Number 116 January 2019

# ERCIM NEWS

www.ercim.eu

*Special theme:*

# Transparency in Algorithmic Decision Making

*Research and Society:*
Ethics in Research

## KEYNOTE



## RESEARCHCH AND SOCIETY

This section about "Ethics in research has been coordinated by Claude Kirchner (Inria) and James Larrus (EPFL)



## SPECIAL THEME

The special theme "Transparency in Algorithmic Decision Making" has been coordinated by Andreas Rauber (TU Wien and SBA), Roberto Trasarti and Fosca Giannotti (ISTI-CNR).

Introduction to the special theme



## RESEARCH AND INNOVATION

This section features news about research activities and innovative developments from European research institutes



## ANNOUCEMENTS, IN BRIEF

# High-Level Expert Group on Artificial Intelligence

*Sabine Theresia Köszegi, Professor of Labor Science and Organization Institute of Management Science, TU Wien, Chair of the Austrian Council on Robotics and Artificial Intelligence, BMVIT, Member of the High-Level Expert Group on Artificial Intelligence of the European Commission.*

On 25 April 2018, the Europaen Commission published a Communication in which it announced an ambitious European Strategy for Artificial Intelligence (AI). The major advances in AI over the last decade revealed its capacity as a general-purpose technology and pushed inventions in areas of mobility, healthcare, home & service robotics, education and cyber security, to name just a few. These AI-enabled developments have the capability to generate tremendous benefits not only for individuals but also for the society as a whole. AI has also promising capabilities when it comes to address and resolve the grand challenges, such as climate change or global health and wellbeing, as expressed in the United Nations Sustainable Development goals. In competition with other key players, like the United States and China, Europe needs to leverage its current strengths, foster the enablers for innovation and technology uptake and find its unique selling proposition in AI to ensure a competitive advantage and a prosperous economic development in its Member States. At the same time, AI comes with risks and challenges associated to fundamental human rights and ethics. Europe therefore must ensure to craft a strategy that maximizes the benefits of AI while minimizing its risks.

The Commission has set out an interwoven strategy process between the development of a European AI Strategy and the development of a Coordinated Action Plan of Member States (hosted under the Digitising European Industry framework). The publication of the European policy and investment strategy on AI is envisaged for Summer 2019. To support this strategy development process and its implementation, the Commission has called for experts to establish a High-Level Expert Group on Artificial Intelligence (AI HLEG). Following an open selection process by DG Connect in spring 2018, the Commission has appointed 52 experts encompassing representatives from different disciplines of academia, including science and engineering disciplines and humanities alike, as well as representatives from industry and civil society. As an expert in labor science and with a research background in decision support systems, I was selected to join the exciting endeavor to lay the foundations for a human-centric, trustworthy AI in Europe that strengthens European competitiveness and addresses a citizen perspective to build an inclusive society.

Our mandate includes the elaboration of recommendations on the policy and investment strategy on ethical, legal and societal issues related to AI, including socio-economic challenges. Additionally, we serve as a steering group for the European AI Alliance to facilitate the Commission's outreach to the European society by engaging with multiple stakeholders, sharing information and gathering valuable stakeholder input to be reflected in our recommendations and work.

On 18 December 2018, we proposed a first draft on "Ethics Guidelines towards Trustworthy AI" to the Commission, setting out the fundamental rights, principles and values that AI has to comply with in order to ensure its ethical purpose. Additionally, we have listed and operationalized requirements for trustworthy AI as well as provided possible technical and non-technical implementation methods that should provide guidance on the realization of trustworthy AI. This draft on ethics guidelines is currently in a public consultation process in the European AI Alliance platform. Through this engagement with a broad and open multi-stakeholder & citizen forum across Europe and beyond, we aim to secure the open and inclusive discussion of all aspects of AI development and its impact on society. The finalised draft will be formally presented in the First Annual Assembly of the European AI Alliance in Spring 2019.

To advise the Commission with regards to the European policy and investment strategy, we are currently preparing a set of recommendations on how to create a valuable ecosystem for AI in Europe in order to strengthen Europe's competitiveness. The draft document of recommendations should be published in April 2019 and will undergo a public consultation process as well. The recommendations will primarily address European policy makers and regulators but also relevant stakeholders in Member States encompassing investors, researchers, public services and institutions. I would like to use the opportunity, to invite the readers of ERCIM News to engage in the European AI Alliance (see the link below) and to contribute your expertise and input to our policy and investment recommendations.

The complexity of AI-related challenges requires to set up a problem-solving process with highest information processing capacities that allows to consider different perspectives and to resolve conflicts of interest between different stakeholders. It can easily be imagined that our discussions as an inter-disciplinary expert and multi-stakeholder group are intense, difficult and at times emotional. In difficult situations, I remind myself of our commitment to the following statement in our ethics guidelines: "Trustworthy AI will be our north star, since human beings will only be able to confidently and fully reap the benefits of AI if they can trust the technology."

AI Alliance:
https://ec.europa.eu/digital-single-market/en/european-ai-alliance

Introcuction to the section "Research and Society"

# Ethics in Research

by Claude Kirchner (Inria) and James Larrus (EPFL)

***Science is in revolution. The formidable scientific and technological developments of the last century have dramatically transformed the way in which we conduct scientific research. The knowledge and applications that science produces has profound consequences on our society, both at the global level (for example, climate change) and the individual level (for example, impact of mobile devices on our daily lives). These developments also have a profound impact on the way scientists are working today and will work in the future. In particular, informatics and mathematics have changed the way we deal with data, simulations, models and digital twins, publications, and importantly, also with ethics.***

Ethics in research has been a field of inquiry since antiquity, in particular in health science and more recently in physics. Today all scientific disciplines, from medicine to biology, humanities, informatics, mathematics, physics and chemistry are troubled by ethical issues. They have been particularly popularized by dilemmas posed by the rise of machine learning, AI, and autonomous mechanized entities such as drones, robots, and vehicles.

In the fields of informatics and mathematics, there is a surge of interest in understanding and studying emerging ethical issues and methods for teaching them. Scientists are currently engaged in dialogues about these ethical issues with peers as well as the general population. Professional organisations, such as ERCIM, Informatics Europe, ACM, IEEE and IFIP, are discussing these issues and many advisory documents are being produced at the national and international levels. Let us point to a few examples of such contributions [3, 2, 1].

Of course, ethics in research is fundamental, but it is only one of the corner stones for the investigation of the ethical consequences of the impact of sciences, technologies, usages, and innovations on the digital evolutions. Ethical issues should be investigated everywhere in the world and in particular in Europe to allow the appropriation by everyone, person, organization, and company of the digital advances and transformations that are arising. We are indeed in a situation similar to bioethics in the 1980's with the necessity to set up ethical committees to consider the consequences of the digital innovations and transformations. Initiatives like the Montreal Declaration [L1], addressing "the responsible development of artificial intelligence, whether it is to contribute scientifically or technologically, to develop social projects, to elaborate rules (regulations, codes) that apply to it, to be able to contest bad or unwise approaches, or to be able to alert public opinion when necessary" are gaining visibility and, we hope, the necessary impact.

To contribute to raising awareness, this section, jointly coordinated by ERCIM and Informatics Europe, features four contributions on "ethics in research", focusing on informatics and mathematics and the possible interactions with other scientific disciplines. The contributions develop the following topics:

- how informatics, logic and mathematics can help to understand ethics in machine learning techniques;
- why and how reproducible research should be central to research developments;
- education in ethics and scientific integrity is now recognized as a priority, but how should we organize the training for doctoral candidates;
- how shall we organize research to avoid waste and bias in the exploration of scientific knowledge?

These contributions address a small but important part of the overall ethical issues, and we believe it is important and urgent to set up a joint working group between ERCIM and Informatics Europe, to contribute to the development of ethics research in our fields. We invite scientists to join the initiative by contacting us.

**Link:**

[L1] https://www.montrealdeclaration-responsibleai.com

**References**

[1] AIHLEG. Draft ethics guidelines for trustworthy AI [online]. December 2018. https://kwz.me/hdq

[2] CERNA. Research Ethics in Machine Learning. Research report, February 2018. https://hal.archives-ouvertes.fr/hal-01724307.

[3] J. Larus, et al.: "When computers decide: European recommendations on machine-learned automated decision making". Technical report, New York, NY, USA, 2018. https://dl.acm.org/citation.cfm?id=3185595/.

**Please contact:**

Claude Kirchner, Inria, France
claude.kirchner@inria.fr

James Larus, EPFL, Switzerland
james.larus@epfl.ch

# How to Include Ethics in Machine Learning Research

by Michele Loi and Markus Christen (University of Zurich)

***The use of machine learning in decision-making has triggered an intense debate about "fair algorithms". Given that fairness intuitions differ and can led to conflicting technical requirements, there is a pressing need to integrate ethical thinking into research and design of machine learning. We outline a framework showing how this can be done.***

One of the worst things to be accused of these days is discrimination. This is evident in some of the fierce responses on social media to recently published reports about the increasing proliferation of "discriminatory" classification and decision support algorithms; for example, the debate about the COMPAS prognosis instrument used to assess the relapse risk of prisoners [1].

This shows that researching and designing machine learning models for supporting or even replacing human decision-making in socially appropriate situations, such as hiring decisions, credit rating or criminal justice, is not solely a technical endeavor. The COMPAS case illustrates this exemplarily, as pointed out by the US-based investigative journalism NGO ProPublica. ProPublica showed that COMPAS made racist predictions, even though information about the race of the offender is not included in the calculation: Based on their risk scores, African Americans charged with a crime who do not reoffend within two years (out of prison) are predicted to be significantly at higher risk than whites who also do not reoffend within two years.. The algorithm thus violates the following idea of fairness: individuals who do not actually relapse should have the same probability that they will (unjustly?) be denied parole.

A standard research ethics answer to such a diagnosis would be the following: The result may point to discriminatory practice in designing COMPAS. One ethics answer would then be to increase the ethical integrity of the machine learning experts e.g. through better training or by increasing diversity in the team.

Unfortunately, the story is more complicated. The mathematicians who analysed the problem after the ProPublica revelations showed that a form of discriminatory distortion is inevitable – even if you program with a clear conscience and the data is free from bias [1]. Indeed, COMPAS was tested for discrimination – but another criterion for fairness was met: people who are denied (or granted) parole should have the same likelihood of relapse. This is achieved by "calibrated risk-scores" and using the same risk-score threshold for deciding whom to release. (Notice that using different thresholds for different groups also seems discriminatory.) The result is an algorithm that achieves "predictive value parity", i.e. the ratio of false positives and false negatives to predicted positives and predicted negatives (also known as Conditional Use Error [2]) is the same for both groups. This also seems intuitively required by fairness.

It turned out that it is mathematically impossible (except for irrelevant borderline cases) to meet both fairness conditions simultaneously [3]. In other words, you can either ensure that the people you release on parole are equally likely to commit crimes again, regardless of their race ("COMPAS-Fairness"). Or you can ensure that those who do not commit crimes are equally likely to be released from prison, regardless of their race ("ProPublica Fairness").

From a research ethics point-of-view this means that teaching the norm "avoid discrimination" to machine learning researchers would not work – as it is inevitable. A further difficulty in the assessment of fairness norms is that an algorithm fulfills a clear fairness constraint for individual decisions may have unexpected implications in the context, e.g. how do judges respond to algorithmic recommendation if they know that predictive value parity is not obtained? How do risk-scores evolve dynamically if more minority citizens are given loans that they are not able to repay?

The role of ethics in such a setting thus goes beyond transmitting norms about what is the right thing to do; it concerns increasing the moral sensitivity of the involved machine learning researchers such that they can identify broader effects of the systems they create. This task cannot be outsourced to those researchers. Rather, we should create working environments ("labs"), where computer scientists collaborate more closely with ethicists and domain experts. The rational analysis of the conceptual relationship between commonsense ideas of fairness and statistical properties of predictions is anything but a trivial task. And we need to answer questions like: What new skills do such ethicists need?

**References:**

[1] A. Chouldechova. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." ArXiv:1610.07524 [Cs, Stat]. http://arxiv.org/abs/1610.07524, 2016.

[2] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. "Fairness in Criminal Justice Risk Assessments: The State of the Art." ArXiv:1703.09207 [Stat], March. http://arxiv.org/abs/1703.09207, 2017.

[3] J. Kleinberg, S. Mullainathan, and M. Raghavan. "Inherent Trade-Offs in the Fair Determination of Risk Scores." ArXiv:1609.05807 [Cs, Stat]. http://arxiv.org/abs/1609.05807, 2016.

**Please contact:**

Michele Loi and Markus Christen
DSI Digital Ethics Lab, University of Zurich, Switzerland
michele.loi@uzh.ch, christen@ethik.uzh.ch

# Fostering Reproducible Research

by Arnaud Legrand (Univ. Grenoble Alpes/CNRS/Inria)

***To accelerate the adoption of reproducible research methods, researchers from CNRS and Inria have designed a MOOC targeting PhD students, research scientists and engineers working in any scientific domain.***

## A Reproducibility Crisis?

As the influence of computer science on society keeps growing, and in particular the recent widespread and enthusiastic use of AI/learning techniques, computer scientists are increasingly concerned with ethical issues. Fairness in recommendation systems, accountability in healthcare decision making process and guarantees in autonomous vehicular systems are only the tip of the iceberg.

The tuning of algorithms with (often heavily biased) real-life data involves lengthy and computationally intensive adjustments of hyper parameters and the classical training/testing procedures are often too costly to be rigorously followed and statistically rigorous. Furthermore, publishing results never requires a full disclosure of all the work, especially when trade secrets or confidential data is at stake. As a consequence many research results that may have worked in a specific restricted context turn out to be very difficult to reproduce by other independent researchers.

Although the root causes may be different, similar difficulties have been under the spotlight in every other scientific domain (in particular biology) and publicised under the terms of "reproducibility crisis". If data science and artificial intelligence are particularly exposed because of the hopes they inspire, it is computer science (operating systems, architecture, image processing) as a whole which suffers from similar reproducibility issues. It is past time that computer scientists adopt a robust and transparent research/experimental methodology.

## Reproducible Research

Reproducible research aims at facilitating the exploitation and reuse of research results by advocating for the use of computerised procedures and laboratory notebooks, for the full disclosure of code, data and provenance as well as for standardised and well-controlled statistical procedures and experimental testbeds. It is often seen as a solution to scientific integrity issues but it is foremost an essential step of modern science that has become increasingly complex and error-prone.

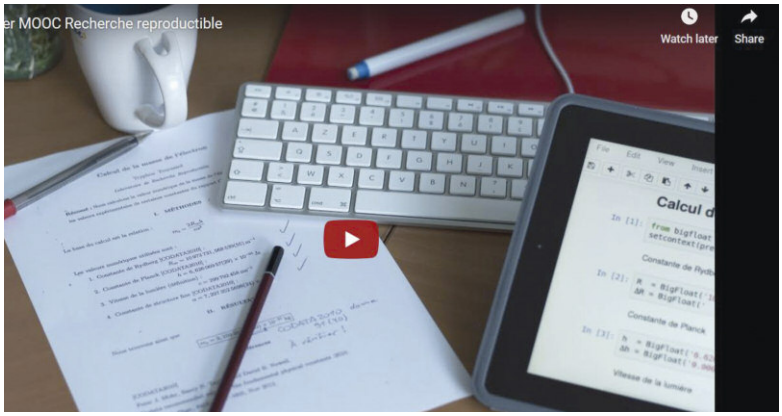
In recent years, many researchers, scientific institutions, funding agencies, and publishers, have started to initiate a change in our research and publication practices. Many conferences have set up artifact evaluation procedures and the ACM has proposed some reproducibility badges. The European Research Council, the National Science Foundation and many national research agencies now require data management plans and open-access publications to promote open science. But no standard practice clearly stands out as yet and researchers are often left unarmed to efficiently answer such requirements.

## Reproducible Research Challenges

Although computer programs are often thought of as deterministic systems, our algorithms, our software stack and our digital infrastructures have become so complex and evolve so quickly that even skilled computer scientists fail to fully control them. Several interesting projects stand out to address some of the key challenges of reproducible research:

*Explainability and tracability:* Computational notebooks have become particularly popular as they are easy to share and allow easy implementation of some form of literate programming which emphasises the narration. Scientific workflows have also become essential to orchestrate complex computations, track provenance and exploit parallel architectures.

*Software environment control and reconstruction:* Notebooks and workflows only track limited information regarding the software environment they build on. Without rigorously preserving this environment, the notebook will hardly be re-executable on another machine and will very likely either fail at runtime in a few months or, even worse, silently produce a different result. Virtual machines or containerization are often considered as a solution but only projects like Nix or Guix propose a clean solution to the traceability and the reconstructability of the environments.




*Screenshot of the teaser of the Mooc "Reproducible research": Methodological principles for a transparent science.*

*Software and data preservation:* Link rot in academic literature is a well-known issue but solutions emerge. The Software Heritage project addresses this by preserving software commons. The Zenodo data warehouse allows any researcher, regardless of their domain, to upload scientific data and share them in a perennial way.

## Accelerating Reproducible Research with a Massive Open Online Course

Carrying out reproducible research is obviously not a one-size-fits-all effort. Mastering all these technologies and integrating them in a daily methodology requires time. To get a majority of researchers started, we have designed with the

support of Inria a MOOC entitled "Reproducible Research: Methodological Principles for a Transparent Science" [L1] which targets graduate and PhD students, research scientists and engineers working in any scientific domain. The first edition of this MOOC started in October 2018 and has been followed by more than 1,000 individuals (out of about 3,200 registered individuals) working mostly in computer science and biology. This MOOC consists of four modules that combine videos and quizzes with exercises for acquiring hands-on experience with open source tools and methods (Markdown for taking structured and exploitable notes, GitLab for version control and collaborative working, Computational notebooks for efficiently combining the computation, presentation, and analysis of data). We propose three paths, each of which uses a different notebook technology: (1) Jupyter notebooks and the Python (or R) language, which requires no software installation on students' computers, (2) RStudio and the R language, and (3) the Org-Mode package of the Emacs editor and the languages Python and R. We also introduce the main challenges of reproducible research (data management, software environment, numerical issues) and present a few alternatives. At the end of this MOOC, students and researchers will have acquired good habits for preparing replicable documents and for sharing the results of their work in a transparent fashion.

**Link:**
[L1] https://learninglab.inria.fr/en/mooc-recherche-reproductible-principes-methodologiques-pour-une-science-transparente/

**Please contact:**
Arnaud Legrand
Univ. Grenoble Alpes/CNRS/Inria, France
arnaud.legrand@imag.fr

# Research Ethics and Integrity Training for Doctoral Candidates: Face-to-Face is Better!

by Catherine Tessier (Université de Toulouse)

***The University of Toulouse and Inria have set up face-to-face training in research ethics and integrity for doctoral candidates based on debates about their own theses.***

In order to meet the requirements of 2016 French legislation about doctoral studies, face-to-face training in research ethics and integrity for doctoral students has been conducted at Toulouse University since 2016 and at Inria since 2017. The training programme is based on the work [1] of CERNA, the French ethics committee for research in ICTs.

Based on the conviction that doctoral candidates should be able to express themselves and question and discuss issues of research ethics and integrity that directly concern themselves and their own theses, the training is organised as a single-day (i.e., six to seven hours) session with a maximum of twenty doctoral candidates per session. The originality of the approach lies in the fact that, on the one hand, classes can be multidisciplinary since they are open to doctoral students from any doctoral school; and on the other hand that the training is facilitated by two trainers preferably with different backgrounds.

The training material includes a set of slides, a collection of dilemma exercises and a trainer's guide [2]. The trainers commit to respecting the structure of the course i.e., to using the slides, supervising a dilemma exercise, asking each participant to write an ethical or integrity issue about their thesis and hosting class discussions about the issues that are put forward.

The slide-based lecture includes headline news examples of cases raising ethical or integrity issues, information about laws and codes of ethics, philosophical and historical background of ethics in science and useful vocabulary and concepts. The doctoral candidates' issues are collected and organised by the trainers in three topics: research integrity, publication and research ethics. Each set of issues is discussed by the whole class, and the trainers refer to the relevant available slides as the discussion goes along e.g., best research practices and misconducts, relationships with thesis supervisors (research integrity), authors, reviewers, citations (publications) and value conflicts during the thesis (research ethics). The slides are sent to the doctoral candidates after the training so that they can refer to them and visit the numerous links that are provided to relevant documents.

In order to increase the size of the pool of trainers, the training of trainers is done in situ, i.e., most often a pair of trainers includes an experienced trainer and a "new" trainer who may in turn train another colleague during a future session. The side effect of this structure is that the people involved as trainers are actually made aware of the issues of research ethics and integrity.

After more than two years' experience and over 1000 trained doctoral candidates, the following points are worth highlighting:

- It is not necessary for a trainer to be an ethics "expert". Nevertheless they must become familiar with the concepts, ask themselves about their own research practice and pay attention to scientific news and associated issues. Indeed a trainer enriches the course with their own experience and thinking.
- A trainer should not be afraid of disturbing issues, difficult situations, conflicting points of view, and where appropriate should be able to cope with emotion. Indeed sensitive issues are likely to be raised by doctoral candidates.
- Classes must be composed exclusively of doctoral candidates so that the issues that they raise can be addressed without the pressure of "senior" colleagues.
- In order for doctoral candidates to talk freely, it should be announced at the beginning of the session that everyone (both the trainers and the doctoral candidates) should commit to confidentiality; furthermore a doctoral candidate may remain anonymous when raising an issue about their

thesis (questions are put down in writing on stickers provided by the trainers).
- Trainers should announce that some issues – especially research ethics issues – do not have simple black and white answers and that the main point is to become familiar with ethical thinking and deliberation.
- Computers and cell-phones should be banned so that everyone can be involved in the debates.

Feedback from doctoral candidates about the training includes some very isolated but interesting comments:
- The dilemma exercise, which aims at creating ethical deliberation through the confrontation of moral values, can be regarded as aggressive and provoke rejection. On the contrary, it can be judged irrelevant because it is considered as being far from reality or pointless.
- Philosophical and historical references can provoke sharp or even hostile reactions, depending on each person's culture and beliefs.

