-defined capabilities and ethical/functional rules. Another challenge is about the assumptions (and obligations) about how humans interacting with the system behave. One source of disturbance is humans deviating from those assumptions.

We argued for the need for a theory for ethics-aware Human-AI collaboration grounded in mathematical modelling. We explored the reasoning features that need to be supported such as time, probabilities, non-determinism, interaction, and conformance. We also explored some of the techniques which might be used to support that reasoning, such as verification, synthesis, or goal-based requirements engineering. We also reviewed some of the available formalism that might support those reasoning features, including Markov and other stochastic models, hybrid process algebra, and fuzzy description logic. As none of the existing formalisms seems to support the reasoning needed for ethics-aware Human-AI collaboration, the question remains regarding how to integrate/unify and extend those different formalisms to address the different dimensions of uncertainty.

The group also explored the main building blocks for supporting the engineering of ethics-aware Human-AI collaboration based upon architectural patterns and their reification into reference implementation and reusable components. Similarly the reification of mathematical models into tools and domain-specific languages is needed for the specification of the ethics and functional requirements. Finally, standards and guidelines may guide the specification and engineering of those ethics-aware Human-AI collaborations.

5.3 Reasoning about antifragility from a control perspective

Mario Gleirscher (Universität Bremen, DE)

Marc Carwehl (Humboldt-Universität zu Berlin, DE)

Ada Diaconescu (Telecom Paris, FR)

Sebastián Uchitel (University of Buenos Aires, AR)

Gricel Vázquez (University of York, GB)

License  Creative Commons BY 4.0 International license

© Mario Gleirscher, Marc Carwehl, Ada Diaconescu, Sebastián Uchitel, and Gricel Vázquez

This birds-of-a-feather group had the objective of *developing a universal notion of antifragility based on the closed-loop control framework* widely applied in control theory and engineering. The discussions were aimed at a method for developing and evaluating controllers for an upcoming next generation of complex adaptive software systems, expected or even required to be increasingly resilient [1].

A collection of autonomous mobile robots working in a hospital was considered as an example following practical trends. These care robots must deliver documents, serve food to patient rooms, and interact with patients and staff.

The group's working hypothesis was that *antifragility* of such an application can be rephrased as a stability property of *resilience* as a quantity measured via an observed quality attribute of the application. This correspondence enables control-theoretic reasoning, for example, verifying whether a particular *adaptation manager* pushes the *resilience error* (i.e., the difference between observed and preferred resilience) below some threshold or whether the resilience level stabilises at a reference value.

During the discussion, we sketched a preliminary formal framework in support of this hypothesis. The framework is based on the notion of a signal, upon which the detection and evaluation of *disruptions* can be defined. The aggregated evaluation of the disruptions should then result in a characterisation of resilience and, moreover, allow one to observe antifragility. In particular, the outlined framework implies the notion of antifragility as the monotonic decrease of the resilience error, respectively, the monotonic increase of resilience over time, relative to a control loop, an adaptation manager, and a resilience profile. Overall, this notion resembles the desire of asymptotic stability of the control loop under consideration.

An important research challenge identified by the group is finding an appropriate adaptation manager for a particular control loop, such that the outlined monotonicity conditions are satisfied. In summary, the proposed control-theoretic perspective of resilience and antifragility enables the utilisation of further tools from control engineering in the search and design of adaptation managers responsible for improving resilience over time.

References

- 1 Jean-Claude Laprie. From dependability to resilience. In *Dependable Systems and Networks (DSN), 38th IEEE/IFIP Int. Conf.*, pages G8–G9, 2008.

