First World Z-inspection® Conference

Ateneo Veneto, March 10-11, 2023, Venice, Italy

Conference Reader

QUOTES

„Last year Provincie Fryslân, Zicari's interdisciplinary team of scientists and Rijks ICT Gilde ran the pilot project 'Assessment for Responsible AI'. I am proud that integrating the FRAIA into the Z-Inspection® method contributed to great conversations about human rights, both in the pilot and during the conference. To me, that really shows the strength of the Z-Inspection® method and its community, and also its value to the Dutch government."
-- Willy Tadema, AI Ethics Lead, Rijks ICT Gilde.

„Accelerated digitalization, replacing human employment, algorithms all over the place. Where do we draw the line? What an amazing world's first Z-Inspection® Conference in Venice Willy Tadema, Gerard Kema and I had! We talked about #Trustworthy, #ethics and #human rights with over 60 attendees with different backgrounds, expertise and knowledge. For me this conference was all about sharing , learning and connecting. An experience I will forever take with me in my heart."
-- Marijke ter Steege, Senior Consultant Data and Strategy, Rijks ICT Gilde.

„Last weekend I had the pleasure to be part of a panel at the World Z-inspection® Conference in Venice, giving Merck's perspective and mechanisms for ethical handling of AI It was a thrilling experience to exchange with leading academics in this field (and some fellow industry experts) - all thanks to the organizer Roberto V. Zicari
Adding to this experience was the unique location of Ateneo Veneto - imagine listening to a debate on AI while sitting in front of a Tintoretto painting..."
-- Jean Enno Charton, Director Digital Ethics & Bioethics, Merck

„What a wonderful event last week at the Ateneo Veneto in Venice! I am very grateful that I could be part of it and share my experience with the #zinspectionCo-Design assessment of our explainable AI project exAID (https://exaid.kl.dfki.de/).
Thank you so much Roberto V. Zicari for bringing together so many inspiring people, motivated to advance the field of #"TrustworthyAI!"
-- Adriano Lucieri, PhD Candidate, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)

„How can we develop #trustworthy #AI applications and bring them into use? And what does a holistic #assessment process look like for assessing the trustworthiness of an AI system used in #medicine, #industry, or the #Public sector?
The first World Z-Inspection Conference, held at ATENEO VENETO Istituto Culturale ONLUS in Venice this month, focused on these and many other opportunities and challenges to be addressed.“
-- Johannes Winter, Chief Strategy Officer, L3S Research Center

„We are about to start a Trustworthy AI for Healthcare Lab in Poland. Stay tuned! Assessment of trustworthy AI systems is always a process and as with every journey, it needs to start with the first step. Thus, we are now forming a use case in healthcare that will be our first hands-on experience with the Z-Inspection®. „
-- Katarzyna Kaczmarek-Majer, Polish Academy of Sciences

“As most recent advances in AI such as generative AI are transforming our society, way of living and work (#futureofwork), the ethics of data & AI will be one of the key themes guiding our future to protect human rights and a human-centric approach to technology overall. This clearly illustrates the importance of interdisciplinary high-profile events like this 1st Global Z-Inspection® Conference.``
**-- Lisa Bechtold, Global Lead AI Assurance & Data Governance,
Zurich Insurance Company Ltd**

„What an amazing first World Z-Inspection® conference in Venice!
Brilliant people coming together from all over the world, so grateful for this experience! Thank you Roberto V. Zicari, Venice Urban Lab, Global Campus of Human Rights and all the amazing attendees!“
-- Hanna Sormunen, Data Scientist and Chairman of the AI ethics board at Verohallinto - Finnish Tax Administration

TABLE OF CONTENTS

PROGRAM

First World Z-inspection® Conference: Ateneo Veneto, March 10-11, 2023, Venice, Italy.

Press Release (Ateneo Veneto):
https://ateneoveneto.org/world-z-inspection-conference-sullintelligenza-artificiale/

In cooperation with Global Campus of Human Rights (https://gchumanrights.org) and Venice Urban Lab (https://www.veniceurbanlab.org/en)

Supporters:
Arcada University of Applied Science,
Merck, Roche, Zurich Insurance Company.

……

Friday (afternoon), March 10, 2023

– Welcome remarks
Antonella Magaraggia, President, Ateneo Veneto,
Sergio Pascolo, President, Venice Urban Lab,
George Ulrich, Academic Director, Global Campus of Human Rights.

– The Z-inspection® initiative
Roberto V. Zicari (Lead Z-inspection® initiative),

– Presentation of selected Affiliated Trustworthy AI Labs
Magnus Westerlund (The Laboratory for Trustworthy AI at Arcada University of Applied Sciences (Helsinki, Finland).
Sune Holm, Boris Düdder (Trustworthy AI Lab at the University of Copenhagen (Copen hagen, Denmark)
Gemma Roig, Karsten Tolle (Trustworthy AI Lab at the Goethe University Frankfurt (Frankfurt, Germany)
Vince Madai (Trustworthy AI in Healthcare Lab at the QUEST Centre for Responsible Research (Berlin Institute of Health at Charité (BIH) (Germany)
Pedro Moreno Sanchez (Trustworthy AI for healthcare Lab, Tampere University (Finland)
Roberto Francischello (Trustworthy AI Lab at the Imaging Lab, University of Pisa (Pisa, Italy)

– Panel: “Human Rights and Trustworthy AI”
Panelists:
George Ulrich (Academic Director, Global Campus of Human Rights),
Elisabeth Hildt (Director, Center for the Study of Ethics in the Professions, Illinois Institute of Technology, and L3S Research Center, Leibniz University Hannover)
Emilie Wiinblad Mathez (Senior Ethics Adviser)
Frédérick Bruneault (adjunct professor, École des médias Unive, Université du Qué bec,Montréal)
Peter G. Kirchschlaeger (Ethics-Professor, Director of the Institute of Social Ethics ISE, University of Lucerne)
Giovanni Sartor (Professor in Legal Informatics at the University of Bologna)
Moderator: Holger Volland CEO brand eins, Germany)

Saturday (all day), March 11, 2024

– Welcome remarks
Gianpaolo Scarante, Past President Ateneo Veneto

– Pilot Project “Responsible use of AI” with Rijks ICT Gilde
(Ministry of Interior and Kingdom Relationships) , and the Province of Friesland, The Netherlands
Willy Tadema (AI Ethics Lead, Rijks ICT Gilde, The Netherlands),
Marijke Steege, Marijke (Senior Consultant Strategy Innovation and Data, Rijks ICT Gil de, The Netherlands),
Gerard Kema (Innovator Manager, Province of Friesland, The Netherlands)

– Trustworthy AI in Practice: Best practices
Mattia Savardi. Davide Farini, Alberto Signoroni
Alberto Signoroni, (University of Brescia, Italy),
Mattia Savardi, (University of Brescia, Italy),
Davide Farina (University of Brescia, Italy),
Hanna Sormunen (Finnish Tax Administration),
Vince Madai (QUEST, Berlin)

– Panel: “How do we trust AI? “
Panelists:
Jean Enno Charton (Director Bioethics & Digital Ethics, Merck),
Bryn Roberts (Global Head of Data & Analytics, Roche),
Sarah Gadd (Head of Data & Artificial Intelligence Solutions, Credit Suisse),
Lisa Bechtold, (Global Lead AI Assurance & Data Governance · Zurich Insurance Com pany)
Moderator: Holger Volland (CEO brand eins, Germany)

– Trustworthy AI in Practice: Best practices
Ulrich Kühne (Hautmedizin Bad Soden, Germany)
Adriano Lucieri (DFKI, Germany)
James Brusseau (Pace University, USA)
Adarsh Srivastava (Roche, India)

- Concluding Remarks
Sergio Pascolo (President, Venice Urban Lab)
Roberto V. Zicari (Lead Z-inspection® initiative)

“Singing tuning with Ahhh.” Sessions: Alessandro Donati

Conference Moderator: Holger Volland (CEO, brand eins, Germany)

Why this conference matters

Holger Volland

In the spring of 2023, a number of people from academia, the tech sector and society issued a call. They called for a moratorium, a pause in the development of artificial intelligence. This was preceded by worldwide media reports about possible dangers and problems, triggered by the latest versions of generative AI like GPT or Midjourney.

While some stakeholders in the moratorium have singular interests, and both the wisdom and feasibility of a six-month „development pause“ are questionable, they were right about one thing: Many developments using machine learning, large language models, or other applications of AI are being pushed forward without thoroughly addressing the trustworthiness of the developments, their ethical, moral, health, social, or even legal implications.

One reason for this is that most applications exclusively involve experts from the respective domain, such as health or mobility, as well as technological expertise. Obviously, this means that the requirements, design and implications of an application also integrate only the limited expertise from these two domains. What is far too often missing are interdisciplinary discussions with experts from other relevant fields. Only these could prevent a certain blindness to the later effects in the application.

Interdisciplinary considerations together with standardized and documented processes are essential components of Z-inspection®. At the First World Z-inspection® Conference in Venice, scientists and representatives of organizations as well as companies from numerous countries and disciplines came together for the first time to discuss issues around trustworthy AI and the findings and work of the first Z-inspection® Labs.

Interdisciplinarity, transparent processes and careful ongoing consideration of all developments allow the necessary trust to emerge in society and politics, which should form the basis of all future AI applications. Only in this way can well-founded discussions be conducted with the participation of all relevant stakeholders.