Nevertheless, the overwhelming majority of feedback is positive, strengthening the pedagogical approach that has been adopted i.e., a face-to-face doctoral training in research ethics and integrity, in small classes, based on debates that are focused on the issues that are raised by the doctoral candidates themselves. Most doctoral candidates report that they have become aware of best practices concerning publications and experiments, that they have learnt to consider their work from an ethics point of view, and that they have appreciated being able to raise issues about their own theses. Clearly this costs money and the recruitment and remuneration policy of the trainers is an integral element of the (ethical) debate about this kind of training.

**References:**

[1] Commission de réflexion sur l'éthique de la recherche en sciences et technologies du numérique d'Allistene (CERNA), "Proposition de formation doctorale – Initiation à l'éthique de la recherche scientifique", 2018. [Online]. Available (in French): http://cerna-ethics-allistene.org/digitalAssets/55/55709_Cahier_CERNA_FormationDoctorale2018.pdf

[2] Formation Doctorale "Éthique de la recherche et intégrité scientifique": une formation proposée par l'école des docteurs de Toulouse, 2017. [Online]. Available (in French): https://hal.archives-ouvertes.fr/cel-01452867v2 (version 3 to come)

**Please contact :**
Catherine Tessier
ONERA/DTIS, Université de Toulouse
+33 5 62 25 29 14
catherine.tessier@onera.fr

# Efficient Accumulation of Scientific Knowledge, Research Waste and Accumulation Bias

by Judith ter Schure (CWI)

***An estimated 85 % of global health research investment is wasted [1]; a total of one hundred billion US dollars in the year 2009 when it was estimated. The movement to reduce this waste recommends that previous studies be taken into account when prioritising, designing and interpreting new research. Yet current practice to summarize previous studies ignores two crucial aspects: promising initial results are more likely to develop into (large) series of studies than their disappointing counterparts, and conclusive studies are more likely to trigger meta-analyses than not so noteworthy findings. Failing to account for these apects introduces 'accumulation bias', a term coined by our Machine Learning research group to study all possible dependencies potentially involved in meta-analysis. Accumulation bias asks for new statistical methods to limit incorrect decisions from health research while avoiding research waste.***

The CWI Machine Learning group in Amsterdam, The Netherlands, develops methods to allow for optional continuation in statistical testing. Thus, in contrast to standard statistical tests, the safe tests we develop retain their statistical validity if one decides, on the spot, to continue data collection and obtain a larger sample than initially planned for – for example because results so far look hopeful but not yet conclusive. Additional research into the application of these safe methods to meta-analysis was inspired by the replicability crisis in science and the movement to red uce research waste.

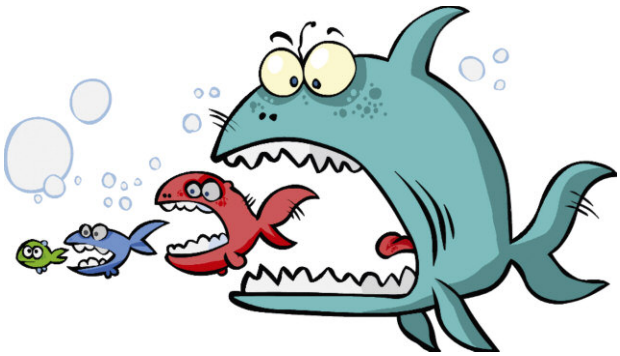
The 85 % research waste estimate is calculated by cumulatively considering waste in four successive stages of health research: (1) the choice of research questions, (2) the quality of research design and methods, (3) the adequacy of publication practices and (4) the quality of research reporting. In two of these stages, design and reporting, research waste is caused by a failure to systematically examine existing research. In terms of research design, the paper that estimated the research waste [1] stresses that new studies "should not be done unless, at the time it is initiated, the questions it proposes to address cannot be answered satisfactorily with existing evidence". Its recommendations about reporting involve that new studies should "be set in the context of systematic assessments of related studies" [1].

In 2014, a series of follow-up papers put forward by the REWARD Alliance showed that the 2009 recommendations remained just as pressing in 2014. The recommendation to always relate new studies to available research in design and reporting acquired the name evidence-based research in 2016

and has since then been promoted by the Evidence-Based Research Network.

Deciding whether existing evidence answers a research question is a difficult task that is further complicated when the accumulation of scientific studies is continuously monitored. This continuously monitoring is key to 'living systematic reviews', which are meta-analyses that incorporate new studies as they come available. Restricting further research when a certain boundary of evidence is crossed introduces bias, while continuous monitoring also creates multiple testing problems. As a result, reducing research waste is only feasible with statistical methods that allow for optional continuation or optional stopping.

Optional stopping is a well-studied phenomenon in statistics and machine learning, with a variety of approaches in the frequentist, Bayesian and online learning realm. These approaches are neatly combined, and much generalised, in the safe testing paradigm developed in our group. What is new to the meta-analysis setting is that dependencies arise even without continuously testing a series of studies. The very fact that a series of studies exists already introduces



*Large study series are more likely when they include initial promising results within the series than when they include very disappointing ones, just like the availability of the big fish in the cartoon depends on specific smaller fish available.*
*Image: Bobb Klissourski (Shutterstock)*

dependencies with results part of it that were at least not unacceptably disappointing to prohibit the expansion into the available series.

Meta-analysis is currently mainly considered when a study series of considerable size is available, with a median number of studies of around 15 in a typical meta-analysis [2]. Large study series are more likely when they include initial promising results within the series than when they include very disappointing ones, just like the availability of the big fish in the cartoon depends on specific smaller fish available. These dependencies introduce accumulation bias that in turn inflates false positive error rates when ignored in statistical testing.

Our research tries to determine how to deal with small-fish-dependent large fish, for various fish sizes. We intend to develop these methods for meta-analysis within the period of my PhD research (2017-2022) and involve other researchers from evidence based medicine, the reducing waste movement, psychology's reproducibility projects, and software projects such as JASP in implementing and communicating the results.

The recommendations of the 2009 research waste paper are increasingly being heard by chief scientific advisors, funders (such as the Dutch ZonMW) and centres for systematic reviews [3]. Now we need to implement them efficiently with the right statistics.

**Links:**

[L1] http://rewardalliance.net/
[L2] https://en.wikipedia.org/wiki/Evidence-based_research
[L3] http://ebrnetwork.org/

**References:**

[1] I. Chalmers, P. Glasziou: "Avoidable waste in the production and reporting of research evidence", The Lancet, 374(9683), 86-89, 2009.
[2] D. Moher, et al.: "Epidemiology and reporting characteristics of systematic reviews, PLoS medicine, 4(3), e78, 2007.
[3] P. Glasziou, I. Chalmers: "Research waste is still a scandal—an essay by Paul Glasziou and Iain Chalmers". Bmj, 363, k4645, 2018.

**Please contact:**

Judith ter Schure, CWI, The Netherlands
+31 20 592 4086
Judith.ter.Schure@cwi.nl

Introduction to the Special Theme

# Transparency in Algorithmic Decision Making

by Andreas Rauber (TU Wien and SBA), Roberto Trasarti, Fosca Giannotti (ISTI-CNR)

***The past decade has seen increasing deployment of powerful automated decision-making systems in settings ranging from smile detection on mobile phone cameras to control of safety-critical systems. While evidently powerful in solving complex tasks, these systems are typically completely opaque, i.e. they provide hardly any mechanisms to explore and understand their behaviour and the reasons underlying their decisions. This opaqueness raises numerous legal, ethical and practical concerns, which have led to initiatives and recommendations on how to address these problems, calling for greater scrutiny in the deployment of automated decision-making systems. Clearly, joint efforts are required across technical, legal, sociological and ethical domains to address these increasingly pressing issues.***

Machines are becoming increasingly responsible for decisions within society. Various groups, including professional societies in the area of information technologies, data analytics researchers, industry, and the general public are realising the power and potential of these technologies, accompanied by a sense of unease and an awareness of their potential dangers. Algorithms acting as black boxes make decisions that we are keen to follow as they frequently prove to be correct. Yet, no system is fool proof. Errors do and will occur no matter how much the underlying systems mature and improve due ever more data being available to them and ever more powerful algorithms learning from them.

This awareness has given rise to many initiatives aiming at mitigating this black-box problem, trying to understand the reasons for decisions taken by systems. These include the "ACM Statement on Algorithmic Transparency and Accountability" [1], Informatics Europe's "European Recommendations on Machine-Learned Automated Decision Making" [2] and the EU's GDPR regulation [3] which introduces, to some extent, a right for all individuals to obtain "meaningful explanations of the logic involved" when automated decision making takes place. One of the first activities of the newly established European Commission's High Level Expert Group on Artificial Intelligence (HLEG-AI) has been to prepare a draft report on ethics guidelines for trustworthy AI [4]. These documents pave the way towards a more transparent and sustainable deployment of machine-driven decision making systems. On the other hand, recent studies have shown that models learning from data can be attacked to intentionally provide wrong decisions via generated adversarial data. In spite of a surge in R&D activities in this domain [5], massive challenges thus remain to ensure automated decision making can be accountably deployed, and the resulting systems can be trusted.

The articles collated in this special theme provide an overview of the range of activities in this domain. They discuss the steps taken and methods being explored to make AI systems more comprehensible. They also show the limitations of current approaches, specifically as we leave the domain of analysing visual information. While visualisation of deep learning networks for image analysis in the form of heat maps, attention maps and the like [6,7] has helped drastically in understanding and interpreting the regions most relevant in image classification, other domains are frequently reverting to extracting rules as surrogates for or explanations of more complex machine learning models. While such rules are, in principle, fully transparent individually, their complexity frequently renders them unusable for understanding the decision making complexity.

We should consider the main question in the field to be: what is an explanation? This question in itself illustrates how new this research topic is. As yet there is no formalism for an explanation and nor is there a way to quantify the grade of comprehensibility of an explanation for humans. The following works are pioneers in this area, creating fertile ground for innovation.

Ricardo Guidotti and his colleagues provide a very good introduction into this black-box problem and approaches to mitigate it. They also propose that explanations be broken into two levels, namely a local level explanation on a data instance level, which is subsequently combined on a global level by synthesising the local explanations, and optimising them for simplicity and fidelity.

On the other hand, as research has frequently pointed out over decades, even human experts find it hard or impossible to provide clear, transparent explanations for their decisions. Fabrizio Falchi in his paper highlights the importance of lessons to be learned from the field of psychology in understanding intuition and how these can

assist with improving our understanding of deep learning algorithms.

Two articles delve deeper into the sources and effects of concerns about the lack of transparency, beyond just the complexity of the underlying machine learning models. Alina Sîrbu and colleagues review the process of opinion formation in society and how the "bias" introduced by selecting specific information being delivered to users influences the outcome of public debate and consensus building. Maliciously manipulated data provided as adversarial input to machine learning algorithms are reviewed by Fabio Carrara and colleagues, highlighting the need for a means to detect such attacks and the importance of making algorithms more robust against them.

Alexander Dür and colleagues go beyond explanations of the decisions made by Deep Learning network, focusing on ways to extract information on the impact of specific inputs on the decision in a text mining domain as they propagate through the network layers.

Owing to the high level of scrutiny it receives, the area of health is one of the dominant application domains. Tamara Müller and Pietro Lio address the need to provide explanations that are meaningful to the specific user, introducing a system that builds on top of a rule extraction system for Random Forests in order to inspect, tune and simplify these rules, and to provide visualisation support for their interpretation. Anirban Mukhopadhyay and colleagues also focus on the need for deeper understanding of the workings and reasons behind decisions made by AI systems. While heat maps allow visualisation of the area most relevant for a decision, little information is provided about the actual reason for a certain decision being made. They identify three challenges that go beyond the technicalities of the algorithms, including the availability of data, a regulatory approval process and the integration of the doctor - patient relationship into the evaluation.

In a similar vein, Carmen Fernández and Alberto Fernández review the ethical and legal implications of the use of AI technology in recruiting software, proposing a separation of concerns via a multi-agent system architecture as a mechanism to regulate competing interests.

The final three papers in this special theme section present examples from other application domains. Ulrik Franke discusses a new project that has been set up to study transparency in the insurance industry, whilst Max Landauer and Florian Skopik highlight issues with the semantic expressiveness of log data elements for cyber threat identification. Last, but not least, Markus Berg and Sebastian Velten provide an example from the scheduling domain where transparency issues do not arise from the complexity of a black-box deep learning model, but from the tardiness of the underlying processes due to resource constraints in the underlying optimisation processes.

We strongly believe that the new wave of interest in the field, coupled with the existing big opportunities and challenges will produce a new era where AI will support many human activities. For this reason, society needs to open the black box of AI to empower individuals against undesired effects of automated decision making, to reveal and protect new vulnerabilities, to implement the "right of explanation", to improve industrial standards for developing AI-powered products, increasing the trust of companies and consumers, to help people make better decisions, to align algorithms with human values and finally to preserve (and expand) human autonomy.

**References:**


[1] ACM Policy Council: Statement on Algorithmic Transparency and Accountability, 2017.

[2] Informatics Europe and EUACM: When Computers Decide: European Recommendations on Machine-Learned Automated Decision Making, 2018.

[3] European Parliament: Regulation (EU) 2016/679 of the European Parliament and Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016.

[4] High-Level Expert Group on Artificial Intelligence. Draft Ethics Guidelines for Trustworthy AI, European Commission, Dec. 18 2018.

[5] R. Guidotti, et al.: "A Survey of Methods for Explaining Black Box Models", ACM Computing Surveys 51(5):93, DOI 10.1145/3236009.

[6] J. Yosinski, et al.: "Understanding neural networks through deep visualization", in International Conference on Machine Learning (ICML) Workshop on Deep Learning, 2015.

[7] P. Rajpurkar, et al.: "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning", arXiv Preprint arXiv:1711.05225v3, Dec. 2017.


**Please contact:**


Andreas Rauber
TU Vienna, Austria
rauber@ifs.tuwien.ac.at

Roberto Trasarti, Fosca Giannotti
ISTI-CNR, Italy
roberto.trasarti@isti.cnr.it
fosca.giannotti@isti.cnr.it

# The AI Black Box Explanation Problem

by Riccardo Guidotti, Anna Monreale and Dino Pedreschi (KDDLab, ISTI-CNR Pisa and University of Pisa)

***Explainable AI is an essential component of a "Human AI", i.e., an AI that expands human experience, instead of replacing it. It will be impossible to gain the trust of people in AI tools that make crucial decisions in an opaque way without explaining the rationale followed, especially in areas where we do not want to completely delegate decisions to machines.***

On the contrary, the last decade has witnessed the rise of a black box society [1]. Black box AI systems for automated decision making, often based on machine learning over big data, map a user's features into a class predicting the behavioural traits of individuals, such as credit risk, health status, etc., without exposing the reasons why. This is problematic not only for lack of transparency, but also for possible biases inherited by the algorithms from human prejudices and collection artifacts hidden in the training data, which may lead to unfair or wrong decisions [2].

Machine learning constructs decision-making systems based on data describing the digital traces of human activities. Consequently, black box models may reflect human biases and prejudices. Many controversial cases have already highlighted the problems with delegating decision making to black box algorithms in many sensitive domains, including crime prediction, personality scoring, image classification, etc. Striking examples include those of COMPAS [L1] and Amazon [L2] where the predictive models discriminate minorities based on an ethnic bias in the training data.

The EU General Data Protection Regulation introduces a right of explanation for individuals to obtain "meaningful information of the logic involved" when automated decision-making takes place with "legal or similarly relevant effects" on individuals [L3]. Without a technology capable of explaining the logic of black boxes, this right will either remain a "dead letter", or outlaw many applications of opaque AI decision making systems.

It is clear that a missing step in the construction of a machine learning model is precisely the explanation of its logic, expressed in a comprehensible, human-readable format, that highlights the biases learned by the model, allowing AI developers and other stakeholders to understand and validate its decision rationale. This limitation impacts not only information ethics, but also accountability, safety and industrial liability [3]. Companies increasingly market services and products with embedded machine learning components, often in safety-critical industries such as self-driving cars, robotic assistants, and personalised medicine. How can companies trust their products without understanding the logic of their model components?

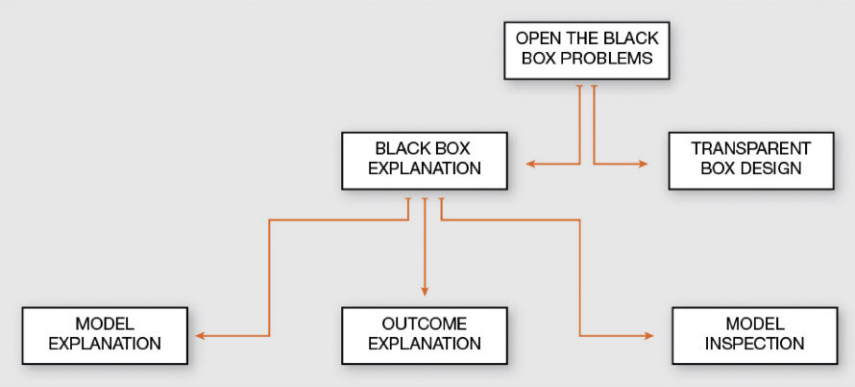



*Figure 1: Open the black box problems taxonomy.*

At a very high level, we articulated the problem in two different flavours:

- *eXplanation by Design* (XbD): given a dataset of training decision records, how to develop a machine learning decision model together with its explanation;
- *Black Box eXplanation* (BBX): given the decision records produced by a black box decision model, how to reconstruct an explanation for it.

In the XbD setting the aim is to provide a transparent machine learning decision model providing an explanation of the model's logic by design. On the other hand, the BBX problem can be resolved with methods for auditing and finding an explanation for an obscure machine learning model, i.e., a black box for which the internals are unknown. In fact, only the decision behaviour of the black box can be observed. As displayed in Figure 1, the BBX problem can be further decomposed into:

- model explanation when the explanation involves the whole (global) logic of the black box classifier;
- outcome explanation when the target is to (locally) understand the reasons for the decision of a given record;
- model inspection when the object is to understand how internally the black box behaves changing the input by means of a visual tool.

We are focusing on the open challenge of constructing a global meaningful explanation for a black box model in the BBX setting by exploiting local explanations of why a specific case has received a certain classification outcome. Specifically, we are working on a new local-first explanation framework that works under the assumptions of: (i) local explanations: the global decision boundaries of a black box can be arbitrarily complex to understand, but in the neighbourhood of each specific data point there is a high chance that the decision boundary is clear and simple; (ii) explanation composition: there is a high chance that similar records admit similar explanations, and similar explanations are likely to be composed together into more general explanations. These

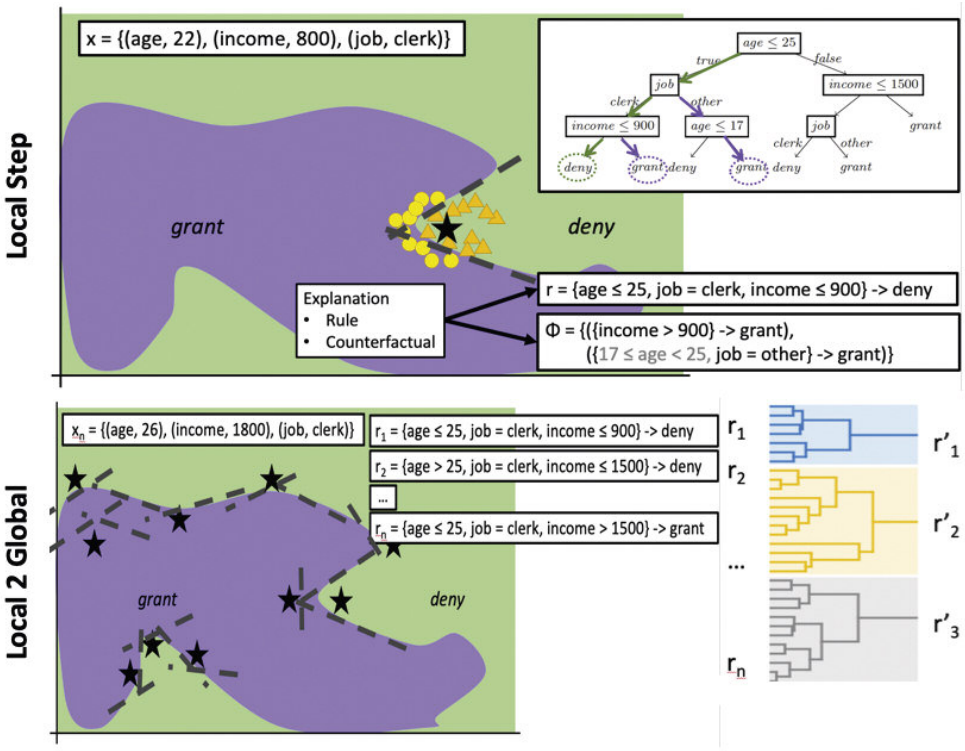



*Figure 2: Local-first global explanation framework.*

assumptions suggest a two-step, local-first approach to the BBX problem:

- Local Step: for any record x in the set of instances to explain, query the black box to label a set of synthetic examples in the neighbourhood of x which are then used to derive a local explanation rule using an interpretable classifier (Figure 2 top).
- Local-to-Global Step: consider the set of local explanations constructed at the local step and synthesise a smaller set by iteratively composing and generalising together similar explanations, optimising for simplicity and fidelity (Figure 2 bottom).

The most innovative part is the Local-to-Global (L2G) Step. At each iteration, L2G merges the two closest explanations e1, e2 by using a notion of similarity defined as the normalized intersection of the coverages of e1, e2 on a given record set X. An explanation e covers a record x if all the requirements of e are satisfied by x, i.e., boundary constraints, e.g. age > 26. L2G stops merging explanations by considering the relative trade-off gain between model simplicity and fidelity in mimicking the black box. The result is a hierarchy of explanations that can be represented by using a dendrogram (a tree-like diagram, Figure 2 bottom right).