5.4 Uncertainty propagation to support achieving antifragility

Sebastian Hahner (KIT – Karlsruher Institut für Technologie, DE)

Matteo Camilli (Politecnico di Milano, IT)

Andreas Heyl (Robert Bosch GmbH – Stuttgart, DE)

Antje Loyal (Continental Automotive Technologies – Frankfurt, DE)

Gabriel Moreno (Carnegie Mellon University – Pittsburgh, US)

Diego Perez-Palacin (Linnaeus University – Växjö, SE)

License © Creative Commons BY 4.0 International license

© Sebastian Hahner, Matteo Camilli, Andreas Heyl, Antje Loyal, Gabriel Moreno, and Diego Perez-Palacin

Uncertainty Flow Diagrams [1] are a recently proposed syntax, inspired by data flow diagrams and activity diagrams, that allows representing the system from the view of uncertainty to understand the existence of different uncertainties, their propagation along the operations and data flow, and to analyze uncertainty interaction. Some examples of systems that can benefit from the study of uncertainty propagation in their antifragility process implementation are self-adaptive systems such as *znn.com* and the software in the autonomous driving perception and decision modules. The former can combine different tactics to form strategies enhancing the service quality (e.g., by using different cloud providers, or changing the content quality), thus evolving its adaptation strategy at runtime, which also alters the uncertainty propagation. The latter uses sensor fusion to adapt to different environmental conditions and unanticipated change at runtime, e.g., due to sensor failures.

A possible benefit is to use the results of the uncertainty propagation analysis to measure the system antifragility, together with other metrics. A second benefit is to use the information resulting from an uncertainty propagation upstream analysis to identify the system elements that are increasingly contributing to the effect of uncertainty in the system decisions and trigger a system improvement process focusing on those elements.

In this group, we built on the aforementioned examples to compare three alternative approaches to measure antifragility: quality vs. time, uncertainty propagation depth, and covered conditional space. The first approach is the “classical” way to measure antifragility.

The second approach assumes that uncertainty which is mitigated earlier indicates a more antifragile system. The third approach regards the resilience of the system as being associated with properly handled uncertainty scenarios. Our initial findings indicate that all three approaches are equally suited for measuring antifragility, and can be interchanged. We proposed that approaches like the uncertainty propagation depth can also help the managing system (in a state-of-the-art MAPE-K model) assess and enhance its antifragility. Further research should investigate the flow and combination of different uncertainty sources and representations, and use the outcome of this investigation to build more resilient and antifragile systems.

References

- 1 Javier Cámara, Sebastian Hahner, Diego Perez-Palacin, Antonio Vallecillo, Maribel Acosta, Nelly Bencomo, Radu Calinescu, and Simos Gerasimou. Uncertainty Flow Diagrams: Towards a Systematic Representation of Uncertainty Propagation and Interaction in Adaptive Systems In *19th International Conference on Software Engineering for Adaptive and Self-Managing Systems*, 2024.

Participants

- Lee Barford
Keysight Technologies –
London, GB
- Amel Bennaceur
The Open University –
Milton Keynes, GB
- Simon Burton
University of York, GB
- Radu Calinescu
University of York, GB
- Matteo Camilli
Politecnico di Milano, IT
- Marc Carwehl
Humboldt-Universität zu
Berlin, DE
- Ana Cavalcanti
University of York, GB
- Felicita Di Giandomenico
CNR – Pisa, IT
- Ada Diaconescu
Telecom Paris, FR
- Kerstin I. Eder
University of Bristol, GB
- Mario Gleirscher
Universität Bremen, DE
- Vincenzo Grassi
University of Rome “Tor
Vergata”, IT
- Sebastian Hahner
KIT – Karlsruher Institut für
Technologie, DE
- Andreas Heyl
Robert Bosch GmbH –
Stuttgart, DE
- Antje Loyal
Continental Automotive
Technologies – Frankfurt, DE
- Ravi Mangal
Carnegie Mellon University –
Pittsburgh, US
- Lina Marusso
University of Toronto, CA
- Raffaella Mirandola
KIT – Karlsruher Institut für
Technologie, DE
- Gabriel Moreno
Carnegie Mellon University –
Pittsburgh, US
- Elena Navarro
University of Castilla –
La Mancha, ES
- Shiva Nejati
University of Ottawa, CA
- Diego Perez-Palacin
Linnaeus University – Växjö, SE
- Ralf H. Reussner
KIT – Karlsruher Institut für
Technologie, DE
- Patrizia Scandurra
University of Bergamo –
Dalmine, IT
- Catia Trubiani
Gran Sasso Science Institute –
L’Aquila, IT
- Sebastián Uchitel
University of Buenos Aires, AR
- Grisel Vázquez
University of York, GB