Holger Volland
Holger Volland is CEO of brand eins Medien AG in Hamburg. Responsible technological growth, impact business and diversity in boardrooms and supervisory boards are his topics. As an author on the topics of transformation and AI, he is published in major national and international publishing houses. He is also an experienced keynote speaker and lecturer (Bits & Pretzels, St. Gallen, Goethe-Institut, DLD, Art Directors Club, SXSW, Frankfurt Book Fair, etc.). He is Advisory Board Member of the Sonophilia Foundation and active supporter at Z-Inspection, the interdisciplinary science network for Mindful Use of AI, among others.

Welcome Remarks

Sergio Pascolo

Let the future of the world be discussed in Venice, a harmonious city. (Le Corbusier)
On initiative of Roberto Zicari as scientific coordinator and me as President of the Venice Urban Lab has been established in 2022 the Trustworthy AI Venice Lab, with the goal to combine the contribute for a better world with the one for a better future of this fragile and beautiful city in which we are.

We remember Le Corbusier's famous exhortation in the 1960s when he was commissioned to design the new hospital in Venice, a thoughtful and intelligent project unfortunately never realized, because the grand master of 20th-century architecture, concerned about the fate of this city, urged us to do what we did in this important conference at the Ateneo Veneto in Venice on March 10 and 11, 2023: a great collective work for a better world. The theme was an ethical approach to deal with the exponential development of artificial intelligence, a big challenge for the future.

AI is in fact already changing the way we live and work. It's revolutionizing industries from healthcare to finance, from transportation to agriculture and last not least to cities governance and the life of citizens.

AI systems in urban environments can raise concerns around privacy, so we must focus on transparency and accountability about the data being collected and how it is being used. This is particularly important in the context of urban planning, where minor changes can have significant impacts on people's lives.

AI systems should be subject to appropriate oversight and regulation, with mechanisms in place to address any issues or concerns that may arise. Lastly, we must engage the community throughout the development process, be transparent about the use of AI systems, and ensure that they are developed in a fair, transparent, and accountable manner. There is a need to prioritize sustainability and ethics in all areas of our lives.

The host of Trustworthy AI Venice Lab, Venice Urban Lab, is an organization that deals with the sustainable transformation processes of the city with a holistic and multidisciplinary approach. We are Partner of the New European Bauhaus, an European Commission project promoting the improvement of the quality of living spaces and coexistence of citizens with the three keywords: beauty, sustainability together. On the trail of the SDGs, Sustainable Development Goals of the U.N., we strive to address global challenges with local solutions.

Our work in Venice is centered around preserving and exporting values such as proximity, walkability, beauty, quality of time, solidarity and inclusion. We aim to create a model of a diversified economy enabling Venice to remain an inhabited city which is not guaranteed given the depopulation of the last 6 decades during which the city lost more than 120.000 Inhabitants. Now the city has less than 50,000 inhabitants and the increasing pressure of the global mass tourism economy threatens to turn the city into a large resort with no inhabitants. This risk is very high and the challenge very complex because the economic interests are enormous and the social consequences equally so.

Rather, we envision Venice as one of the most attractive small cities on the planet where people can live well because of its high and sustainable quality of life, transmitting values of balance, harmony and peaceful coexistence to the world.
The values we are preserving and exporting are about creating a better world for everyone. These values encourage us to consider the needs of others and the impact of our actions on the environment and society. They promote a sense of responsibility towards the world and its inhabitants, both human and non-human.
This work that we are carrying out together with a network of other local associations aligns closely with the principles of ethics.

Related to the ethics of AI in urban issues, an important example for the topic we are dealing with, is the New York AI Localism, a group of

researchers working on bottom-up approaches to AI regulation. In this case cities are presented not so much as centers of government, but as networks, resilient, adaptive, collaborative, and thriving ecosystems, sharing best practices that can be emulated by communities around the world.

And this is what many cities need, what the city of Venice needs and what we want to improve with the Trustworthy AI Venice Lab.In the last years the governance challenges of AI applications have captured the best attention of the world's cities. The concept of smart city is well known and more and more developed, opening many ethical questions. About this topic many cities around the world are doing an amount of regulatory work and concrete efforts in setting rules to promote ethical and sustainable use of AI.

London, Barcelona, and Amsterdam launched in 2021 the Global Observatory on Urban AI with the intention of monitoring trends and promoting its ethical and responsible use. Together with Barcelona and Amsterdam, Montreal, San Francisco, Porto, New York, Helsinki, Toronto, and Seattle are involved in building a human rights-based AI and technology model consonant with a democratic digital society.
As a good example, the themes of the Montréal Declaration on Responsible Development of Artificial Intelligence 2018 are the principle of well-being, respect for autonomy, protection of privacy and confidentiality, the principle of solidarity, democratic participation and equity, the principle of inclusion of diversity, the principle of caution and responsibility, the principle of sustainable development
Similarly the city of Toronto developed the Digital Infrastructure Strategic Framework (DISF – DIP) wth Principles of Equity and Inclusion, A Well-run City, Economy and the Environment, Privacy and Security, Democracy and Transparency, Digital Autonomy.

We acknowledge a great convergence of ethical principles between these examples and the most advanced protocols of urban quality, such as the New European Bauhaus and the Declaration of Davos 2018 in which the European Ministers of Culture are promoting a vision for a high-quality „Baukultur" recognising that we urgently need a new, adaptive approach to shaping our built environment; one that is rooted in different culture, actively builds social cohesion, ensures environmental sustainability, and contributes to the health and well-being. This shows us that just as on a spiritual level we are all connected, even in urban matters everything is connected - quality of living space, inclusion, mobility, solidarity, environment, energy, waste management, and many other aspects of a sustainable coexistence.
The development of AI ethics through Z-Inspection® processes can contribute precisely in creating virtuous connections.

Many things are not because they are difficult that we dare not do them, but it is because we dare not do them that they are difficult
Seneca

Sergio Pascolo
architect and urbanist, Adjunct Professor of architectural and urban design at the Iuav University of Venice, author of the book "Venezia secolo Ventuno. Visioni e strategie per un rinascimento sostenibile", Founder and President of Venice Urban Lab.

Z-INSPECTION®: A PROCESS TO ASSESS TRUSTWORTHY AI

Roberto V. Zicari

There are a number of frameworks and guidelines for Responsible AI and/or Trustworthy AI. For examples: Ethics Guidelines for Trustworthy AI. Independent High-Level Expert Group on Artificial Intelligence. European commission,UN,OECD Recommendation of the Council on Artificial Intelligence, UNESCO Recommendation on the ethics of artificial intelligence, Fundamental Rights and Algorithm Impact Assessment (FRAIA), United Nations Framework for Ethical AI , to name a few, give guidelines but do not tell you how to assess in practice the use of AI in a given context and domain.
This is where Z-Inspection® comes into play, either as a co-design, self-assessment, or auditing method. Its ultimate goal is to foster high levels of trustworthiness of AI systems, entailing them to be fair, safe, transparent, as well as socially acceptable.

Z-inspection® is a participatory process, developed by a non commercial initiative, which helps stakeholders to assess the risks of using AI in a given context and map such risks to a given Framework (e.g. EU Trustworth AI).

Z-Inspection® is a holistic process based on the method of evaluating new technologies, where ethical issues need to be discussed through the elaboration of socio-technical scenarios. In particular, Z-Inspection® can be used to perform independent assessments and/or self-assessments together with the stakeholders owning the use case.

Relevant Links:
https://z-inspection.org

Z-Inspection® is listed in the new OECD Catalog of AI Tools & Metrics
https://oecd.ai/en/catalogue/tools/z-inspection

.....

Best Practices

The Z-inspection® Process has been used by a team of interdisciplinary experts to assess a number of use cases in different domains.

Healthcare

- On Assessing Trustworthy AI in Healthcare. Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls.
In cooperation with the Emergency Medical Services Copenhagen - responsible to manage the 112 Health emergency calls for the City of Copenhagen- Denmark
https://www.frontiersin.org/articles/10.3389/fhumd.2021.673104/full

- Co-Design of a Trustworthy AI System in Healthcare: Deep Learning Based Skin Lesion Classifier.
In cooperation with the German Research Center for Artificial Intelligence GmbH (DFKI)
https://www.frontiersin.org/articles/10.3389/fhumd.2021.688152/full

- Deep Learning for predicting a multi-regional score conveying the degree of lung compromise in COVID-19 patients.
In cooperation with Department of Information Engineering and Department of Medical and Surgical Specialties, Radiological Sciences, and Public Health – Brescia Public Hospital (ASST Spedali Civili) Brescia, Italy.
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9845195

Nature/ Biodiversity

Pilot Project: Assessment for Responsible Artificial Intelligence together with Rijks ICT Gilde -Ministry of the Interior and Kingdom Relations (BZK)- and the province of Fryslân (The Netherlands)
https://z-inspection.org/general/pilot-project-assessment-for-responsible-artificial-intelligence-together-with-national-ict-guild-and-the-province-of-fryslan-the-netherlands/

What Does It Take to Use the Process?

Depending on the use as co-design, self-assessment, or auditing method, it requires first to build a *team of interdisciplinary experts* in various disciplines (ranging from applied ethics, to domain experts, legal scholars, machine learning engineers, social scientists, etc) The team will then work on the *identification and discussion of ethical issues and tensions through the elaboration of socio-technical scenarios.*

It will then map such *„risks" to a framework of choice*, e.g The EU Framework for Trustworthy AI with four ethical principles and seven requirements.

It will finish *giving a set of recommendations for key stakeholders.*

A typical best practice lasts between 3-6 months. A light version of the process can last a few weeks.

Resources:

How to Assess Trustworthy AI in Practice.