We argue that the L2G approach has the potential to advance the state of art significantly, opening the door to a wide variety of alternative technical solutions along different dimensions: the variety of data sources (relational, text, images, etc.), the variety of learning problems (binary and multi-label classification, regression, scoring, etc.), the variety of languages for expressing meaningful explanations. With the caveat that impactful, widely adopted solutions to the explainable AI problem will be only made possible by truly interdisciplinary research, bridging data science and AI with human sciences, including philosophy and cognitive psychology.

This article is coauthored with Fosca Giannotti, Salvatore Ruggieri, Mattia Setzu, and Franco Turini (KDDLab, ISTI-CNR Pisa and University of Pisa).

**Links:**

[L1] https://kwz.me/hd9
[L2] https://kwz.me/hdf
[L3] http://ec.europa.eu/justice/data-protection/

**References:**

[1] F. Pasquale: "The black box society: The secret algorithms that control money and information", Harvard University Press, 2015.
[2] R. Guidotti, et al.: "A survey of methods for explaining black box models", ACM Computing Surveys (CSUR), 51(5), 93, 2018.
[3] J. Kroll, et al.: "Accountable algorithms", U. Pa. L. Rev., 165, 633, 2016.

**Please contact:**

Riccardo Guidotti, ISTI-CNR, Italy
+39 377 9933326,
riccardo.guidotti@isti.cnr.it

Anna Monreale, University of Pisa, Italy
+39 328 2598903,
anna.monreale@unipi.it

Dino Pedreschi, University of Pisa, Italy
+39 348 6544616,
dino.pedreschi@unipi.it

# About Deep Learning, Intuition and Thinking

by Fabrizio Falchi, (ISTI-CNR)

***In recent years, expert intuition has been a hot topic within the discipline of psychology and decision making. The results of this research can help in understanding deep learning; the driving force behind the AI renaissance, which started in 2012.***

"Intuition is nothing more and nothing less than recognition" [1], is a famous quote by Herbert Simon, who received the Turing Award in 1975 and the Nobel Prize in 1978. As explained by Daniel Kahneman, another Nobel Prize winner, in his book Thinking, Fast and Slow [2], and during his talk at Google in 2011 [L1]: "There is really no difference between the physician recognising a particular disease from a facial expression and a little child learning, pointing to something and saying doggie. The little child has no idea what the clues are but he just said, he just knows this is a dog without knowing why he knows". These milestones should be used as a guideline to help understanding decision making in recent AI algorithms and thus their transparency.

Most of the recent progress in artificial intelligence (AI) has been on recognition tasks, and this progress has been achieved through the adoption of deep learning (DL) methods. The AI renaissance started in 2012 when a deep neural network, built by Hinton's team, won the ImageNet Large Scale Visual Recognition Challenge. Deep learning methods have been, and still are, the driving force behind this renaissance. Like the little child mentioned by Kahneman, a state-of-the-art deep neural network is able to look at something and say "doggie", without knowing why it knows. In other words, the task of recognition, especially in computer vision, has been solved by DL methods with a form of artificial intuition. And this is not a surprise given that important researchers such as Simon have accepted the equivalence between intuition and recognition.

Even if many people feel a sense of magic talking about DL, the research conducted in recent years has proven that there is no magic at all in intuition and the same holds for DL.

Within the discipline of psychology and decision making, expert intuition has been discussed a lot in recent years, dividing researchers into believers and skeptics. However, after six years of discussion, a believer, Gary Klein, and a skeptic, Daniel Kahneman, wrote an important paper in 2009 whose subtitle was "A failure to disagree". Trying to answer the question When can we trust intuition? they agreed on a set of conditions for trustable intuitive expertise. Among these conditions, the most important ones are:

- an environment that is sufficiently regular to be predictable;
- an opportunity to learn these regularities through prolonged practice.

I believe most of the researchers working on DL would agree that those are also good conditions for the question: When can we trust deep learning? In fact, in order for a DL method to learn, we need a large training set (prolonged practice) and this set must be representative of the application scenario in which the environment must be sufficiently regular to be predictable.

What degree of transparency can we ask for from DL methods? Following the metaphor between DL and intuition, we can look at what Simon said about human recognition capabilities: "we do not have access to the processes that allow us to recognise a familiar object or person". I believe the same is true for DL. Even if we can monitor the flow of information in a deep neural network, we don't understand the "process".

Nevertheless, DL methods can be transparent in some terms: knowledge about the used training set and an in-depth analysis of the statistical outcomes can help in making them trustable for a specific task, in a specific context at a specific time.

As humans should not rely on intuition for all decisions, DL methods should be used as part of more complex AI systems that also involve non-intuitive processes. Kahneman has used the metaphor of two systems in his research about human thinking:

- System 1: fast, automatic frequent, emotional, stereotypic, unconscious.
- System 2: slow, effortful, infrequent, logical, calculating, conscious.

It is not by chance that DeepMind AlphaGo, the program that in 2016 defeated South Korean professional Go player Lee Sedol, combines DL with Monte Carlo tree search. As Michael Wooldridge, chair of the IJCAI Awards Committee said, "AlphaGo achieves what it does through a brilliant combination of classic AI techniques as well as the state-of-the-art machine learning techniques that DeepMind is so closely associated with". Following our metaphor, AlphaGo is a good example of collaboration between System 1 and System 2. AlphaGo uses DL to provide an intuitive estimation of the likelihood that the next stone will be placed in a specific place and of the final outcome of the game given the current status. However, the final decision about where to put a stone is made using the Monte Carlo tree search (AlphaGo System 2). In other words, AlphaGo uses the outcome of artificial intuition implemented using DL methods (its System 1) but takes decisions with logical reasoning (its System 2).

The few examples discussed here show that psychology can help in understanding AI. When Simon was conducting his research, psychology and AI were closely linked. This is a link that we need to revisit.

**Link:** [L1] https://kwz.me/hdj

**References:**

[1] H. A. Simon: "What is an 'explanation' of behavior?" Psychological science 3.3, 150-161, 1992.

[2] D. Kahnemann: "Thinking, Fast and Slow", Farrar, Straus and Giroux, 2011.

**Please contact:**

Fabrizio Falchi, ISTI-CNR, Italy
+39 050 315 29 11
Fabrizio.falchi@cnr.it

# Public Opinion and Algorithmic Bias

by Alina Sîrbu (University of Pisa), Fosca Giannotti (ISTI-CNR), Dino Pedreschi (University of Pisa) and János Kertész (Central European University)

***Does the use of online platforms to share opinions contribute to the polarization of the public debate? An answer from a modelling perspective.***

The process of formation of opinions in a society is complex and depends on a multitude of factors. Some relate to personal preferences, culture or education. Interaction with peers is also relevant: we discuss important issues with friends and change or reinforce our opinions daily. Another significant effect is that from the media: external information reaches and influences us constantly. The choice of the people we interact with, and of the news we read, is thus crucial in the formation of our opinions. In the last decade, the patterns of interaction with peers, and of news consumption, have changed dramatically. While previously one would read the local newspaper and discuss with close friends and neighbours, nowadays people can interact at large distances and read news across the world through online media. Social media in particular is increasingly used to share opinions, but also, as the Reuters 2018 Digital News Report shows, to read and share news. This means that the peer and external effects in the dynamics of opinions are becoming susceptible to influences from the design of the social media platforms in use. These platforms are built with a marketing target in mind: to maximise the number of users and the time they spend on the platform. To achieve this, the information that reaches the users is not randomly selected. A so called 'algorithmic bias' exists: we see the news related to topics that we like and the opinions of friends that are close to our opinions, so that we are driven to read them and come back to the platform. However, a question arises: does that interfere in any way with the formation of our opinions? Do we ever change our minds any more, or we just keep reinforcing our positions? Do we ever see things in a different perspective, or we are now closed in our little information bubbles?

A group of researchers from the Knowledge Discovery and Data Mining Laboratory in Pisa, and the Department of Network and Data Science of the Central European University, funded by the European Commission, are trying to answer some of these questions by building models of opinion dynamics that mimic formation of opinions in society. Based on the evolution of opinions, a group of people can reach consensus, i.e. agreement on a certain topic, or fragmentation and polarisation of society can emerge. This process can be modelled by representing opinions with continuous numbers, and simulating interactions with specific rules to change opinions at regular intervals. The population typically starts from a random configuration and evolves in time to either consensus or a fragmented state. One very popular model for such simulations is the 'bounded confidence' model, where peers interact only if their opinion is close enough. In this model, clusters of opinion appear if the confidence is low, while for large confidence consensus emerges. This model has been modified to include algorithmic bias. Instead of selecting peers to interact with in a random way, they are selected with a bias: a person is more likely to choose to interact with a peer that has an opinion close to their

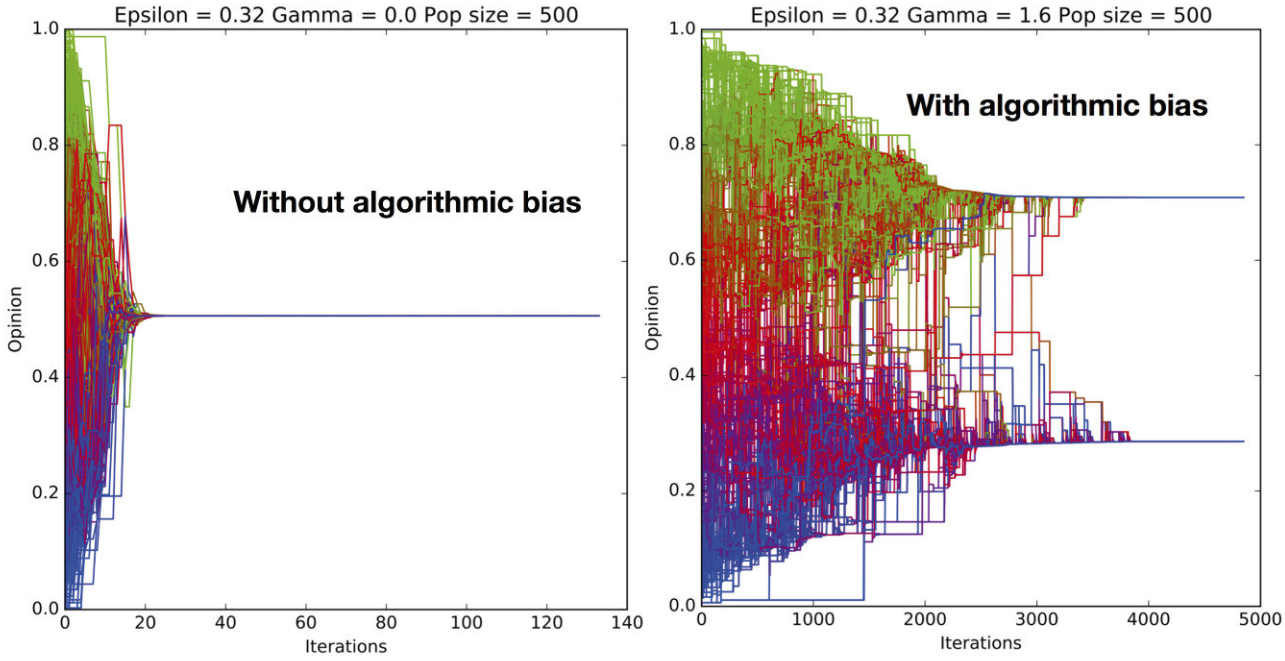



*Figure 1: Simulation of opinion formation with and without the bias: The bias slows down the process and leads to the formation of two clusters instead of one.*

own, while it will have a low probability of interaction with opinions far from their own.

Simulations of the algorithmic bias model show several results that suggest that online platforms can have important effect on opinion formation and consensus in society. First, the number of opinion clusters grows when algorithmic bias grows (see illustration). This means that online platforms can favour fragmentation of opinions. Second, this leads also to polarisation, where the distance between the opinions of the people is larger compared to the situation without algorithmic bias. Third, the changes in opinion are much slower when the bias is in operation. Even when consensus is obtained, the time to reach it becomes very long. In practice, this means that it could take years for people to agree on an issue, being in a highly fragmented state while this occurs.

These results bring important evidence that algorithmic bias may affect outcomes of public debates and consensus in society. Thus, we believe measures are required to at least stop its effects, if not reverse them. Researchers are investigating means of promoting consensus to counteract for the algorithmic bias effects. In the meantime, users could be informed of the way platforms feed information and the fact that this could affect their opinions, and maybe the mechanisms implemented by the platforms could be slowly withdrawn.

**Reference:**
[1] Alina Sîrbu, et al.: "Algorithmic bias amplifies opinion polarization: A bounded confidence model", arXiv preprint arXiv:1803.02111, 2018. https://arxiv.org/abs/1803.02111

**Please contact:**
Alina Sîrbu,
University of Pisa, Italy
alina.sirbu@unipi.it

# Detecting Adversarial Inputs by Looking in the Black Box

by Fabio Carrara, Fabrizio Falchi, Giuseppe Amato (ISTI-CNR), Rudy Becarelli and Roberto Caldelli (CNIT Research Unit at MICC – University of Florence)

***The astonishing and cryptic effectiveness of Deep Neural Networks comes with the critical vulnerability to adversarial inputs — samples maliciously crafted to confuse and hinder machine learning models. Insights into the internal representations learned by deep models can help to explain their decisions and estimate their confidence, which can enable us to trace, characterise, and filter out adversarial attacks.***

Machine learning and deep learning are pervading the application space in many directions. The ability of Deep Neural Network (DNN) to learn an optimised hierarchy of representations of the input has been proven in many sophisticated tasks, such as computer vision, natural language processing and automatic speech recognition. As a consequence, deep learning methodologies are increasingly tested in security- (e.g. malware detection, content moderation, biometric access control) and safety-aware (e.g. autonomous driving vehicles, medical diagnostics) applications in which their performance plays a critical role.

However, one of the main roadblocks to their adoption in these stringent contexts is the diffuse difficulty to ground the decision the model is taking. The phenomenon of adversarial inputs is a striking example of this problem. Adversarial inputs are carefully crafted samples (generated by an adversary — thus the name) that look authentic to human inspection, but cause the targeted model to misbehave (see Figure 1). Although they resemble legitimate inputs, the high non-linearity of DNNs permits maliciously added perturbations to steer at will the decisions the model takes without being noticed. Moreover, the generation of these malicious samples does not require a complete knowledge of the attacked system and is often efficient. This exposes systems with machine learning technologies to potential security threats.

Many techniques for increasing the model's robustness or removing the adversarial perturbations have been developed, but unfortunately, only a few provide effective countermeasures for specific attacks, while no or marginal mitigations exist for stronger attack models. Improving the explainability of models and getting deeper insights into their internals are fundamental steps toward effective defensive mechanisms for adversarial inputs and machine learning security in general.

To this end, in a joint effort between the AIMIR Research Group of ISTI-CNR and the CNIT Research Unit at MICC (University of Florence), we analysed the internal representations learned by deep neural networks and their evolution throughout the network when adversarial attacks are performed. Opening the "black box" permitted us to characterise the trace left in the activations throughout the layers of the network and discern adversarial inputs among authentic ones.

We recently proposed solutions for the detection of adversarial inputs in the context of large-scale image recognition with deep neural networks. The rationale of our approaches is to attach to each prediction of the model an authenticity score estimating how much the internal representations differ from expected ones (represented by the

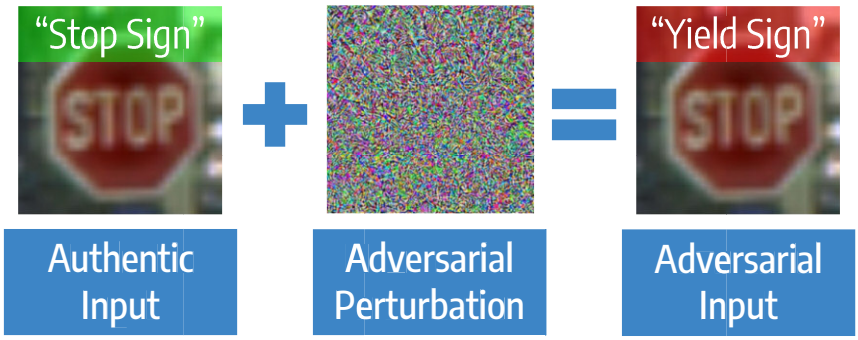



*Figure 1:Example of a common adversarial attack on image classifiers. The adversarial perturbation added (magnified for visualization purposes) fools the network to predict a wrong class with high confidence.*

model's training set). In [1], such a score is obtained by analysing the neighbourhood of the input with a nearest-neighbour search in the activation space of a particular layer. Our experiments on adversarial detection permitted us to identify the internal activations which are influenced the most by common adversarial attacks and to filter out most of the spurious predictions in the basic zero-knowledge attack model (see [L1]).

Building on this idea, in [2] we improved our detection scheme considering the entire evolution of activations throughout the network. An evolution map is built by tracing the positions an input occupies in the feature spaces of each layer with respect to most common reference points (identified by looking to training set inputs). Experiments showed that adversarial inputs usually tend to deviate from reference points leading to different activation traces in the network with respect to authentic inputs (see Figure 2). Thus, conditioning our detector on such information permitted us to obtain remarkable detection performance under commonly used attacks.

We plan to extend our analysis in order to fully characterise the effect of adversarial attacks on internal activations even in stricter attack models, i.e. when the attacker is aware of defensive systems and tries to circumvent it.

Despite our experimentation on adversarial input detection, both the presented approaches actually aim to cope with a broader problem, which is assigning a confidence to a model's decision by explaining it in terms of the observed training data. We believe this is a promising direction for reliable and dependable AI.

**Links:**
[L1]
http://deepfeatures.org/adversarials/

**References:**
[1] Carrara et al.: "Adversarial image detection in deep neural networks", Multimedia Tools and Applications, 1-21, 2018
[2] Carrara et al.: "Adversarial examples detection in features distance spaces", ECCV 2018 Workshops, 2018.

**Please contact:**
Fabio Carrara, ISTI-CNR, Italy
fabio.carrara@isti.cnr.it

Roberto Caldelli, CNIT Research Unit at MICC – University of Florence, Italy
roberto.caldelli@unifi.it

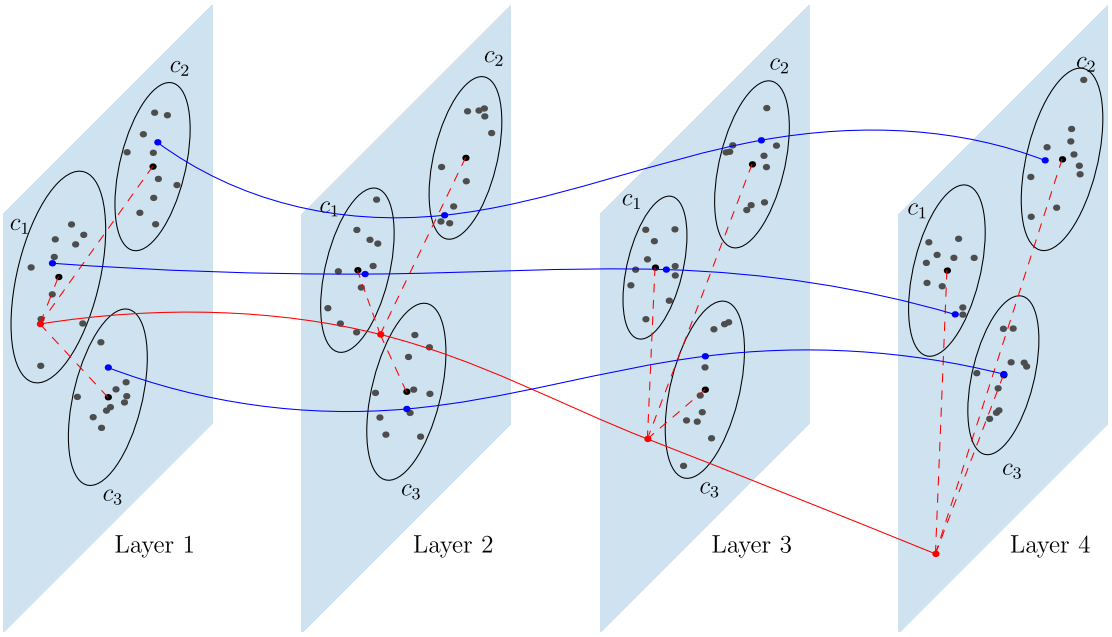



*Figure 2: Conceptualisation of the evolution of features while traversing the network. Each plane represents a feature space defined by the activations of a particular layer of the deep neural network. Circles on the features space represent clusters of features belonging to a specific class. Blue trajectories represent authentic inputs belonging to three different classes, and the red trajectory represent an adversarial input. We rely on the distances in the feature space (red dashed lines) between the input and some reference points representatives of the classes to encode the evolution of the activations.*

# Inspecting the Behaviour of Deep Learning Neural Networks

by Alexander Dür, Peter Filzmoser (TU Wien) and Andreas Rauber (TU Wien and Secure Business Austria)

***With the desire and need to be able to trust decision making systems, understanding the inner workings of complex deep learning neural network architectures may soon replace qualitative or quantitative performance as the primary focus of investigation and measure of success. We report on a study investigating a complex deep learning neural network architecture aimed at detecting causality relations between pairs of statements. It demonstrates the need to obtain a better understanding of what actually constitutes sufficient and useful insights into the behaviour of such architectures that go beyond mere transformation into rule-based representations.***

Recently there has been increased pressure by legislators as well as ethics boards and the public at large to ensure that algorithmic decision making systems also provide means for inspecting their behaviour and are able to explain how they arrive at any specific decision. This is basically well-aligned with researchers' desire to understand the workings of an algorithm as this usually constitutes the most structured and only viable approach to improving an algorithm's overall performance.

Naïve approaches such as untargeted parameter sweeps and architecture variations leading to higher performance in some specific benchmark settings turn out to be useless as they do not provide any rationale guiding the process of how to repeat such optimisations for specific tasks in different settings. Inspection of deep learning (DL) networks to understand what they are doing, which parts of the input are most influential in the final decision making, provide valuable insights both in understanding a specific model as well as guiding the targeted design of improved architectures, data representations and learning routines. Examples such as attention highlighting in image processing revealed both insights into the characteristics learned by a network, as well as uncovered errors and bias in the resulting systems, such as, for example, the unintentional focus on logos embedded in images as a clear separator between positive and negative class images due to the construction of the training data base.

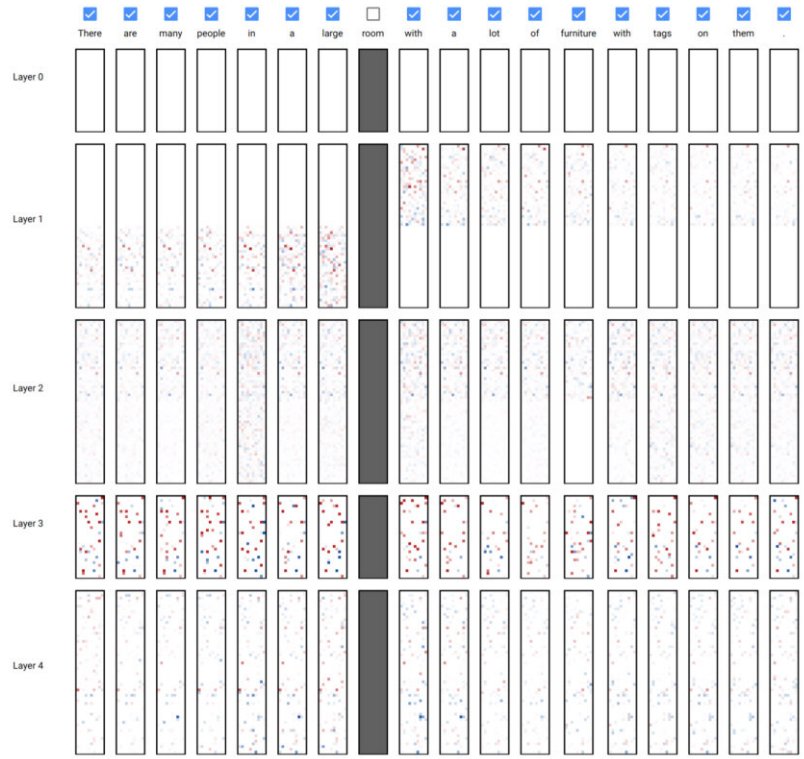



*Figure 1: Activation differences for the first four layers of a neural natural language inference model trained on the SNLI corpus [3]. Layer 0 is an embedding lookup, layer 1 and 2 are bidirectional RNNs, layer 3 is a combination of an attention and a feed forward layer, layer 4 is a bidirectional RNN layer processing the two prior layers.*

Image analysis settings offer themselves for inspection as the data processed can conveniently be displayed in visual form, thus becoming early candidates for identifying areas of attention [1]. Discovering the structure learned by various network layers, such as the focus on edges and orientation of these in subsequent layers of a convolutional neural network (NN) architecture [2], provide intuitive insights into their behaviour. Settings that do not offer themselves for direct visual inspection provide way harder challenges. Even more so, tasks that go beyond mere classification of data based on individual, independent attributes increase the challenge in devising interpretable representations of the inner workings of such complex DL architectures. Here we review the challenges of trying to devise such an inspection in a setting of a neural language inference model, where the goal is to correctly classify the logical relationship between a pair of sentences into one of three categories: entailment, neutral or contradiction.

Current attempts at explaining and understanding neural language processing models are primarily based on the visualisation of attention matrices. While this technique provides valuable insights into the inner workings of such models it is only focused on their attention layers and ignores all other types of layers.

Our approach to understanding complex neural network architectures is based on the analysis of the interaction patterns of a single input word with all

other words. We do this by comparing the network's activations on the original input to its activations when removing individual words. The resulting differences in activations show the interaction between different words of the input in different layers. The interactive tool we built allows users to enter a baseline input and directly perturbate this input by excluding words and observing the influence on activations through all layers of the network including the model's predictive output.

Figure 1 shows how the initial removal of a single noun affects the processing of the activations belonging to all other words. The words most strongly influenced are those that have a linguistic relationship with the removed word, like a preposition referencing a noun. Similarly, we analysed (amongst other perturbations) the effect of changing the word order in the source or target sentences revealing the impact of positional characteristics.

State-of-the-art models for many natural language processing tasks are artificial neural networks which are widely considered to be black box models. Improvements are often the result of untargeted parameter sweeps and architectural modifications. In order to efficiently and systematically improve such models a deeper understanding of their inner workings is needed. We argue that interactive exploration through input perturbation is a promising and versatile approach for inspecting neural networks' decision processes and finding specific target areas for improvement.

**References:**
[1] M. D. Zeiler, R. Fergus: "Visualizing and understanding convolutional networks", European Conference on Computer Vision, Springer, Cham, 2014.
[2] J. Yosinski, et al.: "Understanding neural networks through deep visualization", in International Conference on Machine Learning (ICML) Workshop on Deep Learning, 2015.
[3] S. Bowman, et al.: "A large annotated corpus for learning natural language inference", in Proc. of EMNLP, 2015.

**Please contact:**
Alexander Dür, TU Wien, Austria
+4369919023712,
alexander.duer@tuwien.ac.at

# Personalisable Clinical Decision Support System

by Tamara Müller and Pietro Lió (University of Cambridge)

***We introduce a Clinical Decision Support System (CDSS) as an operation of translational medicine. It is based on random forests, is personalisable and allows a clear insight into the decision making process. A well-structured rule set is created and every rule of the decision making process can be observed by the user (physician). Furthermore, the user has an impact on the creation of the final rule set and the algorithm allows the comparison of different diseases as well as regional differences in the same disease.***

Neurodegenerative diseases such as Alzheimer's and Parkinson's impact millions of people worldwide. Early diagnosis has proven to greatly increase the chances of slowing the diseases' progression [2]. Correct diagnosis often relies on the analysis of large amounts of patient data, and thus lends itself well to support from machine learning algorithms, which are able to learn from past diagnosis and see clearly through the complex interactions of a patient's symptoms. Unfortunately, many contemporary machine learning techniques fail to reveal details about how they reach their conclusions, a property considered fundamental when providing a diagnosis. This is one reason why we introduce our personalisable CDSS that provides a clear insight into the decision making process on top of the diagnosis. Our algorithm enriches the fundamental work of Masheyekhi and Gras [1] in data integration, personal medicine, usability, visualisation and interactivity.

Our algorithm performs by extracting a rule set from a random forest, which is then minimised within several steps. The algorithm can be divided into three major steps. (1) Firstly, a random forest, an ensemble of decision trees, is created as the foundation of the algorithm. (2) Secondly, a set of rules is extracted from the random forest. (3) Thirdly, this rule set is reduced significantly. The user can influence step (3) by preferring as important considered features. The algorithm is implemented in Python and we trained it to predict whether patients are likely to suffer from Alzheimer's or Parkinson's disease, but it is a generic algorithm that can be applied to any kind of disease. Three main factors are taken into account during the reduction process: (a) the performance of individual rules, (b) the rules' transparency, and (c) the personal preferences of the user. The reduction leads to a new, smaller set of rules, with predictive performance comparable to the original set, but much easier to comprehend. A clear understanding of the process behind a diagnosis is crucial for both doctor and patient, and it is hoped that systems like this one will become increasingly prevalent as we continue to improve the state-of-the art in predictive medicine.

Mashayekhi and Gras [1] introduced two methods called RF+HC and RF+HC_CMPR which allow to extract a rule set from random forests. The main idea of their work is to reduce the number of rules dramatically and therefore increase the comprehensibility of the underlying model. We expanded this idea and added another personalisable layer to it to allow the user to add data driven and personal evidence to the algorithm.

We applied our algorithm to different data sets with a variety of data types like magnetic resonance images, bio-

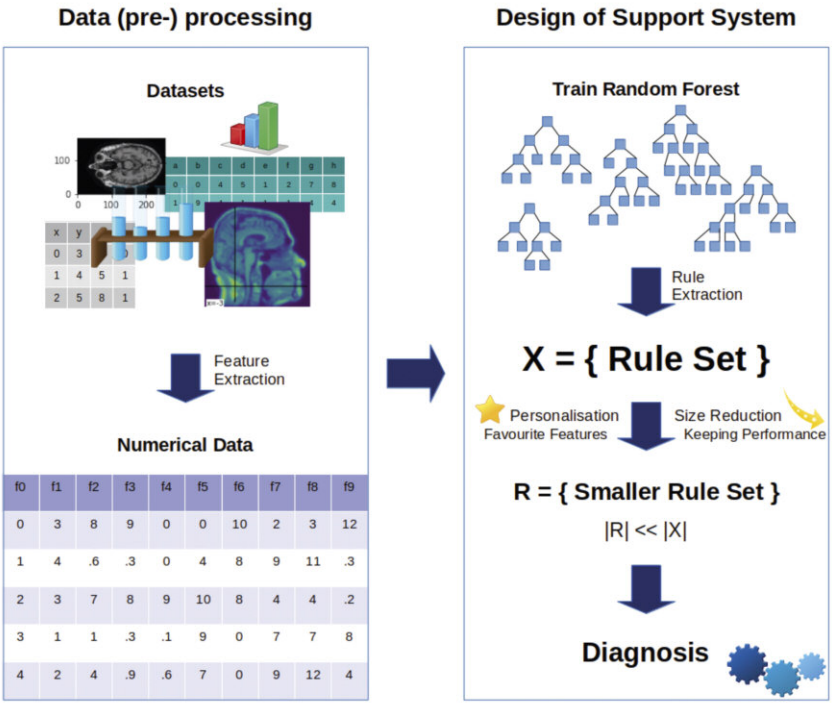



*Figure 1: A visual overview of the algorithm.*

medical voice measurements, drawings, demographic characteristics, etc. After the rule set is extracted from the forest, each rule is assigned with a score. This score depends on the rule's length, its performance on the training set, and whether it contains any preferred features. Based on this score, the weakest rules are eliminated to minimise the rule set. The application to Alzheimer's and Parkinson's data sets has shown that the reduction of the rule set combined with the consideration of preferred features, does not generally impair the performance. It sometimes even has a positive impact on the prediction, as setting preferred features can diminish the risk of over-fitting and take regional characteristics and expertise into account. Furthermore, a deliberately reduced set of rules is less likely to contain noisy rules [1]. The algorithm also reveals information about the importance of features, which allows to draw conclusions about diseases and their indicators. We achieved accuracies of up to 100% and one rule set could be reduced to 0.5% of the original rule set size without reducing its performance, e.g.

A graphical user interface allows the algorithm to be used easily, where data of a new patient can be added intuitively. Furthermore, all rules can be inspected by the physicians. It is also possible to show statistical distributions to get a better understanding of which features are more or less important in the decision making process. It allows to compare different diseases and detect regional differences like urban characteristics versus rural ones or international variations in diseases.

**References:**

[1] M. Mashayekhi, R. Gras: "Rule extraction from random forest: the RF+ HC methods", 2015.

[2] Alzheimer's Association: "2017 Alzheimer's disease facts and figures" Alzheimer's & Dementia 13, no. 4 (2017): 325-373.

**Please contact:**
Tamara Müller, Pietro Lió
University of Cambridge, UK
contact@tamaramueller.com
pl219@cam.ac.uk

# Putting Trust First in the Translation of AI for Healthcare

by Anirban Mukhopadhyay, David Kügler (TU Darmstadt), Andreas Bucher (University Hospital Frankfurt), Dieter Fellner (Fraunhofer IGD and TU Darmstadt) and Thomas Vogl (University Hospital Frankfurt)

***From screening diseases to personalised precision treatments, AI is showing promise in healthcare. But how comfortable should we feel about giving black box algorithms the power to heal or kill us?***

In healthcare, trust is the basis of the doctor-patient relationship. A patient expects the doctor to act reliably and with precision and to explain options and decisions. The same accuracy and transparency should be expected of computational systems redefining the workflow in healthcare. Since such systems have inherent uncertainties, it is imperative to understand a) the reasoning behind such decisions and b) why mistakes occur. Anything short of this transparency will adversely affect the fabric of trust in these systems and consequently impact the doctor-patient relationship.

Current solutions for transparency in deep learning (used synonymously with AI) centre around the generation of heatmaps. These highlight high-impact image regions on deep learning decisions. While informative in nature, direct adaptation of such methods into healthcare is insufficient, because the actual reasoning patterns remain opaque and leave a lot of room for guesswork. Here, deep generative networks show promise by generating visual clues as to why a decision was made [1].

We believe transparency in image based algorithmic decision making can only be achieved if expert computer scientists and healthcare professionals (radiologists, pathologists etc.) closely collaborate in an equal-share environment.

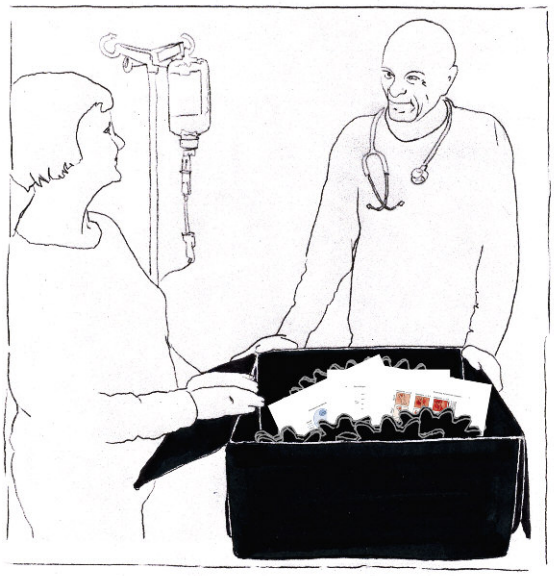


*Figure 1: Doctors and patients are important stakeholders in the discussion about safe and transparent AI. Involving them in solutions is crucial for successful applications.*

In Central Germany, TU Darmstadt and Goethe University Frankfurt have formed an interdisciplinary expert working group of computer scientists and radiologists.

We identified three challenges that render interpretable and robust AI in medical applications difficult, and started researching systematic solutions:

- Bias, quality and availability of medical data for data-driven algorithms,
- Strict requirements by the regulatory approval process and
- Integration of AI into the doctor-patient relationship.

### Bias, quality and availability of medical data for data-driven algorithms

Human beings are unique. The considerable inherent variability of human physiology is generally addressed by curated medical datasets and large case numbers. Additional variability is introduced by the complexity of healthcare facilities. When dealing with diseases, often the number of relevant variables is unknown in addition to the causality. Furthermore, data acquisition is sometimes limited by the need to restrict some investigations or treatments, owing to negative side effects on human health (e.g. radiation exposure of computer tomography). In comparison, other AI application areas can use vast Big Data databases by crawling the web or collected from users. This is compounded by the fact that indication, acquisition and in part interpretation of medical image data has by and large not undergone the standardisation already present in other clinical tests (e.g. blood sample tests). As such, clinics are not currently structured to collect data for traditional data-driven algorithms. Our initial work has shown specialised handling of such data significantly improves the performance of deep learning algorithms [2].

In a day to day context, medical annotations often remain guesses, whose uncertainty is acceptable to justify individual treatment decisions based on averaged cut-off values. To solve this contradiction, we model and incorporate the uncertainty of annotations into AI-based methods.

### Strict requirements by the regulatory approval process

The volume of documentation and the regulatory complexity for licensing in health care are high for good reason: the stakes are high. Through regularisation and a high number of parameters, deep learning achieves high accuracy and good generalisation, but also obscures the insights into reasoning patterns and learned knowledge. One unexplained phenomenon can be found in adversarial examples: only minimally perturbed, crafted images lead to AI-assessments that contradict those (Figure 2) of the unperturbed image. Our work on single pixel flips introduces a systematic analysis for patterns of adversarial examples [3].

Effective adversarial detection and prevention methods are required to avoid harm to patients. If left ignored, a malicious attacker, be it competitor or criminal organisation, can use this vulnerability to damage or extort medical device manufacturers and hospitals. Improvements are required to meet GDPR and documentation requirements. Based on our initial work [3], we focus on a systematic understanding of such adversarial examples.

### Integration of AI into the doctor-patient relationship

How would you feel if you saw your doctor "google" all your questions? Theoretical machine learning research on interpretability often ignores the crucial "humane" aspect of trust between patient and medical expert – and the role a doctor would take as "black-box vocalizer" in this context. When implemented wisely, AI should enhance, not replace, the decision making process. The ethical obligation to make the most informed decisions on what often resembles a life-altering or -ending change for the patient might make AI a necessity in the near future.

In light of the GDPR regulations for a "Right-to-Explanation" and the required high standards for the documentation of all medical examinations

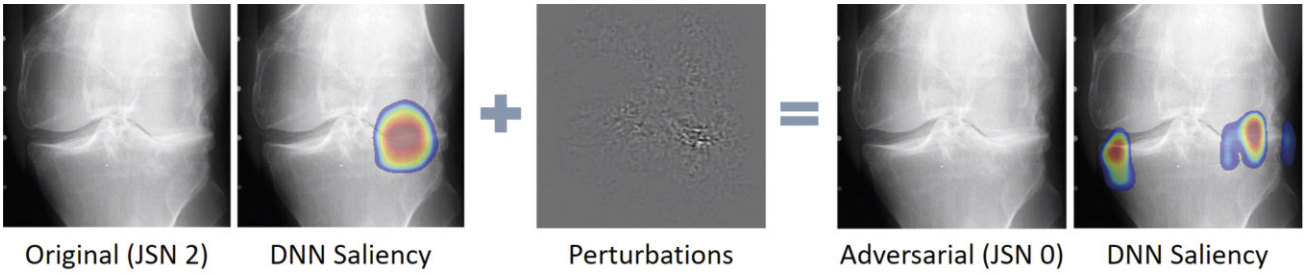



*Figure 2: With perturbations beyond human perception, deep learning often leads to both wrong decisions and wrong corresponding interpretation.*

and incidents, this black-box behaviour of obscuring the reasoning is a significant obstacle for the clinical implementation of AI in healthcare. As (closely) collaborating experts, we are developing courses and guidelines to make doctors AI-ready.

In retrospective in-silico studies, deep learning-based algorithms are already showing promise to be the single most successful concept in medical image analysis. In order to realise the potential impact these algorithms can have on our healthcare, we need to introduce trust-focused prospective clinical studies. Transparency in decision making has developed to be a key component we expect from our doctors. As deep learning is introduced into healthcare, we need standardised methods to guarantee results and transparency to doctors.

**References:**

[1] J. C. Y. Seah, et al.: "Chest radiographs in congestive heart failure: visualizing neural network learning", Radiology, 2018.
[2] M. Sahu, et al.: "Addressing multi-label imbalance problem of surgical tool detection using CNN", IJCARS, 2017.
[3] D. Kügler, et al.: "Exploring Adversarial Examples", Understanding and Interpreting Machine Learning in Medical Image Computing Applications, Springer, Cham, 2018.

**Please contact:**

Anirban Mukhopadhyay
Informatik, Technische Universität Darmstadt, Germany
anirban.mukhopadhyay@gris.tu-darmstadt.de

# Ethical and Legal Implications of AI Recruiting Software

by Carmen Fernández and Alberto Fernández (Universidad Rey Juan Carlos)

***Artificial Intelligence (AI) applications may have different ethical and legal implications depending on the domain. One application of AI is analysis of video-interviews during the recruitment process. There are pros and cons to using AI in this context, and potential ethical and legal consequences for candidates, companies and states. There is a deficit of regulation of these systems, and a need for external and neutral auditing of the types of analysis made in interviews. We propose a multi-agent system architecture for further control and neutral auditing to guarantee a fair, inclusive and accurate AI and to reduce the potential for discrimination, for example on the basis of race or gender, in the job market.***

## Image analysis in human resources: pros and cons

There has been a recent trend towards video-interview analysis in HR departments. Traditionally, AI played no more than an assistant role in HR, e.g. resume and CV scanning. But lately, apps and systems like HireVue [L1], Montage [L2], SparkHire [L3] and WePow [L4] have been changing how recruitment is carried out. An AI-based video interview system could be programmed to check, during an interview, features such as age, lighting, tone of voice, cadence, keywords used (substantial conversation), mood, behaviour (eccentric, movement or quite calm and not talkative), eye contact and, above all, emotions. AI targets the specific traits of a customer-oriented role that employers want in their teams.

AI has produced benefits for HR so far, including:

- Reduction of interview time per candidate, thus recruiting time.
- Customised candidate experience and customised questions and answers.
- Attention to detail (eye contact time, emotions-intonation and body language) and lack of interviewer bias (physical appearance, tattoos, etc.).

But there are several problems that accompany the use of these technologies:

- Candidates are unfamiliar with video interview analysis (for example, lighting, settings), which could affect global performance.
- Gender and racial bias: traditionally, machine learning algorithms were trained with data from white people.
- Imprecisions of technology. Training classifiers with biased datasets. For instance, Affectiva [L5] dataset of human emotions was fed with data from Superbowl viewers, and could presumably have culture-bias.

## Controversial characteristics

We studied several potential controversial characteristics, among them, facial symmetry, race, gender, sexual orientations in voice and image recordings. The problem of racial-bias in AI is not new, just like the detection of mixed race in bad lighting conditions according to Siyao et all [1]. MIT researchers acknowledged race-bias in learning algorithms mainly trained with data from white people.

As an illustration of the advances in sexual orientation recognition both in images and sound, one study [2] needed ethical supervision due to the opaque invasive nature of the research and the use of real user data from dating applications. The study argues that there is a relationship between homosexuality and exposure to particular concentrations of hormones in the womb, and that sexual orientation can be determined by morphological features (for example, the jawline and forehead).