This report is a methodological reflection on Z-Inspection®. Z-Inspection® is a holistic process used to evaluate the trustworthiness of AI-based technologies at different stages of the AI lifecycle. It focuses, in particular, on the identification and discussion of ethical issues and tensions through the elaboration of socio-technical scenarios. It uses the general European Union's High-Level Expert Group's (EU HLEG) guidelines for trustworthy AI. This report illustrates for both AI researchers and AI practitioners how the EU HLEG guidelines for trustworthy AI can be applied in practice. We share the lessons learned from conducting a series of independent assessments to evaluate the trustworthiness of AI systems in healthcare. We also share key recommendations and practical suggestions on how to ensure a rigorous trustworthy AI assessment throughout the life-cycle of an AI system.
https://arxiv.org/abs/2206.09887

Trustworthy AI Labs are established, based on the Z-Inspection® process

The following Labs are affiliated with the Z-Inspection® Initiative:
https://z-inspection.org/affiliated-labs/

Z-Inspection® teaching certification

Benefits:
A "Z-Inspection® Teaching Certificate" allows you to teach the Z-Inspection® process for non-commercial purposes.
List of Certified Z-Inspection® Teaching Experts:
https://z-inspection.org/z-inspection-teaching-certification/

What are we interested in?

1. We are interested to look for non-commercial entity who are interested to set up Trustworthy AI Labs affiliated with the Z-Inspection® Initiative

2. We are looking for governments and/or non commercial organizations who are interested to conduct a Trustworthy AI self-assessment together with the Z-Inspection® Initiative and some of the affiliated Labs.

Prof. Roberto V. Zicari
Z-Inspection® Initiative Lead

Roberto V. Zicari is an affiliated professor at the Yrkeshögskolan Arcada, Helsinki, Finland, and an adjunct professor at the Seoul National University, South Korea.

Roberto V. Zicari is leading a team of international experts who defined an assessment process for Trustworthy AI, called Z-Inspection®.

Previously he was professor of Database and Information Systems (DBIS) at the Goethe University Frankfurt, Germany, where he founded the Frankfurt Big Data Lab.
He is an internationally recognized expert in the field of Databases and Big Data. His interests also expand to Ethics and AI, Innovation and Entrepreneurship. He is the editor of the ODBMS.org web portal and of the ODBMS Industry Watch Blog. He was for several years a visiting professor with the Center for Entrepreneurship and Technology within the Department of Industrial Engineering and Operations Research at UC Berkeley (USA).

Tracing the Contours of a Human Rights-Based Approach to AI

George Ulrich

Introduction

Let me begin by complimenting the organizers on a highly relevant conference theme, which touches upon some of the most pressing challenges of our era. In addition to familiar concerns about social control, the safeguarding of privacy and manipulation of our personal data, Artificial Intelligence profoundly affects, and will increasingly continue to affect (for better and for worse), issues as diverse as cultural creativity, democratic governance, deepening social inequalities, environmental sustainability and intergenerational justice. It is a field fundamentally defined by intersections between private entrepreneurship and public policy, and, as elaborated by Giovanni Sartor in an insightful paper circulated before the conference,[1] a field which gives rise to fundamental questions about the delineation and interplay between ethical evaluations and legal regulation.

My modest contribution to the present debate will be to explore the issues at hand from the perspective of international human rights law, or – on a slightly less ambitious note – to trace the contours of a possible human rights-based approach to AI. Some significant engagements and developments in this field that inform my presentation are: an earlier involvement in elucidating the normative implications of the International Covenant on Economic, Social and Cultural Rights (ICESCR) Art. 15(1)(b), which affirms the right of everyone 'to enjoy the benefits of scientific progress and its applications' – this led to the adoption of the *Venice Statement on the Right to Enjoy Benefits of Scientific Progress and its Applications* (2009);[2] a more recent examination of intersections between science and human rights (cf. CESCR General Comment 25 on science and economic, social and cultural rights[3]); and current developments in the area of Business and Human rights which address the question of engaging Non-State Actors in the protection and promotion of recognised international human rights.

Relevant standards and developments in the field of human rights

My starting point, as anticipated by way of introduction, will be to take the established international normative framework for granted. Certainly, this can be subject to questioning and criticism (I have personally been anchorperson on a recent Global Campus of Human Rights podcast series on 'engaging with human rights skepticism'[4]), but that is not our primary concern in the present context. Our focus is rather how generally agreed international human rights standards apply to and implicate developments in the area of AI. It may in this connection be relevant to recall that the main United Nations human rights treaties (ICCPR and ICESCR) have each been signed and ratified by approximately 170 states; they have, in other words, been voluntarily embraced by the great majority of members of the international community. This, without contest, is the closest we come to a common international ethical and legal standard, which is further reinforced by several other UN human rights treaties and by powerful parallel human rights frameworks at regional level in Europe, Africa and the Americas.

In setting about to explore the implications of such normative standards for the regulation of AI, an important point of reference is the different types of government obligations related to human rights. It is now common practice for experts to distinguish between obligations

to 'respect', 'protect' and 'fulfill'. The obligation to respect requires States to not violate established human rights standards. This is a negative obligation fundamentally restricting the abuse of government power. The obligation to *protect*, by contrast, is a positive obligation of States to adopt legislative measures to prevent third parties from violating or adversely impacting the enjoyment of rights of others. This constitutes a central focus of the emerging international business and human rights normative framework. The obligation to *fulfill*, finally, is a positive obligation for States to adopt broader policies, facilitate awareness-raising and infrastructure changes, etc., aimed at gradually enhancing and expanding the enjoyment of rights in the given area. This is a less clearly defined typology of obligations, more open to creative policy engagements. In procedural terms, it typically requires meaningful stakeholder consultation and participation in decision-making.

This tripartite taxonomy defines government obligations related to human rights, and governments in turn impose both legal and extra-legal (ethical) obligations on private actors. A common societal commitment to universal human rights and related objectives (including the Sustainable Development Goals), by extension, requires all of us to affirm and help foster a culture of respect for everyone's fundamental rights. This commitment may be sanctioned by law but is essentially ethical in nature.[5]The *UN Guiding Principles on Business and Human Rights*, which were unanimously adopted by the UN Human Rights Council in 2011,[6] in the same spirit requires corporate enterprises to undertake a 'human rights due diligence' assessment of all planned and ongoing business activities. A draft UN treaty currently under negotiation, and a related draft EU directive on 'mandatory due diligence', both impose a formal/procedural obligation on business enterprises to take stock of any possible adverse human rights impacts, and it is assumed that this in turn will facilitate an improved record of substantive human rights compliance (going beyond the minimum prohibitions and obligations defined by law). The key premise of my presentation today is that a similar formal requirement to undertake meaningful human rights due diligence should be extended to all significant AI applications, whether public, private or a combination of both. This will give structure to the amorphous field of 'ethical' AI guidance/regulation and will render the latter consistent with the existing normative framework regulating public administration and private enterprise.

Mapping intersections between AI and human rights

A very interesting approach to identifying human rights risks and implications in the field of AI is found in the Dutch *Fundamental Rights and Algorithm Impact Assessment* tool (FRAIA), published in March 2022.[7] To facilitate an overview of the issues at stake, FRAIA distinguishes between four main areas of concern, or, phrased differently, four relevant clusters of rights. These are:

1. Fundamental rights relating to the person (including a number of social and economic fundamental rights)

2. Freedom-related fundamental rights

3. Equality rights

4. Procedural fundamental rights

Each primary cluster is further associated with more specific sub-clusters of rights, which are detailed in an annex to the FRAIA framework, and the aim of an impact assessment is to determine specifically 'which sub-clusters an algorithm affects or may affect.' Methodologically, '[t]he idea is to go through the explanations and the clusters and note down which fundamental rights may be affected by the use of the algorithm.' Clearly this makes for a thorough and well-informed human rights impact analysis, but a potential downside of the approach is that it may be perceived as excessively demanding and cumbersome by practitioners not specifically trained in the area of human rights and for whom this field of inquiry remains secondary to the objectives driving the original engagement with AI. Experiences shared by some Z-Inspection focal groups echo this concern. An important

forward-looking challenge will therefore be to devise relatively simple and intuitive, yet comprehensive templates for human rights impact assessment.

In taking stock of the wider field of intersections between AI and human rights, I wish to suggest four levels of consideration that should be taken into account. These are:

1. Possible *direct* adverse human rights impacts; this, in fact, is the exclusive focus of the FRAIA assessment tool.

2. Possible *indirect* adverse impacts in the form, e.g., of reinforcement of existing inequalities, patterns of structural discrimination, and further marginalization of disadvantaged social groups, etc., due to algorithmic biases and to a gradual remodeling of work, employment, and social and economic access.

3. Capacity of AI to facilitate and *positively contribute* to the (progressive) realization of human rights (e.g., in the health sector) and other related societal objectives as, e.g., defined by the Sustainable Development Goals.

4. Fundamental challenges posed by AI to some of the core underlying premises of normative reasoning such as, notably, the concepts of human dignity, agency and autonomy, free will, intentionality and accountability. (One may note that a similar conceptual challenge is being posed by the discourse of climate justice and nature rights, which confronts us with a deeply problematic anthropocentric bias in our cultural and intellectual heritage, as in fact is manifest in the normative framework centered on human rights.)

When further elaborating this analytical approach, it will be relevant to adopt both an *upstream* and a *downstream perspective* on AI impacts, and a further essential human rights requirement will be to ensure meaningful *stakeholder consultation* and *participation* in the planning and monitoring of initiatives with far-reaching social consequences.
A human rights-based approach to AI, as advocated here, may be expected to shape government policy and regulation, and may, at the same time, be seen as complementary to other ethical perspectives on AI. I have argued elsewhere[8] that a human rights approach can inform and enrich the wider field of professional ethics, and ethical deliberations generally, for example by qualifying the principles of *non-maleficence* ('do no harm'), *beneficence* (contribute to a greater social good), and what concretely is implied by principles of *autonomy and dignity of the human person*. These, clearly, are central to our deliberations.