## Ethical and legal aspects of AI

Whilst the use of AI in this context may have its benefits, it also strips away aspects of humanity, reducing a human recruit to a set of descriptors. The automation of HR processes could lead to potential ethical and legal implications that cannot be ignored. In some countries, companies are not allowed to ask a candidate's age

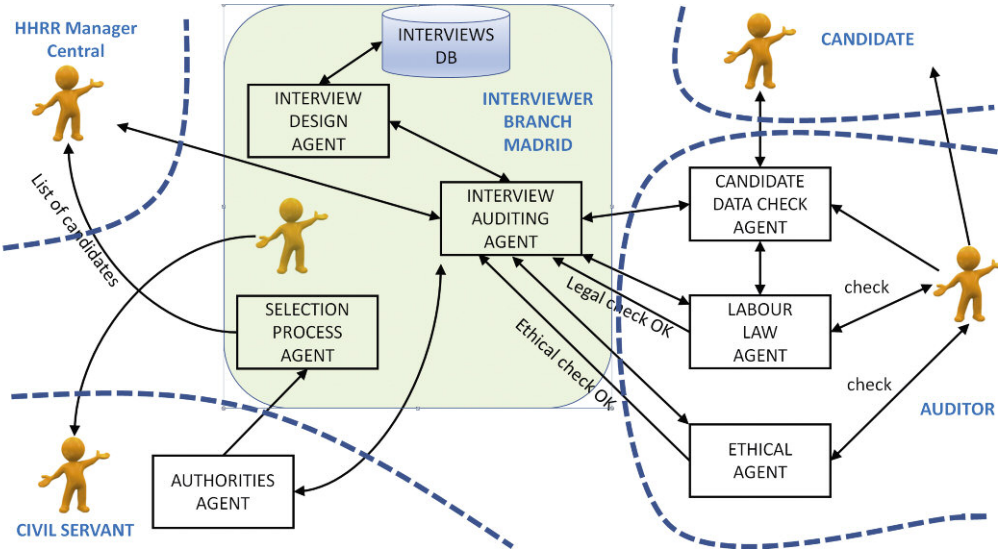



*Figure 1: Multiagent System architecture for auditing.*

during recruitment. Traditionally, United States legislation has been particularly protective of racial differences and discrimination in the workplace (the Civil Rights Act, 1964, forbids "improperly classifying or segregating employees by race"). And yet, even while these regulations exist to reduce discrimination, enterprises are given more and more freedom to customise their systems. We conclude it is risky to blindly follow the adoption of AI in recruiting.

### Multi-agent system architecture

At CETINIA (URJC) we are working on a multi-agent system architecture for auditing (Figure 1). The core of the architecture comprises three different parties that must collaborate: (i) a recruiter/company, (ii) external auditor, and (iii) government/authorities.

An Interview design agent, based at the company central headquarters, is responsible for designing a general interview. The Interview auditing agent is based in company branches and applies the general interview format to a regional scenario of the country where the recruiting is taking place. The Selection process agent can cancel the process due to controversies or give back a list of candidates to the central office if the process is fair. It is also capable of running checks with authorities and auditors.

If the features analysed in the recruiting process break any law or if the process contravenes basic civil rights, the interview process agent would ask for the approval of the Labour Law Agent or Ethical Agent if necessary. If the recruiting process is dealing with a candidate's personal information, it would require the candidate's approval for data handling. If a company is recruiting in another country it would need to register with the Authorities agent.

**Links:**

[L1] https://www.hirevue.com/
[L2] https://www.montagetalent.com/
[L3] https://www.sparkhire.com/
[L4] https://www.wepow.com/es/
[L5] https://www.affectiva.com/

**References:**

[1] Siyao Fu, Haibo He, Zeng-Guang Hou: "Learning Race from Face: A Survey", IEEE Trans. Pattern Anal. Mach. Intell. 36(12): 2483-2509 (2014)
[2] M. Kosinski, Y. Wang: "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images", Journal of Personality and Social Psychology 114 (2), 246-257, 2018

**Please contact:**

Carmen Fernández
Universidad Rey Juan Carlos, Spain
carmen.urjc@gmail.com

# Towards Increased Transparency in Digital Insurance

by Ulrik Franke (RISE SICS)

***Automated decision-making has the potential to increase both productivity and competitiveness as well as compensate for well-known human biases and cognitive flaws [1]. But today's powerful machine-learning based technical solutions also bring about problems of their own – not least in terms of being uncomfortably black-box like. A new research project at RISE Research Institutes of Sweden, in collaboration with KTH Royal Institute of Technology, has recently been set up to study transparency in the insurance industry, a sector that is poised to undergo technological disruption.***

At the Danish Insurance 2018 conference in Copenhagen in February, an IBM partner in the exhibition area marketed Watson with the tagline "Insurance company employees are working with Watson to assess insurance claims 25% faster". The fact that such efforts are underway is no surprise to the informatics or mathematics

communities. Yet, it serves as an excellent illustration of both the promise and the perils that the insurance industry faces in the digital world.

Start with the promise. Insurance still largely relies on manual labor throughout its value-chain. Though digitally supported, assessing risk, setting premiums, and adjusting claims most often involves human decision-making. With smarter tools, whether based on rules or machine-learning, more work could certainly be automated, cutting costs for companies and premiums for customers.

However, the road to digital insurance can be rocky. In the past few years, there have been abundant reports of flawed machine-learning applications, e.g., the infamous misclassification of people as monkeys in image-recognition systems, signs of racial bias in systems predicting the probability of criminal recidivism, and sexism in natural language processing. How can an insurance company that starts employing machine-learning at larger scale be sure that risk is properly assessed, that premiums levels are sustainable, and that customers are fairly treated when claims are adjusted?

Such questions prompted Länsförsäkringars forskningsfond, the research arm of a major Swedish insurer, to issue a call for research proposals related to digitalization and insurance. The resulting project, Transparent algorithms in insurance [L1], started in October 2018 and aims to explore issues of algorithmic transparency, broadly construed, in the insurance industry. During the project's two-year stint, several research questions will be studied, some technical, some more of a social science nature.

One important part of the project will be to study conditions for algorithmic transparency in insurance. Many techniques for transparency in machine-learning systems exist, but they are mostly designed to fit technologies such as deep neural networks or support vector machines. However, meaningful transparency also needs to account for industry, e.g., insurance, and stakeholders. In other words, meaningful algorithmic transparency probably looks different to a CEO, to an actuary, to a software engineer, and to a consumer looking for a car insurance.

Another important strand is ethics. Ethicists have argued that the implementation of algorithms entails ethical judgements and trade-offs, though not always explicit. The project will engage with ongoing software projects with the funder and work with development teams to make explicit the ethically relevant choices made. A promising technique is the paradigm of value-sensitive design [2], which has been successfully employed in other areas.

A third project cornerstone will be to explore the consequences of increased transparency in insurance. Some are expected to be decidedly positive, ranging from better requirements engineering in software projects to better customer value and even growing market shares. Indeed, insurance startups and would-be disruptors such as Lemonade in the US and Hedvig in Sweden use transparency, in various forms, as selling points.

However, the insurance industry is particularly interesting because there is also a significant potential downside to transparency. Insurance companies need to design their products so that customers' incentives don't change for the worse, e.g., a car owner driving recklessly because the insurance coverage will take care of the consequences. Mechanisms such as deductibles are designed precisely to prevent this moral hazard. Similarly, the fact that the insured typically knows more than the insurer, e.g., experiencing an ache that a physician cannot detect, might lead to more risk among insureds than expected. Again, mechanisms such as indemnity limits are designed to prevent such adverse selection. While these conceptual problems inherent to insurance have been known for a long time, an increased dependence on algorithms paired with greater transparency might invite insureds to novel forms of undesirable strategic behavior, which ultimately leads to higher premiums for customers. For example, it is probably self-defeating to be transparent about algorithms designed to detect insurance fraud. Thus, the project will investigate needs and techniques for selective transparency.

The project is still in its infancy, but in two years the research team will have taken first steps in designing the modern software quality assurance processes that will enable Länsförsäkringar and the rest of the insurance industry to make more informed decisions about how to best make use of the advances in machine-learning. As "insurers face unprecedented competitive pressure owing to technological change" [3], it will certainly be an interesting road ahead.

**Link:**

[L1] https://www.ri.se/en/what-we-do/projects/transparent-algorithms-insurance

**References:**

[1] A. Tversky, D. Kahneman: "Judgment under uncertainty: Heuristics and biases", Science 185.4157, 1974, 1124–1131. DOI: 10.1126/science.185.4157.1124

[2] B. Friedman, et al.: "Value sensitive design and information systems", Early engagement and new technologies: Opening up the laboratory. Springer, Dordrecht, 2013. 55–95. DOI: 10.1007/978-94-007-7844-3_4

[3] "The future of insurance: Counsel of protection", The Economist, March 11, 2017, 67–68.

**Please contact:**

Ulrik Franke
RISE ICT/SICS, Sweden
+46 72 549 92 64
ulrik.franke@ri.se

# INDICÆTING – Automatically Detecting, Extracting, and Correlating Cyber Threat Intelligence from Raw Computer Log Data

by Max Landauer and Florian Skopik (AIT Austrian Institute of Technology)

***"Cyber threat intelligence" is security-relevant information, often directly derived from cyber incidents that enables comprehensive protection against upcoming cyber-attacks. However, collecting and transforming the available low-level data into high-level threat intelligence is usually time-consuming and requires extensive manual work as well as in-depth domain knowledge. INDICÆTING supports this procedure by developing and applying machine learning algorithms that automatically detect anomalies in the monitored system behaviour, correlate affected events to generate multi-step attack models and aggregate them to generate usable threat intelligence.***

Today's omnipresence of computer systems and networks provides great benefit to the economy and society, but also opens the door to digital threats, such as data theft. Hackers and cyber-criminal groups carry out attacks by exploiting vulnerabilities of the deployed systems, with little or no time for the victims to react.

The growing interconnectedness of digital services helps to enable cyber-attacks. The Internet of Things and Industry 4.0 entail the emergence of highly complex system landscapes that offer numerous entry points for infiltration. As a simple measure of protection, most modern systems are equipped with blacklists containing indicators of compromise, such as IP addresses, known to correspond to adversarial entities. Detection systems monitor infrastructure states and outbound connections, compare the observed events with predefined signatures specified in the blacklists to raise alerts if certain thresholds are exceeded.

However, this basic approach suffers from a serious shortcoming: simple indicators of compromise (IoC) such as malicious IP addresses are highly volatile and only valid for a short period of time, since it is easy for attackers to circumvent their detection. Tactics, techniques and procedures (TTP) on the other hand are valid for a longer time, because it is difficult to change the modus operandi of attacks [1]. However, compiling threat intelligence on TTPs is difficult and requires manually analysis of complex attack patterns. All manual analyses are tedious and time-consuming, and since attacks are often carried out on a large-scale and affect multiple organisations, too much time passes until blacklists are updated with information on immanent attacks.

INDICÆTING aims to solve this problem by automatically generating complex "threat intelligence", i.e., more expressive than simple IoCs, but rather complex TTPs. INDICÆTING thereby pursues anomaly detection rather than blacklisting, i.e., instead of relying on an existing knowledge base, INDICÆTING makes use of self-learning algorithms that capture the normal system behaviour over time and detect deviations from the expected patterns [2]. This way, continuously generated low-level log data that documents almost all events occurring in the observed system is continuously monitored as soon as it is generated.

Log data contains semantically expressive parameters describing the current

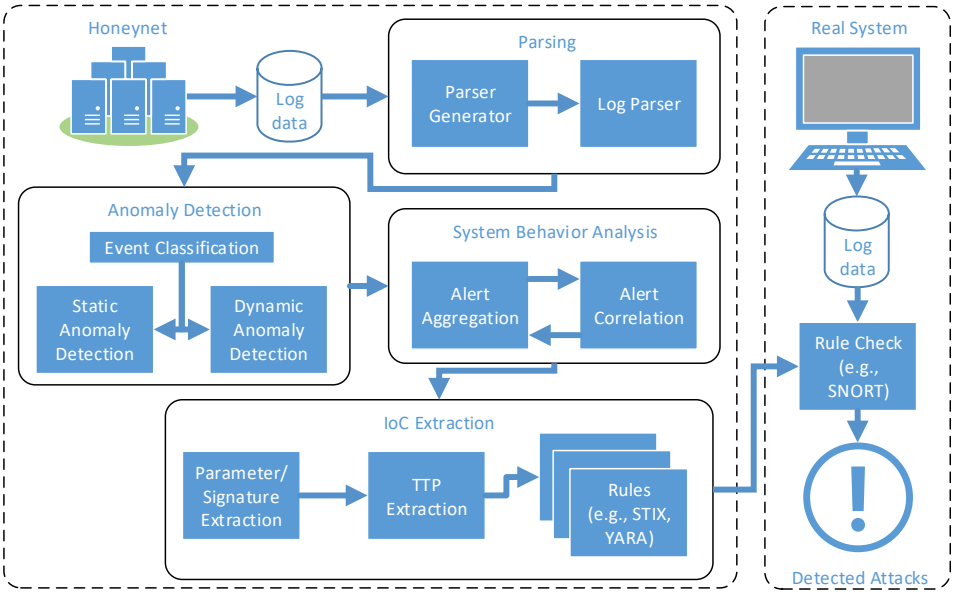



*Figure 1: Process for the automatic generation of Cyber Threat Intelligence.*

system state and is thus suitable for analysing the roots of system failures in hindsight after incidents occurred – a task that has been carried out by software engineers for decades. Only recently, have system logs been analysed in real-time in order to indicate system problems almost as they occur. However, this is a highly non-trivial activity: the main issue with processing system logs is that they are unstructured and different on every system, and it is challenging to automatically extract parameters and map individual log lines to more abstract event classes without human intervention. INDICÆTING achieves this by parsing the data, i.e., determining which parts of the log lines correspond to constant (textual) parts, and which correspond to parameters such as usernames, IDs, IP addresses, etc. As shown in Figure 1, a parser generator learns the structure of log data collected from a honeynet, i.e., a system specifically set up to attract attackers. Once the parsing model is established, INDICÆTING is able to retrieve parameter values and reason on their static distributions, discover dynamic dependencies between events and construct process models. Based on these values, anomalies are then detected by comparing each newly incoming log line with a corresponding model that was trained over a long time, i.e., a baseline model. Thereby, efficiency of the proposed algorithms is a key feature, because log data is typically produced in enormous amounts and fast rates.

Since anomalies reported on individual parameters or events are not sufficient for describing complex attack patterns, a subsequent step is required that analyses the overall system behaviour. For this purpose, the identified anomalies are correlated over multiple data channels and architectural layers in order to model multi-step attacks and derive abstract TTPs that affect several components in a narrow time window. For example, consider an employee entering login credentials after receiving a phishing email that contains a URL to a malicious website. Using the credentials, the attacker then infiltrates the network from a remote connection. This is a multi-step attack that can be detected by correlating URLs in mails, DNS, and web proxy logs. After appropriately modelling all involved steps and additional information on parameters, timing and context, the resulting high-level threat intelligence is suitable to be shared with other parties in cybersecurity communities [L1].

INDICÆTING is financially supported by the Austrian Research Promotion Agency (FFG) under grant number 868306. The project is carried out in course of an industry-related PhD thesis at the Austrian Institute of Technology (AIT) in cooperation with the Vienna University of Technology (TU WIEN). During the runtime of the project, AIT's Automatic Event Correlation for Incident Detection (AECID) [L2] tool will be further developed and used for evaluating the proposed concepts.

**References:**

[1] D. Chismon, M. Ruks: "Threat Intelligence: Collecting, Analysing, Evaluating", MWR InfoSecurity, 2015.
[2] V. Chandola, et al., "Anomaly Detection: A Survey", in ACM Comput. Surv., 2009.

**Links:**

[L1] http://misp-project.org/
[L2] https://aecid.ait.ac.at/

**Please contact:**

Max Landauer
AIT Austrian Institute of Technology, Austria
+43 664 88256012
max.landauer@ait.ac.at

# Why are Work Orders Scheduled too late? – A Practical Approach to Understand a Production Scheduler

by Markus Berg (proALPHA) and Sebastian Velten (Fraunhofer ITWM)

***In complex production environments, understanding the results of a scheduling algorithm is a challenging task. To avoid tardy work orders, proALPHA and Fraunhofer ITWM developed a component for identifying practical tardiness reasons and appropriate countermeasures.***

The proALPHA group is the third largest ERP provider for small and mid-sized manufacturing and trading companies in Germany, Austria, and Switzerland. For more than 25 years, proALPHA has offered a powerful ERP solution as well as consulting, support, training, and maintenance services from one source. The flexible and scalable ERP solution features a wide range of functions that allow all processes along the value-added chain to be controlled.

One of the modules available for the proALPHA ERP solution is Production. The goal of this module is the organization of the whole production process (from prefabrication to final products) of a company by providing decision support for production planners. To this end, work orders (consisting of structured process steps), calendars for production resources (workforces, machines, tools), material levels (with reorder times and known in- and outflows) and resource as well as material requirements can be managed. In addition, the module contains various production planning and detailed scheduling algorithms.

These algorithms can be used to execute a complete optimization of all work orders with respect to current resource and material availabilities. The considered problem is a (multi-) resource constrained multi-project scheduling problem (see [1]) with several additional features: resource and overload decisions, producer/consumer assignments and time bounds among others. The objectives are the minimization of tardiness, earliness and throughput time of the work orders. To model and solve the problem a Constraint Programming approach (see [2]) is used. For the implementation, the CP-library IBM ILOG CP Optimizer (see [3]) is employed.

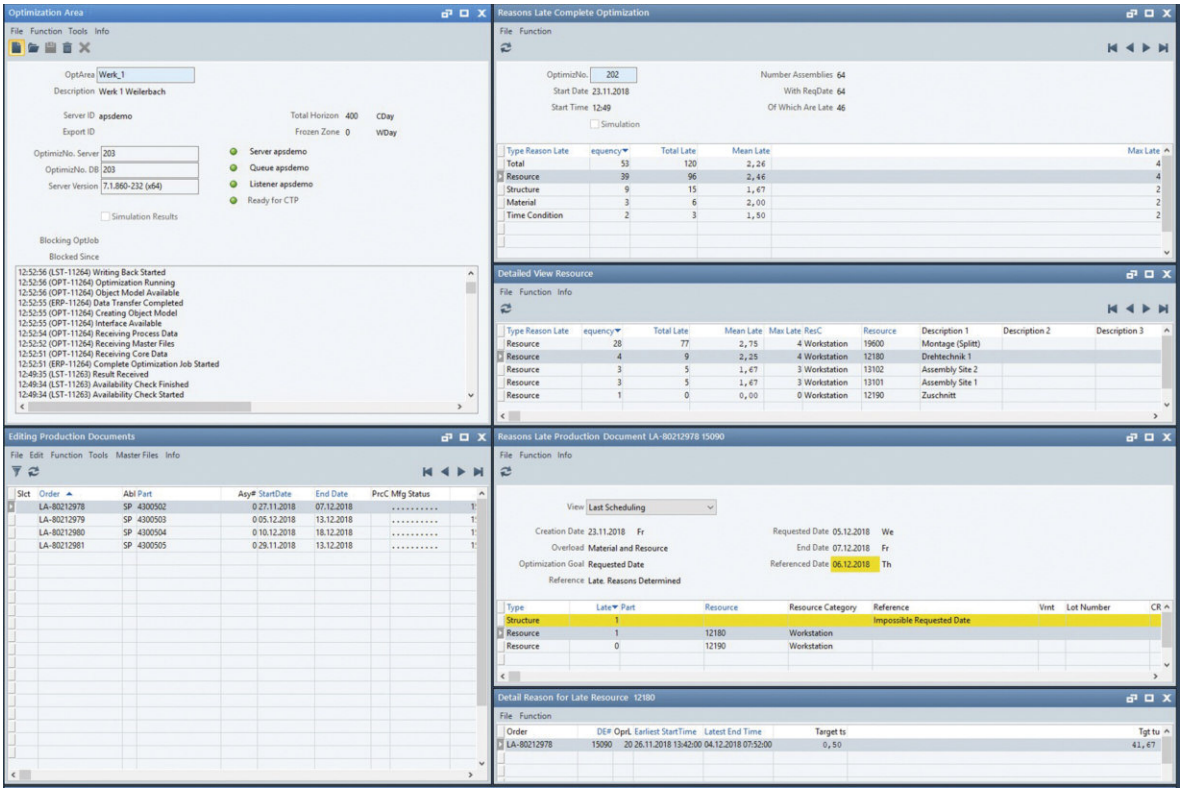



*Figure 1: Result of the tardiness-reasoning component for a single work order.*

Since the complete optimization takes limited resource and material capacities into account, tardy processes can typically not be avoided. In many cases, this is not acceptable, as tardy processes violate contracts or delivery promises made to important customers. Therefore, one of the main tasks of a production planner is to identify why a particular process is scheduled too late and what he or she can do to make a production in time possible. This task, however, is not easy as, for example, customer orders often require many production resources, raw materials and prefabrication steps.

To support the production planner in this critical task, the complete optimization has been extended by a tardiness-reasoning component, which provides tardiness reasons with respect to the optimized schedule. From a mathematical point of view, a tardiness reason of a process is a minimal set of constraints, which, together with the constraint that the process ends in time, is infeasible. However, the determination of such minimal infeasible sets is, in general, a hard problem and minimal sets alone are of limited use.

From a practical point of view, tardiness reasons are nearly always rather simple and can be divided in different reason-types:

- Structure: Processing times together with precedence constraints lead to a delay.
- Time Bound: A time bound (release dates) leads to a delay.
- Single Material: A missing material leads to a delay.
- Single Resource: A resource with missing capacity leads to a delay.
- Combined: A missing material together with a resource or two resources together lead to a delay.

Based on these reason-types, the tardiness-reasoning component works as follows. First, all potential reasons for a tardy process are identified (for example, each consumed material is a potential material reason). Then, for each potential reason it is tested, if it is indeed responsible for a delay. For each reason-type, special constraint based models are used for this test. Finally, if a reason is verified, further information is calculated that help the production planner to evaluate the importance of the reason and give hints on how the problem can be fixed. Examples are the delay caused by a reason, the amount of a material and the latest point in time when it is needed as well as the amount of free time that is missing on a resource (see Figure 1).

Tests with practical data show that far more than 90% of the delays can be explained, even if not all possible reasons are covered. The component helps production planners to identify tardiness reasons as well as bottlenecks (materials, resources). As a result, it leads to a higher acceptance of the schedules provided by the complete optimization.

**References:**

[1] S. Hartmann, D. Briskorn: "A survey of variants and extensions of the resource constrained project scheduling problem", European Journal of Operational Research, 207(1), 1-14, 2010.

[2] P. Baptiste, C. Le Pape, W. Nuijten: "Constraint-Based Scheduling – Applying Constraint Programming to Scheduling Problems", Kluwer, 2001.