.....

[1] Sartor, G. 'Artificial intelligence and human rights: Between law and ethics', *Maastricht Journal of European and Comparative Law* 2020, Vol. 27(6) 705–719
[2] https://unesdoc.unesco.org/ark:/48223/pf0000185558
[3] https://digitallibrary.un.org/record/3899847
[4] https://gchumanrights.org/education/networking-and-outreach/to-the-righthouse-podcast-sessions.html
[5] Ulrich. G. "The Statement of Ethical Commitments of Human Rights Professionals: A Commentary", in M. O'Flaherty and G. Ulrich (eds.). The *Professional Identity of the Human Rights Field Officer*. Ashgate Publishing Ltd. 2010
[6] https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf
[7] https://www.government.nl/documents/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms
[8] Ulrich, G. and Wainwright, T., 'Human Rights and Professional Identity', in Polli Hagenaars, Marlena Plavšić, Nora Sveaass, Ulrich Wagner, and Tony Wainwright (eds.), *Human Rights and Human Rights Education for Psychologists*, Routledge 2020

Prof. George Ulrich

Prof. George Ulrich is currently EIUC Academic Director and Programme Director of the European Master's Degree in Human Rights and Democratisation (EMA). He held the position of Rector and Professor of Human Rights at the Riga Graduate School of Law from 2009-2016. Prior to this, he served as EIUC Secretary General from 2003-2009 and as Academic Coordinator / Programme Director of EMA from 2001- 2004. From 1999-2001 he was Senior Researcher at the Danish Centre for Human Rights and from 1996-1998 Research Fellow at the Institute of Anthropology, University of Copenhagen, and visiting researcher at Makerere University, Kampala, Uganda.

AI, Ethics and Human rights.

Giovanni Sartor

The ethics and law of AI address the same domain, namely, the present and future impacts of AI on individuals, society, and the environment. Both consider the extent to which AI may enhance or constrain individual and social initiatives and contribute to or detract from valuable individual and social interests. Both are meant to provide normative guidance, proposing rules and values on which basis to govern human action and determine the constrains, structures and functions of AI-enabled socio-technical systems. This raises the issue of how to deal with the demands of ethics and law, which may and should indeed converge, but occasionally may pull in indifferent direction.

The law may have failed to adapt to ethical requirements, for instance, not having been able to cope with technological and social development. As a consequence, behaviour that should ideally be prohibited (e.g., facial recognition in public spaces) may be considered legally permissible, or behaviour that should be permissible (e.g., processing personal data for the purpose of medical research) may be legally prohibited.

An important connection between morality and law is provided by human rights. I believe, following Amartya Sen, that human rights are primary ethical demands. They concern freedoms broadly understood as opportunities for individuals. Such opportunities include both negative liberties — which mainly require non-interference from governments and protection from interference by third parties (as in the case of freedom of movement or freedom of expression)—and positive liberties, which require the active provision of resources (as in the case of social rights, for example, rights to education or health). However, human rights have also a legal dimension, i.e., certain important aspects of ethical human rights are also recognised in binding international, regional and national legal instruments, creating enforceable legal obligations for states and other actors.

Thus, significant overlap exists between (different constructions of) legal and ethical rights, but that the two dimensions are distinct. In particular, it may be the case that certain aspects of the ethical human rights are not legally implemented. This may happen because the law wrongly fails to appropriately enforce ethical standard that it should implement, but it may also happen because the law rightly does not enforce aspects of ethics that are better left to voluntary initiatives.

This distinction between ethics and law does not exclude that the two dimensions may influence each another. Ethico-political arguments can be advanced concerning the need that an ethical right (or aspects or implications of it) should, or should not, be legally recognised, and that the law should change accordingly. Ethical arguments can also be deployed to support the interpretation/construction of legal sources and may thus contribute to determining the way in which the law is applied. On the other hand, ethics can learn from the law, which takes institutional approaches to normative issues, is expressed in publicly accessible sources and in critical commentaries on them and contains vast examples of how (the norms extracted from) such sources are applied to concrete cases. Consider, for instance, how general ideas supporting an ethical right to privacy or an ethical right to free speech and to protection from discrimination have been translated into corresponding legal rights set forth in legislation and upheld in a vast case law.

The continuum between ethics and law is borne out by the fact that when speaking of the impact of AI on broadly scoped rights, such as privacy or freedom of expression, or on collective values, such as democracy, public discourse, health, or culture, there is often no reference to any specific ethical theory or municipal law, but rather to a cluster of issues, claims, and concepts pertaining to different ethical approaches and different international, regional, or national legal systems.

This multiple reference of the rights' language should not be condemned, as it contributes to the richness of the normative debate on the impacts of AI and should be combined with the ability to draw the necessary distinction when needed.

Thus, lawyers should not be worried when the language of rights and values is deployed by ethicists, as when the term 'human rights', or terms such the 'right to autonomy', the 'right to privacy', or 'dignity' appear in documents on the ethics of AI. However, lawyers should refrain from translating ethical claims directly into legal claims, nor should ethical claims be misconstrued as legal claims or rejected for not being affirmed by existing laws. Similarly, ethicists should not be too impatient when lawyers are slow or reluctant to incorporate, into the law, ethical claims concerning present and prospective uses of AI.

Finally, neither ethics nor law should be viewed as functionally equivalent, namely, as interchangeable substitutes in the regulation of AI. It has indeed been observed the enthusiasm of the major commercial players for ethical charters may be motivated by purpose of preventing the enactment of binding laws governing their activity, and consequent institutional controls.

We need thus to firmly assert that the law is needed whenever only a coercible public response can effectively counter abuses and misuses of AI, as well as when the allocation of public funds, and the deployment of governmental resources has to be directed to support the creation and accessibility of valuable technological solutions. Thus, the adoption of ethical guidelines by private actors does not exempt them from being subject to old and new legal constraints. Similarly, even under an adequate legal regulation of AI, still it makes sense to develop ethical frameworks, to guide the legally permissible uses of AI toward socially beneficial outcomes, and to support the application and evolution of the law.

Giovanni Sartor
Giovanni Sartor is professor in Legal Informatics at the University of Bologna, professor in Legal informatics and Legal Theory at the European University Institute of Florence, visiting professor of Artificial Intelligence and Law at the University of Surrey. He holds the ERC-advanced grant (2018) for the project Compulaw (2019 – 2025).

He obtained a PhD at the European University Institute (Florence), was a researcher at the Italian National Council of Research (ITTIG, Florence), held the chair in Jurisprudence at Queen's University of Belfast, and was Marie-Curie professor at the European University of Florence. He has been President of the International Association for Artificial Intelligence and Law. He is co-director of Summer Schools on "Artificial Intelligence and Law" e on "Law and Logic". He hold courses at the University Boccani (Milan), the university Catolica (Lisbon) and Surrey (London)
He has published widely in legal philosophy, computational logic, and computer law, AI & law. He is co-director of the Artificial Intelligence and Law Journal and co-editor of the Ratio Juris Journal. His research interests include legal theory, early modern legal philosophy, logic, argumentation theory, modal and deontic logics, logic programming, multiagent systems, computer and Internet law, data protection, e-commerce, law and technology.

Panel on AI and human rights

Frédérick Bruneault

The question of the relationship between AI and human rights raises the important issue of the relationship between ethics and law. If AI is disruptive in the interpretation and application of the law, the dual aspect of human rights (both positive law and philosophical concepts) allows us to think about the ethical implications of the latter. Insofar as ethics can play different roles in relation to law, it is important to clearly define its different functions, which will at the same time allow us to circumscribe the more precise terrain of the discussion. To do so, the conceptual tools proposed in an article by Luciano Floridi on digital governance, published in 2018, will be used. This discussion is part of the theoretical framework of information ethics developed by Floridi in his 2013 book on the *Ethics of Information*, which also constitutes a novel contribution to the debate in AI ethics, notably because it addresses some of the shortcomings of classical ethical frameworks (Bruneault and Sabourin Laflamme, 2021, 2022). Given the limitations of this presentation, we will retain for the purposes of the discussion only a conceptual distinction proposed by Floridi (2018). For the latter, because of the profound transformations that digital technology in general and AI in particular are inducing in our ways of functioning and also because of the many risks associated with them, it is imperative that we develop an adequate normative framework for these technologies, which will undoubtedly occupy an increasing place in the information societies in which the generations that will follow us will evolve. Although these considerations on digital governance are broader in scope than our object of study, i.e., the governance of AI and the Z-Inspection, they are nonetheless very useful in identifying their components and situating them in relation to each other.
First, for Floridi, digital governance should not be considered as a synonym for digital regulation. In fact, such an adequacy would be a fallacious synecdoche in which the part (regulation) would be unduly substituted for