[3] P. Laborie, J. Rogerie, P. Shaw, P. Vilím: "IBM ILOG CP optimizer for scheduling, 20+ years of scheduling with constraints at IBM/ILOG", Constraints, 23, 210 – 250, 2018.

**Please contact:**

Sebastian Velten
Fraunhofer ITWM, Germany
+49 (0)631 31600 4260
sebastian.velten@itwm.fhg.de

European Research and Innovation

# Using Augmented Reality for Radiological Incident Training

by Santiago Maraggi, Joan Baixauli and Roderick McCall (LIST)

***Training for radiological events is time consuming and risky. In contrast to real sources, a prototype augmented reality system lets trainees and trainers safely learn about the necessary detection, identification and decontamination steps.***

The European Union funded H2020 TARGET project uses mixed reality technologies alongside serious gaming approaches to train security critical agents such as the police, fire brigades and Chemical, Biological, Radiological and Nuclear (CBRN) teams. CBRN response team training is the focus of this article, with the particular system described here using augmented reality to train radiological incident teams. The main advantages of using augmented reality for CBRN training are that the trainees continue to interact with the real world but can see invisible risks e.g. radiation. As there are no real radiation sources it is safer and cheaper. The systems developed within the project were a direct result of an extensive requirements capture process [1].

The training case comprises a two person CBRN team entering a scene where they have to detect, find, classify and isolate simulated radiation sources. They also need avoid over exposure to the radiation sources and have to complete their tasks before their simulated oxygen level is depleted. Once they have completed this task they must enter the decontamination phase, which entails simulating the cleaning of their protective suits.

The system uses a range of technologies, for example the Microsoft Hololens augmented reality headset makes it possible to see augmented virtual holographic objects (holograms) in the environment integrated with the real objects perceived by the user, providing a mixed reality perception. The Pozyx indoor localisation system allows us to track movable objects in the scene. Any object can have a chosen simulated radiation isotope attached with an intensity Bq (Becquerels). A customised 3D printed simulated dosimeter (a replica of an "Identifinder" device) has been developed, it contains a Raspberry pi with a screen and a Pozyx tag to identify isotopes and measure the simulated radiation values.

Radiation contamination is modelled using a discrete particle system in Unity 3D. Particles are attached to contaminated objects; this allows the system to check for collisions between, for example, the Pozyx tag attached to the trainee and the contaminated object. Virtual colliders are set with all scene elements in order to detect collisions with contamination particles (see Figure 1). It is also possible for objects to become contaminated in specific areas and for contamination to spread between objects (see Figure 2).

Visualisation of the radiation field is provided on two levels (see Figure 3), firstly the red particles show the critical zone,

*Figure 1: Colliders Superimposed Over Real Objects.*

*Figure 2: Contaminated Laboratory Instruments.*

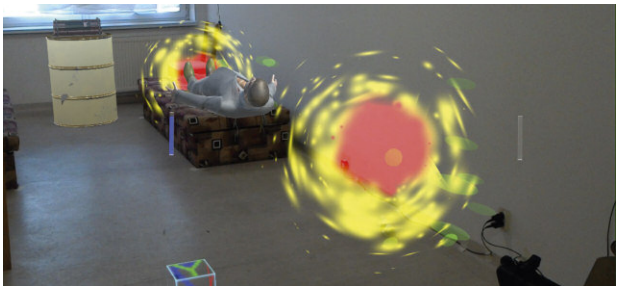


*Figure 3: Radiation Fields.*

while the yellow particles show the external dangerous zone. The extent of the visualisation elements depends on the dose level of the source. For the visualisation of the contamination particles a transparent green spherical model is used. Also, each time a contamination particle comes into contact with a contaminable object a sound is emitted. These visual and aural clues can be activated or deactivated depending on the preferences of the instructor.

All radiation contamination particles have a specific radiation dose value and also a minimum level. This allows for radiation dose levels to fall over time and also for certain levels of radiation to remain even if cleaning (decontamination) is undertaken.

The decontamination or cleaning process involves the application of a physical simulated cleaning brush to contaminated objects. This is similar to using a real brush to undertake the same task. A liquid jet is also simulated and it can be used to clean contaminated particles from a chosen object. Decontamination of the CBRN team members is also an important task. For this, the first trainee has to stand away from the main scene with their arms horizontal. A second trainee uses a simulated cleaning jet to clean the first trainees protective suit, the contamination level is indicated via several augmentations. Decontamination is complete when all the contamination zones have turned green.

To enhance the learning and training experience, some visualisation elements were added. During the exercise, an oxygen level bar with a certain amount of time appears at the left side of the augmented reality display and an accumula-

*Figure 4: CBRN Team Member Wearing a Protective Suit.*

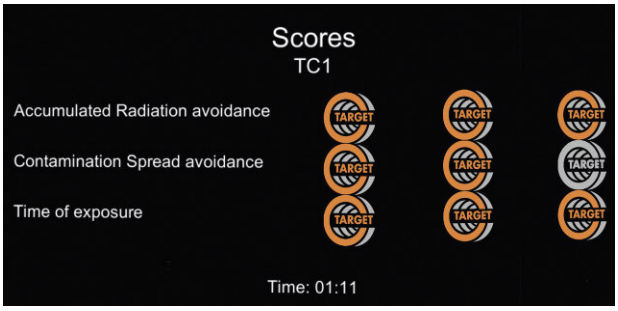



*Figure 5: Gamification Display.*

tive radiation exposure bar is shown on the right and becomes red when the cumulative dose reaches a dangerous level. Several gamification parameters are displayed by the system and these include: Accumulated Radiation Avoidance, Contamination Spread Avoidance and Time of Exposure (Figure 5).

In order to validate the work, different versions of the system have been tested during two trials in Bratislava, Slovakia. The trials identified that the approach is useful for CBRN teams and also pointed to several areas of improvement.

This work has been carried out within the framework of the EU H2020 TARGET project. The project has received funding from the European Union's Horizon research and innovation programme under grant agreement No 653350. We also acknowledge the other partners in the project for their assistance.

**Reference:**

[1] J.L.Huynen et al. "Towards Design Recommendations for Training of Security Critical Agents in Mixed Reality Environments" in BCS HCI 2018.

**Links:**

http://www.target-h2020.eu/
https://www.pozyx.io/
https://www.microsoft.com/en-us/hololens
https://unity3d.com/
https://www.flir.com/products/identifinder-r400/

**Please contact:**

Rod McCall, Environmental Research and Innovation Department (ERIN), Luxembourg Institute of Science and Technology (LIST)
roderick.mccall@list.lu

# Building upon Modularity in Artificial Neural Networks

by Zoltán Fazekas, Gábor Balázs, and Péter Gáspár (MTA SZTAKI)

***In a pilot-study, an urban road environment detection function was considered for smart cars, as well as for self-driving cars. The implemented artificial neural network (ANN) based algorithms use the traffic sign (TS) and/or crossroad (CR) occurrences, along a route, as input. The TS-based and the CR-based classifiers were then merged into a compound one. The way this was accomplished serves as a simple, practical example of how to build upon modularity and how to retain some degree of it in functioning ANNs.***

Modularity is a desired property of built systems; it serves a range of engineering demands throughout the system life-cycle (e.g., traceability, reusability). With respect to artificial neural networks (ANN), the latter property pertains to the network configuration. Modularity, however, comes at a price (e.g., additional layers, lower precision) [1]. In a pilot-study run by the Institute for Computer Science and Control (MTA SZTAKI), Budapest, Hungary, the urban socio-economic road environment detection (RoED) task was considered. The intention was to rely on highly processed, low-volume data as input that are available in smart cars. After the data collection car trips (e.g., in Csepel, see Figure 2), methods were sought that can infer the environment type along a route. The road environment (RE) information could be leveraged both in smart cars and in self-driving cars. A statistical change detection method – applied to traffic sign (TS) occurrences – was presented in [3], and an ANN-based one in [4]. Due to the input choice, the availability of an on-board TS recognition system is essential for any practical utilisation.

The usefulness of the RoED function was confirmed by a driving simulation study [2]. It examined the effect of driving experience on drivers' adaptation to changing RE complexity in urban areas. Three complexity levels – corresponding to the RE types considered herein – were used. The drivers were grouped into three groups according to their experience. The most experienced drivers adapted better to increasingly complex REs than those in the other two groups. Thus, for less experienced drivers, the RoED function would be beneficial indeed; in self-driving cars, the RE information could be drawn upon to choose vehicle speed and acceleration.

In the pilot-study, ANNs exhibiting different levels of modularity were also looked at, and their detection performances were compared. The implemented ANN-based RoED classifier – that relies on recognised TS and crossroad (CR) occurrences – constitutes a simple practical example on how to build upon modularity in ANNs (e.g., by setting up a modular initial ANN), and how to retain some degree of it (e.g., in form of separable subnetworks, or subnetworks with limited interconnections, see Figure 1) even if some performance improvement seems necessary.

The urban RE surrounding an ego-car is classified into one of the three RE categories, namely into downtown, industrial/commercial, and residential areas. The classification is based on TS and CR data (namely, on the types and the along-the-route locations of TSs and CRs). The classification is carried out by an ANN devised for the purpose; this is referred to as a full ANN, as it is merged from a TS-based and a CR-based classifier with the help of a merging module.

An Android application was used during data collection trips for the manual recording of TS data and the RE types. It automatically records the car-trajectory. Eight TS types were identified as occurring frequently in urban areas and prevailing in one of the three REs. A shallow ANN was devised and trained – using back-propagation – to identify the actual RE from TS data [4]. The input features used by the ANN were the average distances between consecutive relevant TSs over the last 250, 500, 1000 and 2000 meters, and the number of occurrences of the typical TSs pertaining to each of the considered three REs in similar fashion. These features provide detailed information of the spatial frequencies of the TSs and CRs. The RoED step is repeated for each 50 m route-segment.

The CR data was added in a post-trip manner to the trajectory data from a public geographical information system. Five

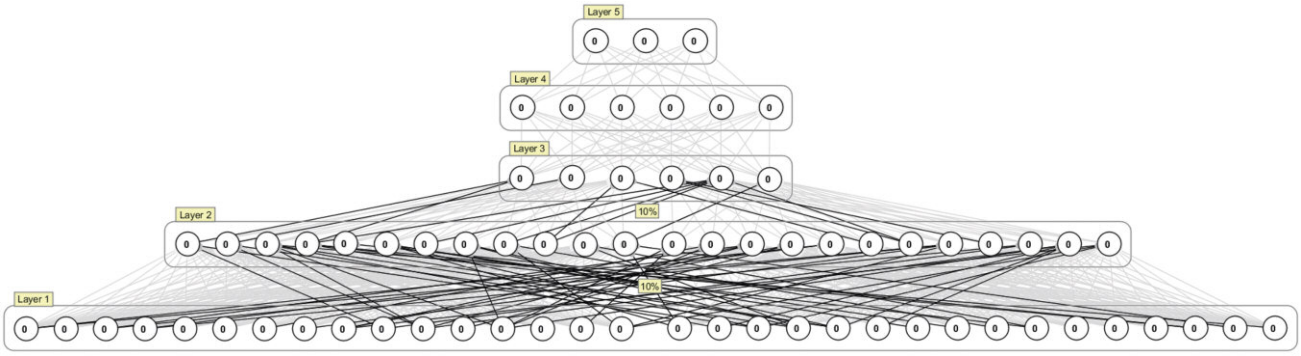



*Figure 1: A full non-modular ANN with a random 10% selection of inter-modular synapses, i.e., the synapses of the modular TS and CR processing network (light grey lines) are augmented with those between the TS and CR processing subnetworks (black lines). The ANN simulations and the diagram was made with the Simbrain 3.0. Neural Networks framework.*

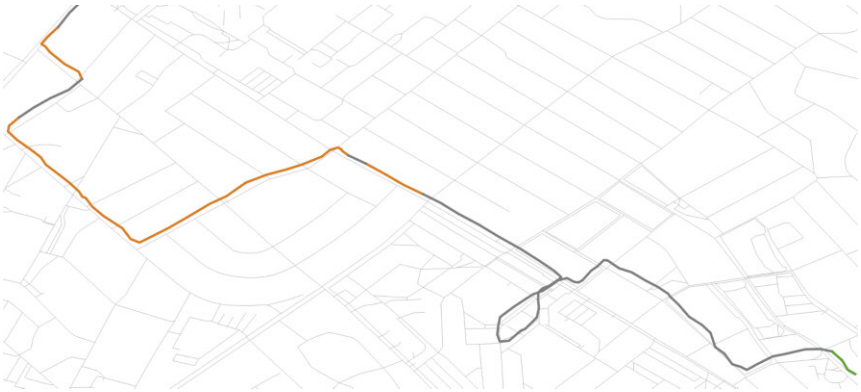


*Figure 2: Road environment types – namely, downtown (orange line), industrial/commercial (dark grey line), and residential (green line) environments – inferred by a modular full ANN along a route in Csepel, Hungary. (Map data: OpenStreetView, map-editor: QGIS 2.8.1.)*

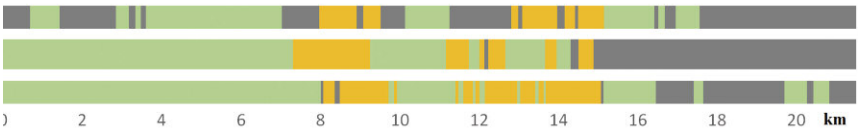



*Figure 3: The road environment types manually recorded along a test-route in Csepel, Hungary (with the path-lengths shown below the stripes) serving as ground truth (middle), and those inferred by a modular full ANN (top) and by a non-modular one (bottom). The latter ANN comprises a random 40% selection of the possible inter-modular synapses. The colour scheme is the same as in Figure 2.*

CR categories were considered. An ANN-based classifier was devised and trained also for the CR data and was re-applied as a functional module/subnetwork. The CR input features were similar to the TS ones. The TS and the CR modules were then merged with a merging module. The resulting modular full ANN appears in Figure 1 as the ANN with light grey synapses.

In some of the experiments, the TS and CR subnetworks were kept separate, while in others increasing percentages of the possible synapses between these subnetworks were allowed (Figure 1). It was hypothesised that the separable ANNs can achieve a reasonable RoED performance, and that the non-modular ones – due to their more intricate data interactions – can achieve even better.

Several training regimes were devised for the merged ANN using the analogy of mechanical grids. These regimes when applied to the full ANNs retain the weights and biases in certain (stiff) parts of the network, while the other (loose) parts can evolve more freely. The good starting values for the weights and biases – inherited from the TS and CR classifiers – shorten the necessary training effort.

After training, the ANN-based classifier processing TS data could achieve a 67.3% agreement with the ground truth, while the CR-processing classifier, on its own, achieved only a 59.7% agreement. As expected, merging these two resulted in higher values: a reasonably good agreement (71.9%) was achieved for a modular full ANN, and an even better (74.1%) for a full non-modular ANN with a random 40% selection of inter-modular synapses. In Figure 3, the RoED results can be visually compared for the modular and the non-modular full ANN.

In our view, the RoED function could be turned into a useful automotive function – both in the context of smart cars and self-driving cars – in the coming years, particularly, if the function is further developed and meticulously tested in various urban environments in different countries.

**Link:**
[L1] https://www.sztaki.hu/en/science/projects/lab-autonomous-vehicles

**References:**
[1] M. Woźniak, M. Graña, E. Corchado: „A survey of multiple classifier systems as hybrid systems", Information Fusion, 16, 3-17, 2014.
[2] C. M. Rudin-Brown, J. Edquist, M. G. Lenné: „Effects of driving experience and sensation-seeking on drivers' adaptation to road environment complexity", Safety Science, 62, 121-129, 2014.
[3] Z. Fazekas, G. Balázs, L. Gerencsér, P. Gáspár: "Inferring the actual urban road environment from traffic sign data using a minimum description length approach", Transportation Research Procedia, 27, 516-523, 2017.
[4] Z. Fazekas, G. Balázs, P. Gáspár: "Identifying the urban road environment type from traffic sign data using an artificial neural network", in Proc. of the Int. Scientific Conference Modern Safety Technologies in Transportation, Kosice, 42-49, 2017.

**Please contact:**
Zoltán Fazekas, MTA SZTAKI, Hungary
+36 1 2796163
zoltan.fazekas@sztaki.mta.hu

# BBTalk: An Online Service for Collaborative and Transparent Thesaurus Curation

by Christos Georgis, George Bruseker and Eleni Tsouloucha (ICS-FORTH)

***BBTalk is an online service designed to support collaborative interdisciplinary development and extension of thesauri. At present, it serves to support the curation of the BackBone Thesaurus (BBT), a meta-thesaurus for the humanities. This service allows for the transparent, community development of the BackBone Thesaurus by enabling users to submit suggestions for changes and additions to the terminology, as well as link specialist thesauri to the meta-thesaurus terms, while enabling the thesauri curators to jointly edit, add and delete terminology. This model of cooperative editing is linked to an online discussion system that allows thesauri curators to confer with one another, exchange views and ideas and finally determine any necessary changes to the BBT.***

The BBTalk service [L1] addresses the need for a means to cooperatively and transparently build and curate a meta-thesaurus for the humanities, the BackBone Thesaurus (BBT) [L2]. The BBT is the research outcome of work undertaken by the Thesaurus Maintenance Working Group (TMWG) [L3] -established in the framework of DARIAH-EU. The BBT offers top-level-concepts (facets and hierarchies) that should serve as common ground for generic thesaurus building. The technique adopted for the BBT of faceted classification is, moreover, considered valid and consistent from a cross-disciplinary perspective. A major advantage of this approach is the potential expansion of the BBT into new scientific domains in a sustainable and manageable fashion. The BBT explicitly does not require domain experts to abandon their specialised terminology in the name of some universally accepted terminology. Rather, the core feature of the BBT is that it promotes alignment of cutting-edge terminology to well-formed general terms of the meta-thesaurus. This enables both collaboration and cross-disciplinary resource discovery while ensuring compatibility with thesauri that cover highly specific and developing areas of knowledge. In order to support this vision of a collaborative and gradually expanded top level meta-thesaurus, it is necessary to have the tools to curate and evolve the standard, adjusting it in an objective and cooperative manner to accommodate new areas of research and thought.

To meet this need, BBTalk has been developed by ICS FORTH [L4] and is offered as an online service that enables the basic functions necessary to support the open development of the BBT. These functions include, inter alia, a) proposing new terms and changes to the meta-thesaurus itself and b) connecting specialist thesauri to the federated system.

Key to the BBTalk service is the offer of a built-in communication system supporting discussions between the different user groups of the BBT regarding submissions for changes of the BBT and/or any connections proposed. Such proposed actions are then considered and managed in a formal but open editorial process. Moreover, BBTalk keeps a record of the different versions of the BBT and the history of the submissions in the form of discussions relating to the evolution of the meta-thesaurus. This ensures the level of transparency which is the conditio sine qua non for a meta-thesaurus to serve as central resource for academic research, where decisions must be reached openly and on principle, not by fiat. The ability to communicate in a structured way in relation to

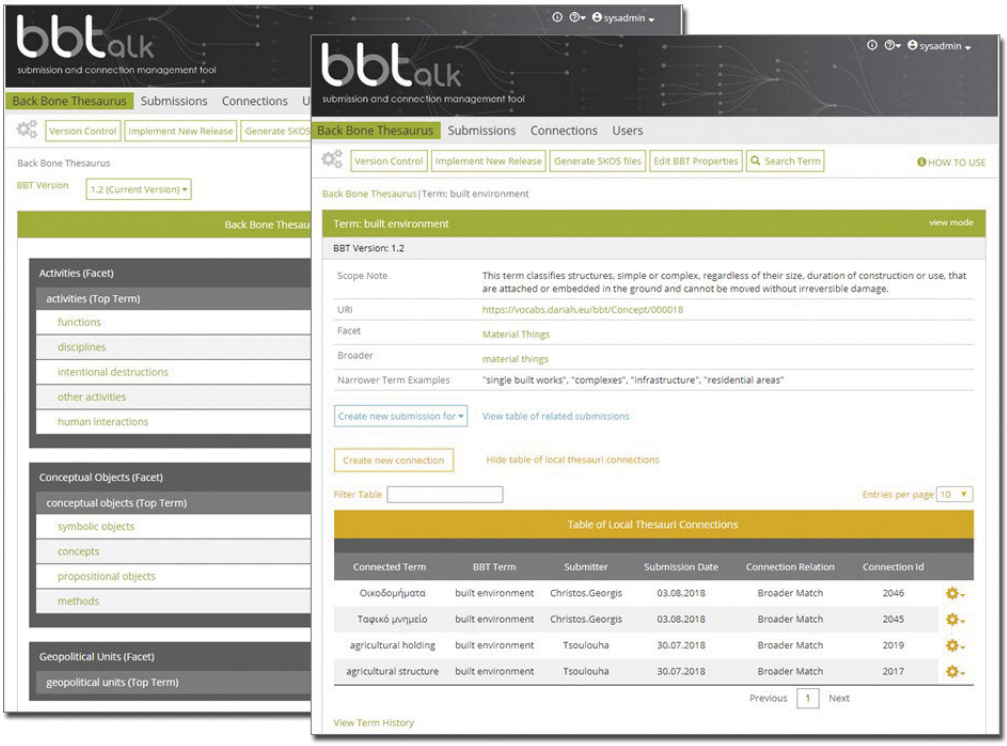



*Figure 1: BBT overview interface.*

the evolution of the thesaurus, and the ability to trace back the decision process allows for a reasoned and criticizable evolution of the thesaurus.

With regard to the functionality for connecting specialist thesauri to the overall BBT schema, BBTalk functions as an alignment tool. Through this alignment functionality, specialist thesauri maintainers are able to align their thesaurus to a high-level and well-structured thesaurus. This alignment process creates a critical feedback relation between the specialist thesauri providers and the BBT maintainers. On the one hand, the proposed links offer to the BBT curatorial team use cases that either corroborate or provide disconfirmations to the top level classifications. On the other hand, specialist thesauri maintainers can test their terms against the BBT classifications to see whether they follow generically robust higher level classifications. The alignments performed in BBTalk can be exported as RDF and then used by researchers in their data transparently.