the whole (governance), thus masking an essential part of what Floridi calls „the normative map“ of digital governance, namely digital ethics. Indeed, digital governance is not just about making laws and regulations: it must also include the moral evaluation of the issues associated with these technologies, with the aim of proposing specific solutions to the problems under analysis, which also directly links his approach to the pragmatist tradition.
According to Floridi, even the presence of appropriate legislative mechanisms would not be sufficient to adequately regulate the development of digital technology and AI, as these mechanisms are limited to determining what is legal and illegal, without questioning the avenues that would be more desirable to follow with regard to technological development. It is of course ethics, drawing on the rich conceptual heritage of moral philosophy, that can assume this function.
For Floridi, the ethics of the digital world can be expressed in two ways in relation to the law: what he calls hard ethics and soft ethics. By hard ethics, he means the discussion on the duties and moral responsibilities of each individual and, more generally, the reflection on the principles and values that should guide moral action. Thought in its relation to law in the governance of digital technology, the function of hard ethics, situated at a higher level of abstraction, is to define the principles that should guide legislative reforms aiming at better framing the conception and deployment of digital technology and AI in society, as well as to question the moral validity of the legislative framework in place. The objective of hard ethics is then to evaluate the coherence of existing laws with the identified ethical principles and to pronounce on their relevance or on the potential need to reform them. In this sense, situated in a way upstream of the law, hard ethics could be likely to influence the legislator‘s orientations and thus indirectly shape the law. Soft ethics, on the other hand, is situated downstream of the law, i.e. it is interested in ethical questions that go beyond the field covered by the regulations and seeks, for example, to determine, through ethical assessment processes, what technological developments are desirable and what are not, beyond what the law permits or prohibits. Soft ethics is therefore a practical exercise in the ethical assessment of specific technological devices in concrete situations.
These reflections can also be carried out on the basis of factual or empirical analyses, by means of ethical risk assessment processes in professional environments or decision support tools (Floridi and Strait, 2020). An assessment of ethical risks and impacts conducted in an appropriate manner from the conception of technological devices is likely to guide their development, and consequently to have a real and effective normative effect on practices, for example where the law is silent or absent. Soft ethics, understood as a source of social regulation, is here clearly envisaged in a pragmatic perspective and can be thought of as a source of normativity complementary to law within the framework of normative pluralism. Even if we have to keep in mind that these two functions of ethics cover the same normative ground and that the interactions between soft and hard ethics can be of several kinds, the interest of the distinction proposed by Floridi is that it allows us to identify and situate two distinct normative functions of ethics in relation to law and human rights in AI governance.

Prof. Frédérick Bruneault
École des médias, Université du Québec à Montréal and Philosophie, Collège André-Laurendeau, Canada

I did my PhD in philosophy at both the University of Ottawa and the Université Paris-Sorbonne (Paris IV) on the ethics of technology.
I have been teaching philosophy since 2011.
I am researcher at the GRISQ (Groupe de recherche sur l'information et la surveillance au quotidien / research group on information and surveillance) and also at OBVIA (International Observatory on the Societal Impacts of AI and Digital Technology).
I am currently conducting two researches which received funding in Québec: 1- a research on the teaching of AI ethics at the higher-education level and 2- a research on student digital literacy.
I am the founder of LEN.IA, the AI & Digital Ethics Lab (lenia.net).
Frédérick was awarded a Z-inspection® Teaching Certificate.

Comparing the Trustworthy AI assessment process with the fundamental rights-based FRAIA assessment tool

AI Assessment: On the Role of Ethics and Fundamental Rights

Elisabeth Hildt
Illinois Institute of Technology and L3S Research Center Hannover (ehildt@iit.edu)

A Z-Inspection® working group was involved in the pilot project "Responsible use of AI" with Rijks ICT Gilde and the Province of Friesland, the Netherlands. The working group used two different assessment approaches: the fundamental rights-based assessment outlined by the document "Fundamental Rights and Algorithms Impact Assessment (FRAIA)", and the Z-Inspection Trustworthy AI assessment based on the European Ethics Guidelines for Trustworthy AI.
The FRAIA document suggests a procedure for assessing AI tools from a fundamental rights perspective that identifies whether an AI tool affects fundamental rights and, if so, facilitates a structured discussion about opportunities to prevent or mitigate this interference. The document defines four main clusters of fundamental rights: Fundamental rights relating to the person; Freedom-related fundamental rights; Equality rights; and Procedural fundamental rights.

.....

https://z-inspection.org/general/pilot-project-assessment-for-responsible-artificial-intelligence-together-with-national-ict-guild-and-the-province-of-fryslan-the-netherlands/

https://www.government.nl/documents/reports/2021/07/31/impact-assessment-fundamental-rights-and-algorithms#:~:text=The%20

Fundamental%20Rights%20and%20Algorithm,deployment%20of%20an%20algorithmic%20system

https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

.....

In contrast, the Ethics Guidelines for Trustworthy AI are based on four mid-level principles: Respect for Human Autonomy; Prevention of Harm; Fairness; and Explicability.
The guidelines describe seven key requirements closely connected to these ethical principles: Human agency and oversight; Technical robustness and safety; Privacy and data governance; Transparency; Diversity, non-discrimination, and fairness; Societal and environmental well-being; and Accountability.
Broadly speaking, the FRAIA assessment relies on fundamental rights, whereas the Trustworthy AI assessment is based on ethical principles.

Here is a short reflection on what we learned from the two assessment approaches:
The fundamental rights assessment and the ethics assessment based on the Trustworthy AI guidelines go hand in hand; both approaches provide critical insights with regard to the AI use case. Reflecting on AI from an ethics perspective clearly overlaps with a fundamental rights assessment. Both ethics and fundamental rights are about norms and fundamental values held in society. As ethics reflection and ethics guidelines influence law, scholars from both fields must work together when thinking about the shaping of technology and its societal implications.
Even though there are great similarities, there are several considerable differences between the two approaches.
Ethics, a branch of philosophy, reflects on what is right and wrong. It seeks to find an answer to questions like “What are we to do?” or “What is the right action?”. In the context of AI applications, an ethics-based approach addresses questions like “What is the right way to design, develop, deploy, and use this type of technology so that it is beneficial for individuals and society”? Questions like these require thinking about the various alternatives for action around an AI application, and involve reflecting on the various options and their potential implications without confining the reflection to those options in line with existing law.
A fundamental rights-based approach is more closely linked to existing law and focuses on aspects that are legally relevant and thus enforceable. Compared to this, an ethics-based approach is much broader and also more open to reflection on potential implications that may not be worth considering from a legal perspective. For example, from an ethics perspective, personal autonomy, freedom of decision-making, and fairness were found to be concepts of clear relevance in the context of the pilot project’s AI tool, whereas, from a rights-based perspective, rights related to personal autonomy in a strictly legal sense were considered not infringed by the AI tool.
While a fundamental rights-based assessment focuses on whether fundamental rights are negatively affected or infringed, from an ethics perspective, both positive and negative implications of AI technology are considered. For example, in the pilot project “Responsible use of AI”, the potential positive implications of the AI tool on the environment proved to be central. This implies the question of whether the right to a healthy living environment may or may not be positively affected by the AI tool.
Furthermore, a fundamental rights-based approach towards protecting the environment is clearly anthropocentric, as can be seen from the wording “right to a healthy environment”. The FRAIA document lists the right to a healthy living environment in the cluster “Rights related to the person”.
In contrast, an ethics-based perspective allows us to bring in biocentric or pathocentric perspectives and address biodiversity-related issues. Also, from an ethics perspective, questions of how to adequately consider future generations can be tackled more easily than from a fundamental rights-based approach.
Overall, the fundamental rights-based approach clearly funnels and constrains the aspects, questions, and issues to be discussed around the AI use case. For example, issues related to transparency or human agen-

cy and oversight can only be addressed in the context of the right to good administration, even though transparency and human agency and oversight are clearly relevant in other contexts as well. As discussed above, similar problems arise in the context of the right to a healthy living environment. Approaching the use case from a fundamental rights perspective implies that ethical and societal aspects and implications of AI are discussed only insofar as they are related to fundamental rights and existing law.

Elisabeth Hildt *is a Professor of Philosophy and the Director of the Center for the Study of Ethics in the Professions at the Illinois Institute of Technology in Chicago, and a guest professor at the L3S Research Center of the University of Hannover. She is interested in philosophical and ethical aspects of science and technology. Elisabeth Hildt has been a Z-Inspection® team member for almost three years, contributing to four use cases so far.*

Trustworthy AI – beyond a human rights assessment?

Emilie Wiinblad Mathez

The views expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations.

In this short presentation I will share some reflections on three points related to ethics, human rights and AI from the experiences with the Z-Inspection® process of assessing trustworthy AI.
These reflections are based on participation in two Z-Inspection use cases [1] , one of which included integration of the fundamental rights framework of the Netherland in the assessment. I also draw on my experience with organizational ethics and human rights. The following three questions arose during the assessment.

1. **Human Rights - framing:** How to frame the human rights assessment? Should this be considered differently from an ethical assessment, or the assessment of trustworthiness?

2. **Assessment - the aims:** What should be the aim or scope of the fundamental rights assessment? Is it to ensure that rights are not infringed or violated only, or does it go beyond to look at how rights are affected more broadly, included, protected or promoted?

3. **Trust and decisions about the AI system:** Some ethical issues impacting trustworthiness of the AI system concern the decisions about the AI system more broadly.

Framing the human rights assessment within the use case

In the pilot project in Friesland, the Z-Inspection use case concerned an AI system to be used by a commune in the Netherlands. In March 2022, the Dutch Ministry of Interior and Kingdom Relations issued an Impact Assess-

ment tool for Fundamental rights and algorithms (FRAIA) as a “discussion and decision-making tool for government organisations”. The tool sets out the questions which must be answered when a government organisation considers “developing, delegating the development of, buying, adjusting and/or using an algorithm.” The tool considers three decision-making stages for an AI system and asks that in all stages, respect for fundamental rights must be ensured.
For the Z-Inspection use case in Friesland this tool and framework was included on a pilot basis. The FRAIA is a tool, which in many ways is at the forefront when it comes to ensuring that government organizations live up to human rights obligations and commitments. It foresees a multi-disciplinary and holistic approach.