Taken together the functionalities of BBTalk allow it to work as a general maintenance system for the BBT, supporting the implementation and documentation of proposed changes and, by storing, the connections of specialist thesauri terms to the BBT together with the contact information of their maintainers, enabling the automated notification of changes to BBTalk users. The entire process is supported by a user-friendly browsing facility that allows users of BBTalk to view the connected terms of specialist thesauri that are linked to the meta-thesaurus and the type of relation they stand in, with respect to BBT's relevant terms, i.e. an exact or a broader match (Figure 1).

The BBTalk tool is currently used by a DARIAH-supported community of scholars, who have undertaken the task of curating the BBT. Members of this community include all members of the DARIAH TMWG; FORTH-ICS, AA [L5], DAI [L6] and FRANTIQ [L7]. As more institutions join, we envisage that the federation of more and more thesauri under a common banner will generate a marketplace of extant thesauri that can be reused by researchers working in the same discipline, thus supporting comparable research and reducing the need for harmonisation and alignment of datasets.

**Links:**

[L1] http://www.backbonethesaurus.eu/BBTalk
[L2] http://www.backbonethesaurus.eu/
[L3] https://www.dariah.eu/activities/working-groups/thesaurus-maintenance/
[L4] http://www.ics.forth.gr
[L5] www.academyofathens.gr
[L6] www.dainst.org
[L7] https://www.frantiq.fr

**Reference:**

[1] M. Daskalaki, L. Charami: "A backbone Thesaurus for Digital Humanities", in ERCIM News, Issue 111, October 2017, Special theme: "Digital Humanities", p. 18.

**Please contact:**

Christos Georgis, ICS-FORTH, Greece
georgis@ics.forth.gr

# Understandable Deep Neural Networks for Predictive Maintenance in the Manufacturing Industry

by Anahid N.Jalali, Alexander Schindler and Bernhard Haslhofer (Austrian Institute of Technology)

***Data driven prognostic systems enable us to send out an early warning of machine failure in order to reduce the cost of failures and maintenance and to improve the management of the maintenance schedule. For this purpose, robust prognostic algorithms such as deep neural networks are used whose put is often difficult to interpret and comprehend. We investigate these models with the aim of moving towards a transparent and understandable model which can be applied on critical applications such as within the manufacturing industry.***

With the development of manufacturing technology and mass production, the methodology of maintenance scheduling and management has become an important topic in industry. Predictive maintenance (PdM) is the recent maintenance management approach after run to failure (R2F), which is applied when observing the failure in the system and preventive maintenance (PvM) that is scheduled based on average life time of the machine. PdM denotes a set of processes that aim to reduce maintenance costs in the manufacturing industry. For this purpose, prognostic systems are used to forecast metrics such as remaining useful life (RUL) and time to failure (TTF), which are often grouped into two data driven and model-based methods.

Data driven approaches are used when no or very little understanding of the physics behind the system operation exists. In order to detect changes within the system, these approaches usually employ techniques such as pattern recognition and machine learning, which compared with model-based methods, are easier to obtain and implement. These novel methods require robust predictive models to capture the health status of a machine by using information extracted from collected sensor data. At the moment models are built manually by feature engineering, which relies extensively on domain knowledge. The effectiveness of data-driven predictive maintenance methods, which predict the health state of some machinery, depends heavily on health indicators (HI), which are quantitative indicators (features) that are extracted from historical sensor data. This relies on the assumption that statistical characteristics of data are relatively consistent unless a fault occurs. A selection of relevant features is then fed into the model to compute a one/multi-dimensional indicator, which describes the degradation of the machine's health to eventually estimate the remaining lifetime of the machine.

Deep neural networks have shown superior performance on a variety of applications such as image and audio classification and speech and handwriting recognition. Similar to other

applications, data assembled for predictive maintenance are sensor parameters that are collected over time. Utilising deep models could reduce manual feature engineering effort and automatically construct relevant factors and the health factors that indicate the health state of the machine and its estimated remaining runtime before the next upcoming downtime.

However, despite the promising features of deep neural networks, their complex architecture results in a lack of transparency and it is very complicated to interpret and explain their outcome, which is a severe issue that currently prevents their adoption in the critical applications and manufacturing domain. Explainable artificial intelligence (XAI) addresses this problem and refers to AI models yielding behaviours and predictions that can be understood by humans. The general goal of XAI research is to be able to understand the behaviour of the model by clarifying under what conditions a machine learning model (i) yields a specific outcome, (ii) succeeds or fails, (iii) yields errors, and (iv) can be trusted.

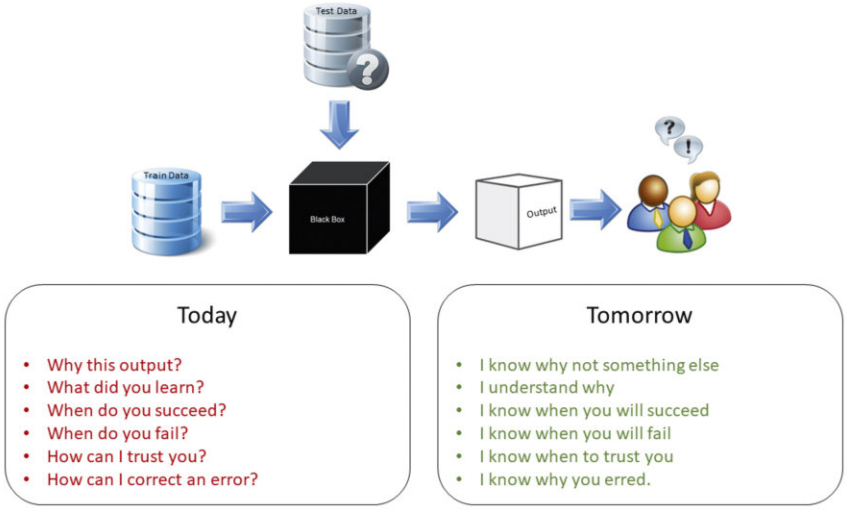



*Figure 1: The machine learning process.*

Therefore, explainable AI systems are expected to provide comprehensible explanations of their decisions when interacting with their users. This can also be considered as an important prerequisite for collaborative intelligence, which denotes a fully accepted integration of AI into society. Early work on machine learning model explanation often focused on visualising model predictions using a common visualisation technique called nomograms, which was first applied to logistic regression models. Later, this technique was used to interpret SVMs and Naive Bayes models. Recently, these visualisation techniques have been used on deep learning algorithms to visualise the output layers of deep architecture such as CNNs and also on RNNs. Besides producing visualisation for model predictions, existing studies investigated two other approaches for explainability of machine learning (ML) algorithms: prediction interpretation and justification as well as interpretable models.

In prediction interpretation and justification, a non-interpretable complex model and prediction are given to produce a justification, which is often done by isolating contributions of individual features of a prediction. This technique is proposed for models such as Multilayer Perceptron, probabilistic radial basis functions and SVM classifier where it was used to extract conjunctive rules from a subset of features. In 2008, a method was proposed that investigates alternative predictions of an unknown classifier in cases where a particular feature was absent and measured the effect of this individual feature. Effects were then visualised in order to explain the main contributions to a prediction and to compare the effect of that feature in various models. Later on in 2016, to interpret deep model's predictions, a long-short term-memory (LSTM) with a loss function that encourages class discriminative information as a caption generation model was used to justify classification results of a CNN model.

The goal of our research is to first provide a methodology that can automatically construct machine health indicators in order improve the effectiveness of prediction models using deep neural networks. Second, it will investigate optimised deep neural networks to predict maintenance measures such as remaining useful life (RUL) and time-to-failure (TTF). Third, it will extend the field of explainable AI by investigating visualisation and interpretation models to justify and explain the outcome of a predictive model. For that task, the characteristics of deep learning architectures such as convolutional neural network, recurrent neural network and deep belief network will be investigated.

First the machine learning algorithm is trained with a training dataset. The trained model is then evaluated using a test/evaluation dataset. However, since the internal behaviour of the algorithm is not clear, the model's output is not interpretable. The goal is to be able to understand a model's success and failure, and to understand its decisions.

**References:**

[1] D. Gunning: "Explainable artificial intelligence (xai)", Defense Advanced Research Projects Agency (DARPA), nd Web (2017).

[2] M. Robnik-Šikonja, et al.: "Efficiently explaining decisions of probabilistic RBF classification networks." International Conference on Adaptive and Natural Computing Algorithms. Springer, Berlin, Heidelberg, 2011.

[3] D. Hendricks. "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description". The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[4] Epstein, Susan L. "Wanted: collaborative intelligence." Artificial Intelligence 221 (2015): 36-45.

**Please contact:**

Anahid N.Jalali, Alexander Schindler and Bernhard Haslhofer, Austrian Institute of Technology GmbH
anahid.jalali@ait.ac.at
alexander.schindler@ait.ac.at
bernhard.haslhofer@ait.ac.at

# Is My Definition the Same as Yours?

by Gerhard Chroust (Johannes Kepler University Linz)
and Georg Neubauer (Austrian Institute of Technology)

***Effective and efficient communication and cooperation needs a semantically precise terminology, especially in disaster management, owing to the inherent urgency, time pressure, stress and often cultural differences of interventions. The European project Driver+ aims to measure the similarities between different countries' terminologies surrounding disaster management. Each definition is characterised by a set of "descriptors" selected from a predefined anthology (the "bag-of-words"). The number of identical/different descriptors serves as a measure of the semantic similarity/difference of individual definitions and is translated into a numeric "degree of similarity". The translation considers logical and intuitive aspects. Human judgment and mechanical derivation in the process are clearly separated and identified. By exchanging the ontology this method will also be applicable to other domains.***

Clear, unambiguous and semantically precise terminology is one of the keys to effective and efficient communication, cooperation and decision making, especially in view of automation and computer support. This is highly relevant for disaster management because urgency, time pressure, stress and often cultural differences create additional complications.

The European project Driver+ (FP7-Project 607798: "DRiving InnoVation in crisis management for European Resilience", May 2014 – April 2020) [L1] aims at "a shared understanding in Crisis Management across Europe". Its objective is "to cope with current and future challenges due to increasingly severe consequences of natural disasters and terrorist threats, by the development and uptake of innovative solutions that are addressing the operational needs of practitioners dealing with Crisis Management".

The DRIVER+ consortium brings together dedicated multi-national practitioners, relief agencies, policy makers, technology suppliers and researchers, representing 14 countries. One of the partners is the Austrian Institute of Technology (AIT). It has ten years' experience in the development and validation of technological solutions for crisis prevention, interoperable systems, and for management of volunteers, together with theoretical and practical research with a focus on the software domain. It is deeply involved in the research on the similarity problem of definitions.

An essential step to an unambiguous terminology is understanding and measuring the similarity of existing definitions (Cregan, 2005) appearing in different documents such as standards, norms and instruction manuals. The measuring process should be easily understood and applied. The result should be compatible with the intuitive notion of similarity and should also fulfil the precision needs of practitioners in disaster management.

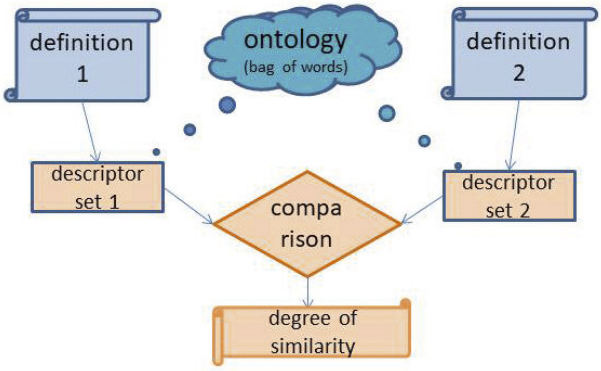



*Figure 1: Basic comparison process.*

It is important to note that our notion of similarity differs from that of literary texts. Definitions require (as far as possible) precise description and/or identification of objects or processes, whilst literary texts evaluate similarity with respect to aesthetics, type of expressions, contents and readability.

Research offers several different methods for measuring the semantic similarity of definitions (Slimani, 2013): For our part of the project the "bag of words" method has been chosen (see Figure 1): A chosen set of parameters is applied to each definition (e.g. size of the disaster, type of actors). For each parameter a set of descriptive concepts is defined (the "bag-of-words"). The similarity between two definitions is computed from the number of concepts which occur in both definitions compared to the number of concepts that only occur in one definition.

As simple as this scheme looks, the difficulties – as usual – lie in the details, some of which include:

- Which concepts are to be included in the "bag of words",
- which parameters are chosen to represent different aspects of a definition,
- how to express and consider synonyms, hierarchies of concepts, and sematic dependencies between concepts,
- how to convert the agreements/differences of concepts into numerical values (Lin, 1998).

The appeal of this method, however, is that, once established, it lends itself to application in other domains: It is "only" necessary to exchange the parameters and the associated bags-of- words.

**Link:** [L1] www.driver-project.eu/

**References:**

[1] A. M. Cregan: "Towards a science of definition", in Proc. of the AOW '05, vol. 58, p. 25–32, Australian Computer Society, Inc., 2005

[2] D. Lin,: "An information-theoretic definition of similarity", in Proc. of ICML 98, p. 296–304, 1998.

[3] T. Slimani: "Description and evaluation of semantic similarity measures approaches", Intl. J. of Computer Applications 80(10), October 2013, p 25–33, 2013.

**Please contact:**

Georg Neubauer
Austrian Institute of Technology GmbH
+43 50 5500 2807, Georg.Neuauer@ait.ac.at

# Science2Society Project Unveils the Effective Use of Big Research Data Transfer

by Ricard Munné Caldés (ATOS)

***The Science2Society project improves collaboration between science and industry by leveraging big research data through sustainable business models.***

The Science2Society project (March 2016 – February 2019) creates pilots and shares good practices, guidelines and training materials that improve awareness and practical performance in seven concrete university-industry-society interfacing schemes that are impacted by new practices derived from Science 2.0 and open innovation approaches and strategies. One of these schemes, led by the Aalto University and involving Atos Spain, Virtual Vehicle and the Joint Institute for Innovation Policy (JIIP) as main partners, is assessing the collaboration between research and industry through big data and science 2.0.

Big data can provide new economic, scientific and social value and new information can be extracted using big data technologies with the integration of additional datasets. The open data initiative together with the potential of big data are pushing many organisations from the government, industry and academia to open their data, so third parties can benefit from the analysis of this existing information. While the benefits of data openness are clear for organisations, the incentives for individual researchers to do so are not so obvious, although this is key to achieving the full potential of open science.

On one hand, this pilot aims to identify the obstacles preventing individual researchers from sharing their big research data and how they may be motivated to share them, and on the other hand, what needs to be done to enable industry to easily use these data. The overarching objective is then to identify what type of sustainable business models can best support both these issues: (i) the sharing of data and (ii) how to enable industry to take advantage of this data.

To this end, we conducted a literature review and examined two real-life cases where big research data are available. The first was a Finnish innovation database, which we used to investigate the challenges of opening the big data. Through a co-creation process based on interviews with database owners, potential users and open science experts, a proposal for opening the database was designed and validated, which resulted in a selection of different solutions as sustainable business cases. The second was the GCAT project, which is a biomedical research initiative with an open database with genetic information, environmental factors, medical records and biological samples from volunteers. In this case, through stakeholder interviews we learnt from the experience of an open big research database and extracted the specific business model underneath, deriving the best practices and lessons learnt.

The literature review revealed that the main obstacles to data sharing by researchers include: lack of perceived benefits, the effort required, and the risks of sharing data. Existing collaborative models in big data ecosystems must be promoted, paving the way for researchers to open their data.

The conclusions from the two case studies were synthetised in two frameworks supporting the development of sustainable business cases. The framework for the Finnish invention database case starts with identifying the benefits of opening the database, the identification of potential users' specific requirements, willingness to pay, and the potential user base. Following, it is important to investigate if there are any potential legal, functional or technical barriers that pose any constraints on data sharing. Finally, a business model consideration step is important to guarantee the sustainability of the process. For the GCAT use case, the framework is focused on

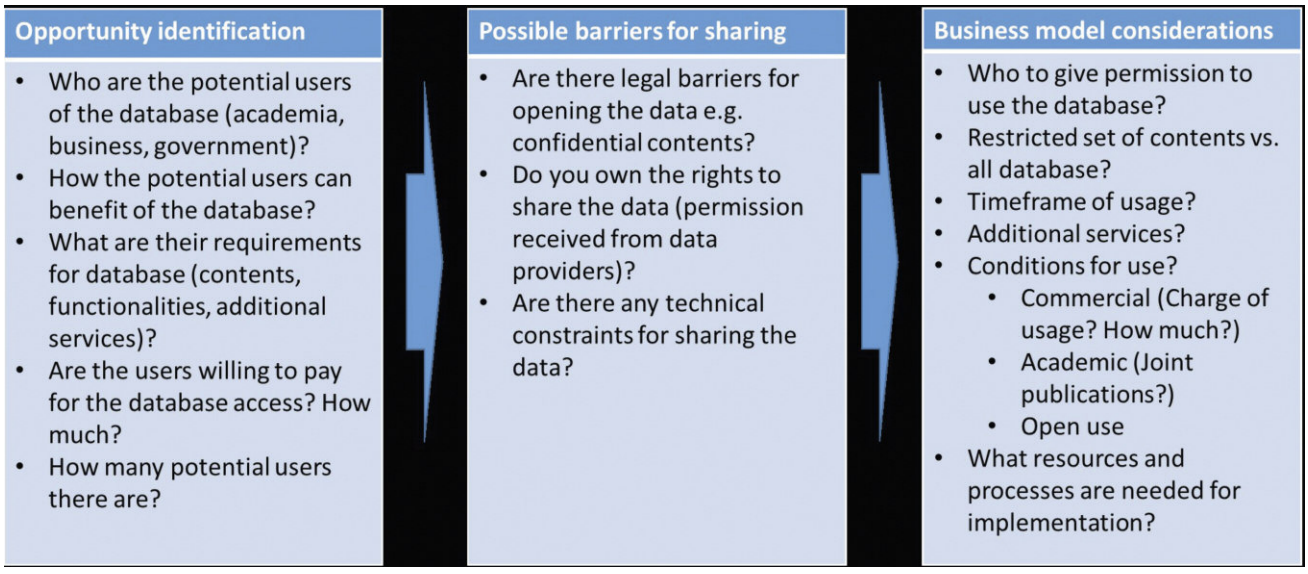



*Figure 1: Framework for opening big research database.*

the management of an existing open database. A successful collaboration can be achieved by minimising the data access fees, attracting the best researchers while allocating enough resources to ensure effective collaboration through data sharing. Privacy is an important factor and it should be reinforced through NDA agreements as part of the standardised informed consent protocols.

As final recommendations for the management of open big research data, researchers should develop a research data management plan prior to the start of the project. The decision to open the data should be based on a careful assessment of the underlying opportunities, barriers and alternative business models. We provide a generic process model for performing that assessment. Based on the GCAT case, lessons learnt and best practices for how to operate a big research database are provided. The key for a sustainable business model is to focus on developing a valuable proposition for data owners and users that benefits both parties.

For more information please refer to the Science2Society Knowledge Database, which contains case studies, methods and tools related to this pilot and the other pilots from the project. Additionally, the document "D3.2 Report on the implementation and evaluation of the UIS interface scheme pilots", will be available soon in the downloads section of the project website.

The knowledge and studies on this collaborative schema and others will be maintained after the project ends with the setting of the Learning and Implementation Alliance.

**Links:**
[L1] http://www.science2society.eu/
[L2] http://www.science2society.eu/kd-front
[L3] http://www.gcatbiobank.org/en_index/

**Please contact:**
Ricard Munné Caldés, Atos, Spain
+34 935485741, ricard.munne@atos.net

# Informed Machine Learning for Industry

by Christian Bauckhage, Daniel Schulz and Dirk Hecker (Fraunhofer IAIS)

***Deep neural networks have pushed the boundaries of artificial intelligence but their training requires vast amounts of data and high performance hardware. While truly digitised companies easily cope with these prerequisites, traditional industries still often lack the kind of data or infrastructures the current generation of end-to-end machine learning depends on. The Fraunhofer Center for Machine Learning therefore develops novel solutions which are informed by expert knowledge. These typically require less training data and are more transparent in their decision-making processes.***

Big data based machine learning (ML) plays a key role in the recent success of artificial intelligence (AI). In particular, deep neural networks trained with vast amounts of data and high performance hardware can now solve demanding cognitive tasks in computer vision, speech recognition, text understanding, or planning and decision-making. Nevertheless, practitioners in industries other than IT are increasingly skeptical of deep learning, mainly because of:

- Lack of training data. Vapnik-Chervonenkis (VC) theory establishes that supervised training of complex ML systems requires substantial amounts of representative data in order to learn reliably. Since details depend on a system's VC dimension, which is usually hard to come by, Widrow's rule of thumb is to train with at least ten times more data than there are system parameters. Modern deep networks with their millions of adjustable parameters thus need many more millions of examples in order to learn well. However, large amounts of annotated data are rarely available in industries that are not yet fully digitised. Even in contexts such as the internet of things or industry 4.0 where data accumulate quickly, we still often face thin data scenarios where labeled data for end-to-end machine learning are inaccessible.
- Lack of traceability. Trained connectionist architectures are black boxes whose inner computations are abstracted away from conceptual information processing. As their decision-making processes may thus be unaccountable, data scientists in industries where regulatory guidelines demand automated decision making to be comprehensible are wary of deep learning. Indeed, recent research shows that silly mistakes made by deep networks might be avoidable if they had "common sense". Even more alarmingly, recent research also shows that silly mistakes can be provoked using adversarial inputs.

The newly established Fraunhofer Center for Machine Learning addresses both these issues and researches methods for informed machine learning that, on the one hand, can cope with thin data and, on the other hand, lead to more explainable AI for industrial applications.