While the tool was useful and clear, the question about how to frame the human rights assessment nevertheless arose and more specifically, how to consider the fundamental rights as part of an assessment of trustworthiness and ethical reflections on and AI system. Should we consider the rights as they are defined in law and interpreted through the courts only? Or should the rights be considered more broadly, as part of the assessment, linking the rights to ethical principles beyond their narrower legal definition? If only the legal definitions are used, an assessment of whether specific legislation applies would be required. This may on the one hand have the advantage of ensuring adherence to existing human rights definition and legislation, while on the other hand be narrow in scope, and therefore risk missing broader ethical question. It might also require specialized legal expertise and could risk excluding other perspectives. In the use case, we adopted a hybrid approach, by first identifying which rights in the framework could be affected, using the legal definitions as part of the argument of why the rights were engaged; before turning to the broader ethical issues related to the rights. In this way, a more open reflection was possible, integrating the fundamental rights framework as part of a broader ethical assessment. This approach was used as the Z-Inspection is not aimed at assessing legal compliance.

If the human rights assessment is defined too narrowly it risks being an assessment separate from the ethical assessment, or the assessment of trustworthiness. If it is too broad, the human rights standards risk being watered down. A two-tiered, integrated approach, looking both at legal requirements and the broader ethical questions, could be envisaged, depending on the organizational set up and use case.

Considering if rights are infringed, protected or promoted?

Frequently raised issues regarding AI and human rights concern how personal data is handled and used in the AI system, right to privacy; or how the data might be biased and can lead to discrimination. The Dutch use case did not concern personal data. The FRAIA suggests that the right to equal treatment, protection of personal data, procedural and good administration rights should always be considered, as these are usually affected by an AI system. However, fundamental rights may also be infringed or affected by the implementation, use, or application of the algorithm, by the context in which the algorithm is used, or by the decisions and measures that are linked to the output of the algorithm. This was highlighted in the FRAIA and is at the centre of the Z-Inspection process which assesses trustworthiness based on the socio-technical scenario.

The FRAIA considers as a first step the identification of fundamental rights which may be affected, or threatened, by the AI system and then a balancing of the seriousness of the rights infringement with the importance of the objectives and the necessity to use the AI to reach the objective. The FRAIA also includes a framework for assessing seriousness. In the practical application of the FRAIA in the Z-Inspection use case two adaptations were made. Firstly, it was decided to also include, in the assessment, the rights which the AI system aimed to affect positively, i.e., the right to a healthy living environment. It was found that the objective of the AI system was largely to promote this fundamental right and that it was useful to include this perspective in the assessment. The objective of promoting a right

was consider as “a claim” in the assessment, rather than a fact, and as such arguments and evidence for the claim were discussed. Secondly, the FRAIA sets out a four-tier framework for assessing seriousness of the rights infringement. This was complex to use and was replaced, in the use-case, by considerations and suggestions for how risks of any infringements could be mitigated. Taking this broader approach to consider if rights were potentially infringed, or sought to be protected or promoted by the AI system was found useful.

Ethical issues arising from the decisions about using an AI system.

The FRAIA asks that the questions about “why an AI system” and “what the system should do” are answered before the question on “how will the AI will do this” is answered. The “why” questions in the tool are questions like “What are the reasons, the underlying motives, and the intended effects of the use of the algorithm? What are the underlying values that steer the algorithm’s deployment?”
The tools provided, in support for answer this, are directly linked to guidance and the parameters for *good governance in policy and law making* [2]. Central to answer this are questions like “who are involved in the problem formulation and the articulation of the solution?” In other words, the ability for an organization, or government, to design, develop, deploy and use trustworthy and human rights aligned AI will be directly affected by its overall ability to answer such questions as part of governance, accountability and pursue of legitimate aims. This in turn is directly linked to how power is distributed and managed both within the organization and with its stakeholders and collaborators. As the experience from the Z-Inspection use-cases shows, this requires not only clarity in the problem formulation, the objective and goal statements, and in the reasons given for why an algorithm is most useful, but also a process to answer this. What came out in the assessment is that, to generate trust, such a process must reflect different viewpoints, ensure accountability and be legitimate and transparent to its stakeholders, both internally and externally. In practice, it may often be unclear how this will be ensured, and what is expected of an AI system developer who is asked to find a solution to a problem which lacks these qualities and legitimacy. However, in the context of the Z-Inspection process the importance of clarity in the problem statement and objectives came to the fore, as did the anthropocentric approach implied in human rights. Experiences from the use-cases indicate that for AI systems to be trustworthy it is important to have

· Clarity in the problem formulation - “what problem do *we* want the AI system to help solve”.

· Have shared values and ethics reflected in the “why” this is a problem and the suggested solution.

· Legitimacy of the “we”. In other words – “do those who are deciding and were consulted on the problem and solution have legitimacy in the eyes of those affected by the decisions both in the short and longer term? And how are they held to account?”

One aspect of this is ensuring clarity on how ethical issues are identified and dilemmas solved.

It follows that where decision-making processes or organizational governance are weak, the decisions related to the design, development, deployment and use of an AI system risk undermining trust or posing ethical questions beyond human rights assessment.

Emilie Wiinblad Mathez

.....

[1] Assessing Trustworthy AI in times of COVID-19. Deep Learning for predicting a multiregional score conveying the degree of lung compromise in COVID-19 patients and the Pilot Project: Assessment for Responsible Artificial Intelligence together with Rijks ICT Gilde -Ministry of the Interior and Kingdom Relations (BZK)- and the province of Fryslân (The Netherlands).
[2] The integrated impact assessment framework for policy and legislation a practical guide

Emilie Wiinblad Mathez
United Nations High Commissioner for Refugees, Geneva, Switzerland.
Emilie has a Master's in International Law and a MBA in cross-cultural management, with 20+ years of experience working with international law and policy, including refugee, humanitarian and human rights law. Her experience includes legal analysis, policy and strategy development, research and evaluation, advocacy, communication and learning as well as project management and ethics advice and guidance in an international organization. She has led the development and implementation of a value-based approach to ethical decision-making as part of organizational change. She has worked in countries in Europe, Asia and Africa.

She has been part of the Z-inspection® since 2020 and was awarded a Z-inspection® Teaching Certificate in 2022.

International Data-Based Systems Agency IDA [1]

Peter G. Kirchschlaeger

1. Digital transformation and so-called "artifi cial intelligence (AI)" – which can more adequately be called *"data-based systems (DS)"* [2] – comprise ethical opportunities and ethical risks. DS can be powerful for fostering human rights but also for violating human rights. Elon Musk warns: "AI is far more dangerous than nukes [nuclear warheads]. Far." [3] Stephen Hawking points out: "Unless we learn how to prepare for, and avoid, the potential risks, AI could be the worst event in the history of our civilization. It brings dangers, like powerful autonomous weapons, or new ways for the few to oppress the many. It could bring great disruption to our economy. [4]" Therefore, it is necessary to identify ethical opportunities and ethical risks as well as opportunities for promoting human rights and human rights risks precisely and at an early stage in order to be able to benefit sustainably from the opportunities and to master or avoid the risks. In the avoidance and mastering of risks, technology-based innovation can in turn play an essential role.

2. Allowing humans and the planet to flourish sustainably and guaranteeing globally that human rights are respected not only *offline* but also *online* and in the digital sphere and the domain of DS requires the following concrete measures:

A. human rights-based data-based systems (HRBDS): Human rights-based data-based systems (HRBDS) means that human rights serve as the basis of digital transformation and DS, e.g., the human rights to privacy and data-protection must be respected. HRBDS exclude the possibility that humans should be able to sell themselves and their data as well as their privacy as products. One possible solution to foster innovation and to make data use legitimate in accordance with the right to privacy and data protection would be the "purpose-driven data use" approach. The "purpose-driven data use" approach starts from the right to privacy and data-pro

tection as a prerequisite and respects this right. In automated driving, for example, people must identify themselves with their data and enter their location and destination in order to enjoy automated driving at all. But this data is only provided to enable the driving process. It may neither be used for other purposes nor sold on to third parties. The users also do not have the option of selling this data themselves (e.g. to obtain a discount). Beyond that, fully anonymized data may only be kept for the optimization of the collective automated mobility with the informed consent of the users.
To illustrate this approach in its feasibility, the following analogy serves: when one goes to the doctor, one also shares personal data so that the doctor knows who she has in front of her and tells her about one's illness in order to hopefully experience relief from suffering as well as healing, without either the doctor being allowed to resell this data or the patient being offered to sell this data in order to receive better medical treatment. The doctor may also keep the patient's file with the medical history strictly confidential – exclusively for the purpose of better treatment of the patient. It is also possible to share completely anonymized data for research purposes if the patient gives informed consent to this sharing.

B. an International Data-Based Systems Agency (IDA): An International Data-Based Systems Agency (IDA) needs to be established at the UN as a platform for technical cooperation in the field of digital transformation and DS fostering human rights, safety, security, and peaceful uses of DS as well as a global supervisory and monitoring institution and regulatory authority in the area of digital transformation and DS.

The establishment of the IDA is realistic because humanity has already shown in its past that we are able to not always "blindly" pursue and implement the technical possible, but also to limit ourselves to what is technically feasible when the welfare of humanity and the planet are at stake. For example, humans researched the field of nuclear technology, developed the atomic bomb, it was dropped several times, but then humans substantially and massively limited research and development in the field of nuclear technology, in order to prevent even worse, despite massive resistance. This suppression was successful to the greatest possible extent, thanks to an international regime, concrete enforcement mechanisms, and thanks to the International Atomic Energy Agency (IAEA) at the UN.

Prof. Dr. Peter G. Kirchschlaeger
Peter G. Kirchschlaeger is Ethics-Professor and Director of the Institute of Social Ethics ISE at the University of Lucerne (Switzerland). He is Research Fellow at the University of the Free State, Bloemfontein (South Africa). Prior, he was Visiting Fellow at Yale University (USA).
2011-2015, he was member of the Board of the Swiss Centre of Expertise in Human Rights, 2013 Visiting Scholar at the University of Technology Sydney (Australia), 2013-2014 Guest-Professor at the Faculty of Theology and Religious Studies at the Katholieke Universiteit Leuven (Belgium), 2013-2017 Fellow at the Raoul Wallenberg Institute of Human Rights and Humanitarian Law (Sweden), 2015-2019 Guest-Lecturer at the Leuphana University Lueneburg (Germany).
He is a consultative expert in Ethics of national and international institutions and organizations (companies, NGOs, …), among others, member of the Swiss Federal Ethics Committee on Non-Human Biotechnology (ECNH).
In his research, he focuses on ethics of digital transformation and ethics of artificial intelligence. His latest book «Digital Transformation and Ethics» was published in 2021 (Nomos).

[1] These points based on a multi-year research-project started at Yale University and finalized at the University of Lucerne published in the book “Digital Transformation and Ethics: Ethical Considerations on the Robotization and Automation of Society and the Economy and the Use of Artificial Intelligence” (Nomos: Baden-Baden 2021, 537 pages)” by Peter G. Kirchschlaeger.

[2] Peter G. Kirchschlaeger, Digital Transformation and Ethics: Ethical Considerations on the Robotization and Automation of Society and the Economy and the Use of Artificial Intelligence. Nomos: Baden-Baden 2021, 102-106.

[3] Clifford, Catherine (2018) Elon Musk: “Mark my words – A.I. is far more dangerous than nukes”. In: CNBC, March 13. Online: https://www.cnbc.com/2018/03/13/elon-musk-at-sxsw-a-i-is-more-dangerous-than-nuclear-weapons.html [25.01.2023].

[4] Kharpal, Arjun (2017): “Stephen Hawking says A.I. could be ‘worst event in the history of our civilization’”. In: CNBC, November 6. Online: https://www.cnbc.com/2017/11/06/stephen-hawking-ai-could-be-worst-event-in-civilization.html [3.2.2023].

There are no dilemmas in AI ethics

James Brusseau

Key finding
Conventional AI ethics is being disrupted by the interdisciplinary approaches pioneered in the Heart Attack, Skin Lesion, and Brescia CXR cases. Because philosophers are now integrating with information engineers and related domain experts, ethics is no longer confined to producing dilemmas and restrictions. Instead, doing ethics now means overcoming those impediments by contributing to increasingly rapid innovation.

Narrative
Conventionally, artificial intelligence ethicists like me have been able to act only through restrictions. Lacking technical knowledge and skills, our influence over innovation has been limited to proposing limitations. The result has been guidelines, regulations, and prohibitions (Guidelines Trustworthy AI, GDPR, AI Act, CCPA, New York City Local Law 144).
Starting from the Heart Attack case and going forward, a strategy has developed for integrating ethical and technical contributions. The key is to force all ethical discussion to fit within the language and principles of the EC Ethics Guidelines for Trustworthy AI. This constriction raised objections and withdrawals from both the humanist and the technical sides. But, there was a reward: the emergence of a viable process for collaboration across disciplines.
The collaboration allows a reorientation. Instead of being limited to describing concerns and dilemmas about innovation, and then responding by prescribing caution and hesitation, the ethicists role can be redefined this way: locate directions for AI innovation that solve the humanist problems AI creates. Because ethicists are linked with information engineers, they – we – are empowered to respond to AI harms with more AI.

The Brescia CXR case illustrates this shift, but the skin lesion case is paradigmatic. There, a traditional dilemma appeared. AI analyses of skin lesions outperformed human diagnoses, but the results could not be explained. So, there was a tradeoff, a human dilemma: accuracy or explainability. One or the other. Along with the dilemma came the possibility of slowing the AI, or restricting its potential since explainability can be increased by decreasing model complexity. But, the dilemma collapses when the same innovative force that created the accurate AI skin analyses is turned toward resolving the question of explainability. This is exactly what happened in the skin lesion case where an AI tool was developed to wrap around the skin analysis model and interpret the process. When that explanation happened, AI innovation solved the ethical problem created by AI innovation.
What we have learned from our cases – what we have learned by being among the few who have done AI ethics instead of just theorizing about it – is that ethics in artificial intelligence is not about forming humanist objections, and not about describing dilemmas, and not about issuing warnings and restrictions. Instead, it is about finding problems and naming dilemmas *in order* to overcome them by speeding advance. Because the reason for problems and dilemmas is to provide directions for innovation, doing AI ethics means the response is never less and slower, but always more and faster.

James Brusseau *(PhD, Philosophy) is author of books, articles, and media in the history of philosophy and ethics. He has taught in Europe, Mexico, and currently at Pace University near his home in New York City. His academic research explores the human experience of artificial intelligence.*
He is also Visiting Professor in the Department of Information Engineering and Computer Science, University of Trento, Italy

jamesbrusseau.net

Co-Design of a Trustworthy AI System for Skin Lesion Analysis – Lessons Learned

Adriano Lucieri

Key findings

- The Holistic Co-Design Process based on z-inspection broadened the exAID teams' understanding of other stakeholders' views and needs.

- It revealed tensions between and within stakeholder groups.

- It highlighted the importance of overdiagnosis, and the need to precisely define target variables for performance evaluation.
- It highlighted the large diversity in patient opinions, and the corresponding need to include patient peer groups in early stages of the development process.

Despite accounting for only one percent of all skin cancers, melanoma is responsible for the majority of skin cancer deaths. The American Cancer Society estimates that around 7,990 people will die from melanoma in the U.S. in the year 2023 (Source: https://www.cancer.org). Artificial Intelligence (AI) bears huge potential to improve prevention and care measures for a growing population of affected people worldwide. However, the use of AI systems is often hampered by their lack of decision explanation.
exAID (explainable AI in Dermatology) is a trustworthy explanation framework developed by a research team from the German Research Center for Artificial Intelligence (DFKI). The system is able to complement the output of well-performing, existing AI classifiers for skin lesion analysis with decision explanations. The framework extends raw diagnosis from a medical image AI with quantitative, visual and textual explanations. exAID is based on the idea of concept-based explanations, meaning that the networks' complex internal decision processes are being mapped onto concepts, commonly used in the diagnostic workflows of domain experts. The computation of these concepts is the basis for the generation of user-friendly concept influence scores, visual concept localization heatmaps and concise textual decision explanations. Besides the explanation of single decisions, exAID also allows the explanation of the global classifier behavior by aggregating dataset-wide information. Being in the early design phase, the team of exAID was supported by Z-inspection through an ethically aligned Co-Design process, aiming at the development of a trustworthy AI roadmap for the further development and implementation of the AI system.
The first phase of the system constituted the definition of the initial aim of the exAID project, as defined by the research team from DFKI. This included the definition of envisioned use cases, as well as the disclosure of training details (e.g., training data composition and distribution) and known limitations of the system.
The actual Co-Design process was held in an interdisciplinary team of 35 experts from diverse fields including philosophers, ethicists, policy makers, social scientists, medical doctors, legal and data protection specialists, computer scientists and machine learning (ML) engineers to draw from the collective creativity of all independent experts. In a total of three workshops, the AI system was examined from a variety of angles using socio-technical usage scenarios. Smaller, independently working focus groups of three to five experts were later formed to work on particular topics. The aim was to identify possible exposures when designing the system, and help create a trustworthy AI roadmap for the design, implementation, and future deployment of the framework.
The Co-Design process led to the identification of a variety of tensions, some of which will lead to trade-offs in the development of the AI. Interestingly, tensions were not only revealed between stakeholder groups, but also within single groups. Moreover, the workshops highlighted the need for definition and communication of common concepts and terminologies between the various stakeholder groups. For the team of exAID, all exchanges and discussions in the multidisciplinary group were highly inspiring and insightful, revealing new,

so far unattended aspects of the development and their effect on later deployment of the system. One such key finding was the importance of the precise definition of any target variable used for the performance evaluation of the system. Although diagnostic success is commonly used as performance criteria in medical applications of AI, discussions revealed that overdiagnosis of diseases might be a relevant aspect which should be considered in the definition of a reasonable target variable. The discussion also revealed that biases are not always problematic if they are aligned with population-wide statistics. Moreover, the analysis highlighted the importance of the requirements for Trustworthy AI as defined by the High-Level Expert Group on AI (AI HLEG) set up by the European Commission, including Privacy and Fairness. Another key finding was the identification of potential effects of such an AI system on the patient stakeholder group, as well as the high diversity in patient opinions. This led to the suggestion to include patient peer groups already in early phases of the development process to receive feedback regarding the acceptance of usage scenarios, the nteroperability of the system but also to assess these potential effects on the patients' well-being.

The interdisciplinary nature of Z-inspection in the context of early Co-Design constituted a great enrichment for the ExAID team. While the development of AI projects in teams specialized on computer science usually revolves around the data and application perspective, the workshops really led to the identification of relevant stakeholder groups, a broadened understanding of the stakeholders' needs, potential tensions and trade-offs between and within groups, as well as the identification of potential conflicts of interest. The application of the Z-inspection process therefore results in benefits related to the overall quality of the project's process, as the researchers were encouraged to formulate the ultimate project goal more precisely, considering all stakeholders' needs. This can reveal relevant system changes in early design phases, which can save time, and additionally ensures the continuous alignment of the development goals with the clinical use case.