A basic observation is that domain experts in industry typically know a lot about the processes and data they are dealing

with and that leveraging their knowledge in the design of ML architectures or algorithms may lessen the need for massive training data. While the idea of hybrid AI that integrates knowledge- and data-driven methods has a venerable history, recent progress in Monte Carlo tree search and reinforcement learning now suggests new approaches. For instance, industrial expert knowledge is often procedural in the sense that there exists experience as to what to do when with measurements in order to achieve actionable results. Given training data for a novel problem and a database of interpretable procedures that have already proven their worth, Monte Carlo tree search or reinforcement learning can automatically compose basic building blocks into larger systems that solve the problem at hand [1]. At the same time, industrial product design or process control often rely on sophisticated knowledge-based simulations. Here, simulated data can augment small amounts of training data, or learning systems can improve existing simulators [2].

Crucially, the Fraunhofer Center for Machine Learning is part of the Fraunhofer Cluster of Excellence Cognitive Internet Technologies [L1]. Also comprising the centers for IoT Communications and Data Spaces, this cluster covers aspects of data acquisition, exchange, curation, and analysis. On the one hand, this ecosystem provides numerous testbeds for informed learning and explainable AI in industry. On the other hand, it provides opportunities for research and development of distributed or federated learning approaches as well as for leaning on the edge. The cluster thus supports digital sovereignty and develops trustworthy technologies for the industrial data economy.

**Link:**
[L1] https://www.cit.fraunhofer.de

**References:**
[1] C. Bauckhage, et al.: "Informed Machine Learning through Functional Composition", Proc. KDML, 2018.
[2] N. Aspiron and M. Bortz: "Process Modeling, Simulation and Optimization: From Single Solutions to a Multitude of Solutions to Support Decision Making", Chemie Ingenieur Technik, 90(11), 2018.

**Please contact:**
Daniel Schulz
Fraunhofer IAIS, Germany
+49 2241 142401
daniel.schulz@iais.fraunhofer.de

Endorsed by the ERCIM Working Group on Formal Methods for Industrial Critical Systems (FMICS)

Announcement and Call for Papers

# FM 2019: 23rd International Symposium on Formal Methods

Porto, Portugal, 7-11 October 2019

## 3rd World Congress on Formal Methods

Every 10 years, the Symposium on Formal Methods takes the form of a World Congress. FM 2019 is the 3rd World Congress on Formal Methods, organised as FM week with a number of distinguished co-located conferences, including LOPSTR, MPC, PPDP, RV, SAS, TAP, UTP and VECoS, 13 workshops, a doctoral symposium, an Industry day, 7 tutorials, and several other interesting events.

## The next 30 years

FM 2019 is the 23rd in a series of symposia organised by Formal Methods Europe, an independent association whose aim is to stimulate the use of, and research on, formal methods for software development. It is now more than 30 years since the first VDM symposium in 1987 brought together researchers with the common goal of creating methods to produce high quality software based on rigour and reason. Since then the diversity and complexity of computer technology has changed enormously and the formal methods community has stepped up to the challenges those changes brought by adapting, generalising and improving the models and analysis techniques that were the focus of that first symposium. The theme for FM 2019 is a reflection on how far the community has come and the lessons we can learn for understanding and developing the best software for future technologies.

## Invited speakers

- June Andronick (CSIRO/Data61 and UNSW, Sydney, Australia)
- Shriram Krishnamurthi (Brown University, Providence, RI, USA)
- Erik Poll (Radboud University, Nijmegen, The Netherlands)

## Submission and publication

Paper submission deadline:
11 April 2019.

FM 2019 encourages submissions on formal methods in a wide range of domains including software, computer-based systems, systems-of-systems, cyber-physical systems, human-computer interaction, manufacturing, sustainability, energy, transport, smart cities, and healthcare. Next to papers on theoretical foundations of formal methods and their use in software and systems engineering, we particularly welcome papers on techniques, tools and experiences in interdisciplinary settings. We also welcome papers on experiences of formal methods in industry, and on the design and validation of formal methods tools.

Accepted papers will be included in the Symposium Proceedings published in Springer's Lecture Notes in Computer Science in the subline on Formal Methods.

Authors of selected papers will be invited to submit an extended version of their paper to a special issue in "Formal Aspects of Computing" or in "Formal Methods in System Design".

## Symposium chairs

- Maurice ter Beek (ISTI-CNR, Pisa, Italy), Annabelle McIver (Macquarie University, Sydney, Australia)
- José Nuno Oliveira (INESC TEC & University of Minho, Portugal)

**More information:**
http://formalmethods2019.inesctec.pt/

Call for Proposals

# Dagstuhl Seminars and Perspectives Workshops

***Schloss Dagstuhl – Leibniz-Zentrum für Informatik is accepting proposals for scientific seminars/workshops in all areas of computer science, in particular also in connection with other fields.***

If accepted the event will be hosted in the seclusion of Dagstuhl's well known, own, dedicated facilities in Wadern on the western fringe of Germany. Moreover, the Dagstuhl office will assume most of the organisational/administrative work, and the Dagstuhl scientific staff will support the organizers in preparing, running, and documenting the event. Thanks to subsidies the costs are very low for participants.

Dagstuhl events are typically proposed by a group of three to four outstanding researchers of different affiliations. This organizer team should represent a range of research communities and reflect Dagstuhl's international orientation. More information, in particular, details about event form and setup as well as the proposal form and the proposing process can be found on

**http://www.dagstuhl.de/dsproposal**

Schloss Dagstuhl – Leibniz-Zentrum für Informatik is funded by the German federal and state government. It pursues a mission of furthering world class research in computer science by facilitating communication and interaction between researchers.

## Important Dates

- Proposal submission: April 1 to April 15, 2019
- Notification: July 2019
- Seminar dates: Between March 2020 and February 2021 (tentative).

# ERCIM "Alain Bensoussan" Fellowship Programme

The ERCIM PhD Fellowship Programme has been established as one of the premier activities of ERCIM. The programme is open to young researchers from all over the world. It focuses on a broad range of fields in Computer Science and Applied Mathematics.

The fellowship scheme also helps young scientists to improve their knowledge of European research structures and networks and to gain more insight into the working conditions of leading European research institutions.

The fellowships are of 12 months duration (with a possible extension), spent in one of the ERCIM member institutes. Fellows can apply for second year in a different institute.

## Why to apply for an ERCIM Fellowship?

The Fellowship Programme enables bright young scientists from all over the world to work on a challenging problem as fellows of leading European research centers. In addition, an ERCIM fellowship helps widen and intensify the network of personal relations and understanding among scientists. The programme offers the opportunity to ERCIM fellows:

- to work with internationally recognized experts;
- to improve their knowledge about European research structures and networks;
- to become familiarized with working conditions in leading European research centres;
- to promote cross-fertilization and cooperation, through the fellowships, between research groups working in similar areas in different laboratories.

## Conditions

Candidates must:

- have obtained a PhD degree during the last eight years (prior to the year of the application deadline) or be in the last year of the thesis work with an outstanding academic record;
- be fluent in English;
- have no obligations with respect to military service which could impact on the fellowship;
- have completed their PhD before starting the grant;
- submit the following required doucments: cv, list of publications, two scientific papers in English, contact details of two referees.

"Success is based on choice, having and maintaining a motivation worth fighting for. I truly appreciate ERCIM Alain Bensoussan Fellowship Program that brought me the position of a Sr. Assistant Professor. I can think of no better postdoctoral fellowship other than ERCIM as it encourages fellows to pursue their own ideas and develop broad collaborations.

**Monalisa MANDAL**
Former ERCIM Fellow

"Thanks to ERCIM for all the support during my Fellowship. It was a wonderful year with a lot of personal and professional learning. All the knowledge that I acquired in Norway is now very useful in Colombia and we have established a great cooperation network.

**Maximiliano Bueno López**
Former ERCIM Fellow

"As the French biologist, Louis Pasteur once said, in the fields of observation chance favors only the prepared mind. Alain Bensoussan Fellowship is a great chance well suited for an early career scientist with high potential within, to build an outstanding research profile. Through the ERCIM Alain Bensoussan Fellowship I could collaborate and network with top scientists and high-tech companies in the field of artificial intelligence and prediction models. Today, as a senior researcher with a higher confidence in leadership I can better pursue a research career in research institutions.

**Amir MOSAVI**
Former ERCIM Fellow

The fellows are appointed either by a stipend (an agreement for a research training programme) or a working contract. The type of contract and the monthly allowance/salary depends on the hosting institute.

Deadlines for applications are currently 30 April and 30 September each year.

Since its inception in 1991, over 500 fellows have passed through the programme. In 2018, 30 young scientists commenced an ERCIM PhD fellowship and 57 fellows have been hosted during the year. In 2005 the Fellowship Programme was named in honour of Alain Bensoussan, former president of Inria, one of the three ERCIM founding institutes.

http://fellowship.ercim.eu

# POEMA
# 15 Doctoral Student Positions Available

on the subject of Polynomial Optimization, Efficiency through Moments and Algebra
at eleven European Research Institutes and Universitites

The Innovative Training Network POEMA is hiring 15 Doctoral Students starting from September 2019. The proposed projects will be investigating the development of new algebraic and geometric methods combined with computer algebra techniques for global non-linear optimization problems. Applications will focus on smarter cities challenges, urban traffic management, water network management, energy flow control, or environmental monitoring.

## Available positions

1. **Algebraic tools for exact SDP and its variants**
   Advisor: M. Safey el Din (Sorbonne University, Paris, France)

2. **Exact algorithms for structured polynomial optimisation**
   Advisor: M. Safey el Din (Sorbonne University, Paris, France)

3. **Polynomial optimization problems with symmetry**
   Advisor: C. Scheiderer (Univ. of Konstanz, Germany)

4. **Hyperbolic polynomials and the Generalized Lax Conjecture**
   Advisor: M. Schweighofer (Univ. of Konstanz, Germany)

5. **Tensor Decomposition by Vector Bundles tools**
   Advisor: G. Ottaviani (Firenze, Italy)

6. **Approximation hierarchies for (non-)commutative polynomial optimisation**
   Advisor: M. Laurent (CWI, Amsterdam, the Nerthelands)

7. **Approximation hierarchies for graph parameters**
   Advisor: M. Laurent (CWI, Amsterdam, the Nerthelands)

8. **Polynomial Optimization Problems in Operations Research and Finance**
   Advisor: E. de Klerk (Univ. of Tilburg, the Netherlands)

9. **Structure of moment problems and applications to polynomial optimisation**
   Advisor: B. Mourrain (INRIA, Sophia Antipolis, France)

10. **Alternative polynomial bases for global optimization**
    Advisor: E. Hubert (INRIA, Sophia Antipolis, France)

11. **Numerical cubature with symmetry and applications to polynomial optimisation**
    Advisor: C. Riener (Univ. of Tromsoe, Norway)

12. **Algorithms and software for nonlinear convex conic optimisation**
    Advisor: M. Stingl (Friedrich Alexander Univ. Erlangen, Germany)

13. **Algorithms and software for structured SDP**
    Advisor: M. Kocvara (Univ. of Birmingham, UK)

14. **Polynomial Optimization: Some challenges from applications**
    Advisor: M. Korda (CNRS, LAAS, Toulouse, France)

15. **Polynomial Optimization Techniques for Energy Network Operation and Design**
    Advisors: J.-H. Hours, M. Gabay (ARTELYS, Paris, France)

The positions have usually a duration of three years. Applicants must have a Master's degree in Computer Science, Mathematics or Engineering (or any equivalent diploma) at the date of recruitment.

POEMA is a Marie Skłodowska-Curie Innovative Training Network funded by the European Union. Its goal is to train scientists at the interplay of algebra, geometry and computer science for polynomial optimization problems and to foster scientific and technological advances, stimulating interdisciplinary and intersectoriality knowledge exchange between algebraists, geometers, computer scientists and industrial actors facing real-life optimization problems.

**More information and application submission: https://easychair.org/cfp/POEMA-19-22**

# Editorial Information

*ERCIM News is the magazine of ERCIM. Published quarterly, it reports on joint actions of the ERCIM partners, and aims to reflect the contribution made by ERCIM to the European Community in Information Technology and Applied Mathematics. Through short articles and news items, it provides a forum for the exchange of information between the institutes and also with the wider scientific community. This issue has a circulation of about 6,000 printed copies and is also available online.*

*ERCIM News is published by ERCIM EEIG*
*BP 93, F-06902 Sophia Antipolis Cedex, France*
*Tel: +33 4 9238 5010, E-mail: contact@ercim.eu*
*Director: Philipp Hoschka, ISSN 0926-4981*

***Contributions***
*Contributions should be submitted to the local editor of your country*



***Advertising***
*For current advertising rates and conditions, see http://ercim-news.ercim.eu/ or contact peter.kunz@ercim.eu*

***ERCIM News online edition:*** *ercim-news.ercim.eu/*

***Next issue:*** *April 2019, Special theme: 5G*

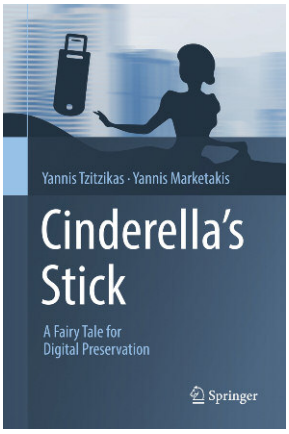
***Subscription***
*Subscribe to ERCIM News by sending an email to en-subscriptions@ercim.eu or by filling out the form at the ERCIM News website: http://ercim-news.ercim.eu/*



***Cover illustrtation:*** *source: Shutterstock.*

New Book

## Cinderella's Stick – A Fairy Tale for Digital Preservation




Authors: Yannis Tzitzikas and Yannis, Marketakis (ICS-FORTH)

This book explains the main problems related to digital preservation using examples based on a modern version of the well-known Cinderella fairy tale. Digital preservation is the endeavor to protect digital material against loss, corruption, hardware/software technology changes, and changes in the knowledge of the community.

The structure of the book is modular, with each chapter consisting of two parts: the episode and the technical background. The episodes narrate the story in chronological order, exactly as in a fairy tale. In addition to the story itself, each episode is related to one or more digital preservation problems, which are discussed in the technical background section of the chapter. To reveal a more general and abstract formulation of these problems, the notion of pattern is used. Each pattern has a name, a summary of the problem, a narrative describing an attempt to solve the problem, an explanation of what could have been done to avoid or alleviate this problem, some lessons learned, and lastly, links to related patterns discussed in other chapters.

The book is intended for anyone wanting to understand the problems related to digital preservation, even if they lack the technical background. It explains the technical details at an introductory level, provides references to the main approaches (or solutions) currently available for tackling related problems, and is rounded out by questions and exercises appropriate for computer engineers and scientists. In addition, the book's website, maintained by the authors, presents the contents of Cinderella's "real USB stick," and includes links to various tools and updates.

Springer 2018, eBook ISBN: 978-3-319-98488-9
DOI 10.1007/978-3-319-98488-9.

## HORIZON 2020 Project Management

A European project can be a richly rewarding tool for pushing your research or innovation activities to the state-of-the-art and beyond. Through ERCIM, our member institutes have participated in more than 90 projects funded by the European Commission in the ICT domain, by carrying out joint research activities while the ERCIM Office successfully manages the complexity of the project administration, finances and outreach. The ERCIM Office has recognized expertise in a full range of services, including identification of funding opportunities, recruitment of project partners, proposal writing and project negotiation, contractual and consortium management, communications and systems support, organization of events, from team meetings to large-scale workshops and conferences, support for the dissemination of results.

**How does it work in practice?**
Contact the ERCIM Office to present your project idea and a panel of experts will review your idea and provide recommendations. If the ERCIM Office expresses its interest to participate, it will assist the project consortium either as project coordinator or project partner.

**Please contact:**
Peter Kunz, ERCIM Office, peter.kunz@ercim.eu

## CWI, EIT Digital, Spirit, and UPM launch Innovation Activity "G-Moji"

The Dutch youth care organization Spirit, CWI and Universidad Politécnica de Madrid (UPM) join forces in the new innovation activity of EIT Digital "G-Moji" – a smartphone application to improve lives of youth at risk. G-Moji aims to support youth with mental health issues. In recent years, in the Netherlands, budgets for youth aid decreased by roughly 24% while more and more young people need mental healthcare with complex needs. The new intervention is targeted at youth professionals working with youth at-risk to support personalized decision making and interventions of young clients with mental problems. The analysis of big data – collected through the usage of the new smartphone application – will help professionals predict the mental state and behavioral patterns of youth aged between 16 and 24 years. By using the smartphone, youth and professionals inform each other real-time about problematic situations which should result in reducing clinical treatments and improving ambulatory treatment. Potentially, this will lead to better tailor-made care, on-distance if suitable. The current use of smartphones by this group allows for a wide range of big data measurements through self-monitoring and hard- and software sensors embedded in the mobile phone. The measurements make it possible to detect mental, emotional and behavioral youth problems quickly, resulting in a more effective and customized intervention than conventional solutions such as questionnaires and interviews. In this project, CWI will be responsible for analyzing the data obtained from mobile and social interaction, physical activity and speech recognition. Based on these analyses CWI will develop and evaluate forecasting models for interventions.
More information: https://kwz.me/hdz

Data Market Services

## New EU Project Data Market Services

ERCIM participates in DataMarketServices (DMS), a new H2020 project whose objective is to overcome the barriers of data-centric European SMEs and start-ups by providing free support services around data skills, entrepreneurial opportunities, legal issues and standardization. The expected project deliverables are a 100-data-based companies portfolio, twelve free support services, webinars and training in topics such as GDPR, IPR, etc.

The DMS consortium is composed of Zabala (ES); the University of Southampton (UK); Spinlab (DE); Spherik (RO); Bright Pixel (PT); The Next Web (NL); Ogilvy (ES); IPTector (DK), and ERCIM which hosts the European branch of the World Wide Web Consortium (W3C).
http://www.datamarketservices.eu/
@datamarketsvc

W3C DEVELOPERS

## New W3C Web Experts Videos

World renowned Web experts were attending the W3C TPAC '18 meeting in October in Lyon, and the W3C developer relations team seized this opportunity to organize a developer meetup featuring five prominent speakers and twelve demonstrations. Hosted by the University of Lyon, this event was made possible with the support from Qwant, Microsoft, Mozilla, NTT Communications, StickerMule, Webcastor and WithYou. Each of the five speakers delivered an expert talk that was recorded: Rachel Andrew on CSS Layout, Ali Spivak on documenting Web standards on MDN, Manuel Rego on how to participate in the CSS specs development, Tristan Nitot on privacy and online services, and


*Screenshot from Ada Rose Cannon's presentation "Why bringing Virtual and Augmented Reality (XR) to the Web?".*

Richard Ishida on making the Web truly world wide. These videos, as well as short videos shot during the meeting, such as Luke Wagner talking about Web Assembly, Elika J. Etemad (Fantasai) talking about specification editing best practices, Ada Rose Cannon on WebXR, and others, are now available on W3C's Vimeo page.
https://vimeo.com/w3c
https://www.w3.org/developers/

## Celebrate the Web@30


In 1989, CERN was a hive of ideas and information stored on multiple incompatible computers. Tim Berners-Lee envisioned a unifying structure for linking information across different computers, and wrote a proposal in March 1989 called "Information Management: A Proposal". By 1991, this vision of universal connectivity had become the World Wide Web! To celebrate 30 years since Tim Berners-Lee's proposal and to kick-start a series of celebrations worldwide, CERN will host a 30th anniversary event on the morning of 12 March 2019 in partnership with the World Wide Web Consortium (W3C) and the World Wide Web Foundation. This anniversary event will be webcast.
https://cern.ch/web30

ERCIM
European Research Consortium for Informatics and Mathematics

**ERCIM – the European Research Consortium for Informatics and Mathematics is an organisation dedicated to the advancement of European research and development in information technology and applied mathematics. Its member institutions aim to foster collaborative work within the European research community and to increase co-operation with European industry.**

**ERCIM is the European Host of the World Wide Web Consortium.**

CNR
Consiglio Nazionale delle Ricerche
Area della Ricerca CNR di Pisa
Via G. Moruzzi 1, 56124 Pisa, Italy
www.iit.cnr.it

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology, Mathematics and Electrical Engineering, N 7491 Trondheim, Norway
http://www.ntnu.no/

CWI
Centrum Wiskunde & Informatica
Science Park 123,
NL-1098 XG Amsterdam, The Netherlands
www.cwi.nl

RI.SE
RISE SICS
Box 1263,
SE-164 29 Kista, Sweden
http://www.sics.se/

Fonds National de la Recherche Luxembourg
Fonds National de la Recherche
6, rue Antoine de Saint-Exupéry, B.P. 1777
L-1017 Luxembourg-Kirchberg
www.fnr.lu

SBA Research
SBA Research gGmbH
Favoritenstraße 16, 1040 Wien
www.sba-research.org/

FORTH
Foundation for Research and Technology – Hellas
Institute of Computer Science
P.O. Box 1385, GR-71110 Heraklion, Crete, Greece
www.ics.forth.gr

MTA SZTAKI
Magyar Tudományos Akadémia
Számítástechnikai és Automatizálási Kutató Intézet
P.O. Box 63, H-1518 Budapest, Hungary
www.sztaki.hu/

Fraunhofer IUK-TECHNOLOGIE
Fraunhofer ICT Group
Anna-Louisa-Karsch-Str. 2
10178 Berlin, Germany
www.iuk.fraunhofer.de

TNO innovation for life
TNO
PO Box 96829
2509 JE DEN HAAG
www.tno.nl

inesc
INESC
c/o INESC Porto, Campus da FEUP,
Rua Dr. Roberto Frias, nº 378,
4200-465 Porto, Portugal
www.inesc.pt

University of Cyprus
P.O. Box 20537
1678 Nicosia, Cyprus
www.cs.ucy.ac.cy/

Institut National de Recherche en Informatique et en Automatique
B.P. 105, F-78153 Le Chesnay, France
www.inria.fr

University of Warsaw
Faculty of Mathematics, Informatics and Mechanics
Banacha 2, 02-097 Warsaw, Poland
www.mimuw.edu.pl/

I.S.I. – Industrial Systems Institute
Patras Science Park building
Platani, Patras, Greece, GR-26504
www.isi.gr

VTT Technical Research Centre of Finland Ltd
PO Box 1000
FIN-02044 VTT, Finland
www.vttresearch.com