Adriano Lucieri *completed his BE in Mechatronic Engineering from Duale Hochschule Baden-Württemberg (DHBW) Mannheim and MS in Mechatronic Systems Engineering from Hochschule Pforzheim in Germany. He is presently pursuing his PhD from Rheinland-Pfälzische Technische Universität Kaiserslautern (RPTU), Germany and is also working as Research Assistant at Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI). His research focus lies on improving the trustworthiness of Computer-Aided Diagnosis (CAD) systems based on Deep Learning for medical image analysis. His work includes the topic of explainable AI, where he focuses on the concept-based and feature-based explanation of skin lesion classifiers as well as the localization of concept regions in input images. In the topic of Privacy-preserving Machine Learning (PPML) he is working on the investigation of effects of common private training methods including Differential Privacy and Federated Learning, on model accuracy and explainability of Deep Learning models.*

https://exaid.kl.dfki.de

Co-Design of a Trustworthy AI System in the detection of Melanoma, what did we learn?

Results of an interdisciplinary working group

Ulrich Kühne,

Key findings

- A multidisciplinary team is necessary for the Z-inspection process. In this case AI engineers, computer scientists, specialists in dermatology, evidence based medicine, family medicine, legal and data protection specialists, philosophers, ethicists, statisticians, social scientists, patients.

- The roles of the different stakeholders have to be defined.

- The multidisciplinary team should be involved at an early stage of the design of the AI System.

- Trade-offs and tensions between the different specialties must be identified.

- There are biases in the datasets that were used to train the AI system. They must be identified and considered when using the system. They are not necessarily harmful.

- The aim of the AI system should be made a claim that can be verified or falsified.

- Autonomy of the physicians and patients over the AI system has to be preserved.

What is it all about?

As a dermatologist, I'd like to share my experience of being part of the co-design process. In my daily routine examining patients for skin cancer is one of the most common tasks. The usual diagnostic algorithm is as follows: asking the patient if he has noticed anything suspicious, visual inspection of the skin, use of dermoscopy (microscope that is placed directly on the skin) and high resolution video dermoscopy for suspicious lesions. Further non-invasive diagnostic methods are electro-impedance-spectroscopy or confocal laser-microscopy. These tools already use AI. The aim of the examination is to determine if a lesion is potentially malignant and thus should be removed. This is important, because early detection of melanomas before they have metastasized bears a favorable prognosis compared to later stages. The average dermatologist performs 6– 30 excisions of benign lesions to remove one melanoma (number needed to treat). Any method that would make our decision to operate more precise and thus reduce the number of unnecessary operations would be very welcome.

Tensions

A hot topic which was a source of lively discussions between the dermatologists and evidence-based medicine physicians was the problem of overdiagnosis. Overdiagnosis in our case is defined as correctly diagnosing a melanoma that would never have harmed the patient and thus exposing him to unnecessary treatments and labeling him as a "cancer-patient" with all the psychological implications. If the AI system would lead to a higher rate of detecting melanomas it could thus contribute to overdiagnosis.

On the other hand a melanoma detected at an early stage has a far better prognosis than in later stages (five year survival rates from 99% down to 27%). Now there are no preoperative markers that could tell us if a given melanoma has the potential to metastasize, so a definitive diagnosis can only be made after excision by histological examination. The demand to remove only "clinically relevant" melanomas cannot be met. The question if screening programs for melanoma are useful is still controversial.

Bias

The datasets used to train the AI system were derived mainly from patients with Fitzpatrick skin types I – IV (celtic skin types to olive skin types). Skin types V and VI (dark brown and

black) were scarcely represented. This is partly due to the fact that benign moles and melanomas have a higher prevalence in the lighter skin types. This bias must be considered but might be acceptable. Personal comment: doctors and their expertise are biased too depending on their experience and population of patients. Many European dermatologists have limited experience in examining patients with skin of color.

Dr. med. Ulrich Kühne
u.kuehne@hautmedizin-badsoden.de,
www.hautmedizin-badsoden.de,
https://www.linkedin.com/in/ulrich-kühne-657831a5

Address:
Hautmedizin Bad Soden, Kronberger Str. 36a, D-65812 Bad Soden, Germany, u.kuehne@hautmedizin-badsoden.de, www.hautmedizin-badsoden.de

Professional career:
1979 – 1985 studies of Medicine at Johann Wolfgang Goethe-University Frankfurt, Germany. 1987 – 1992 Residency at the Dept. of Dermatology and Allergology at University Hospital Frankfurt. Board Certification for Dermatology and Allergology. Since 1993 in private practice “Hautmedizin Bad Soden”, Bad Soden, Germany. Consultant Dermatologist at Main-Taunus-Clinic, Bad Soden. 1994-2000 Medical Advisor for Procter & Gamble Central Europe. Examining patients for skin cancer is the most common task in daily routine.

Scientific Activities:
Investigator and co-investigator in clinical studies. Publications and speaker at national and international conferences on aesthetic dermatology. Since 2020, member of Z-Inspection.

Best practices: Lessons learned

What did we learn in three AI in healthcare best practice assessments?

Vince I. Madai

https://www.bihealth.org/en/quest/service/service/trustworthy-ai-in-healthcare-lab
https://www.linkedin.com/in/vince-madai/

Key findings
- Identifying ethical issues is crucial for trustworthy AI.

- Z-Inspection® is a unique approach to assess AI.

- Challenges include ambiguity, patient perspectives, and organization compliance. Ensure high-quality data, feedback mechanism, risk management, human oversight.

The use of artificial intelligence (AI) is becoming increasingly prevalent in various industries, including healthcare, making trustworthy AI a critical concern. Trustworthy AI systems are designed, developed, and implemented in a way that aligns with ethical and societal values and does not pose any harm to individuals or groups. In this text, I will discuss the lessons learned to make trustworthy AI a success, based on three best practice use cases applying the Z-inspection methodology. I will focus on what we learned, how it was different from usual research work, and the challenges we faced, and will give an outlook.

What Did We Learn?

The first significant lesson learned is the importance of gaining experience with mapping the various ethical issues related to AI systems to key requirements for trustworthy AI. The mapping process allowed us to identify and consolidate related issues from different groups and ensure that ethical principles were considered in the design and development of the AI system. We also learned the importance of a diverse team of experts, including

interdisciplinary experts, to ensure that different perspectives are considered. Another key lesson was the importance of limiting the set of ethical principles and approaches employed to convert theoretical discussions into practical and applicable results. It is also important to ensure that the proposed trustworthiness definition matches the expectations of affected parties. Our co-design approach for developing trustworthy AI in healthcare using a holistic approach involved an interdisciplinary team of experts from different fields. Co-designing trustworthy AI with a holistic approach required some unique aspects in the structure and design of the process, such as the involvement of an interdisciplinary team and a co-design methodology. The evaluation of the AI system with a holistic approach created benefits related to general acceptance or concerns inside and outside the institution that applies an AI project, as well as benefits related to the quality of the project‘s processes, transparency about possible conflicts of interest, and in general comprehensibility of the system.

How Was It Different Than Usual Research Work?

The Z-Inspection® process for the assessment of the AI system is a unique approach that sets it apart from traditional research work. The Z-Inspection® process involves a team of interdisciplinary experts who assess the AI system based on specific guidelines and requirements for trustworthy AI. This approach ensures that ethical and societal values are considered throughout the development and implementation of the AI system. The Z-Inspection® framework differs from usual research work in that it is a holistic and multidisciplinary evaluation process that considers ethical, legal, and societal concerns related to the use of AI.

Challenges We Faced

One of the significant challenges we faced was handling the ambiguity of the mapping from issues to key requirements. Different groups frequently mapped issues to different key requirements, which made the consolidation process less effective as planned. Another challenge was ensuring that the mapping process did not overlook other ethical concepts and principles relevant to the use case. The lack of patient perspectives was also a challenge in the assessment process. Another challenge was the identification of a list of specific evaluation criteria that is complete and as exhaustive as possible. Another challenge of the Z-Inspection® framework is that the evaluation cannot guarantee that the organization administering the AI system necessarily sticks to the recommendations that are given. However, participation in the inspection is voluntary, and organizations come with a high openness for proposed changes. Another challenge was the potential misinterpretation of the results of the exchanges and of the ethical evaluation established by Z-Inspection®.

Outlook

Based on the findings and lessons learned, a few common recommendations can be identified to ensure trustworthy AI in the development and implementation of AI systems. These include the need for a large dataset with diverse, high-quality data, a feedback mechanism to prevent bias, a detailed risk management plan, and policies to secure informed consent and protect patient rights. We also recommend a test branch and service with a public repository that allows external parties to test the models directly with test data. Additionally, an external audit should test the model to certify ethical and healthcare standards. Additionally, the need for human oversight in AI is essential to avoid the problem of automation bias, where an AI system becomes optimal in suggesting decisions to humans and reducing human attention and responsibility. A human-centric approach is required, where machines complete human actions and assist human decision-making rather than replacing it. Z-Inspection® is a highly suited method to ensure this for trustworthy design and implementation of AI tools in healthcare.

Vince Madai

Work Experience:
- Project team lead and Principal Investigator, Berlin Institute of Health, QUEST Center (Oct 2022 - Present)
- Senior Researcher and Principal Investigator, Berlin Institute of Health, QUEST Center (Mar 2021 - Oct 2022)
- Visiting Professor, Birmingham City University (Mar 2020 - Present)
- CSO, ai4medicine (Sep 2018 - Feb 2023)
- Senior Researcher, Charité - Universitätsmedizin Berlin (Nov 2016 - Apr 2022)
- Researcher, Charité - Universitätsmedizin Berlin (Sep 2008 - Oct 2016)
- Visiting Scholar, McGill University (Apr 2016 - Jun 2016)

Education:
- PhD in Medical Neuroscience, Charité (2014 - 2017) - Grade: summa cum laude
- Data Analyst Nanodegree, Udacity (2017 - 2019)
- MA in Medical Ethics, Johannes Gutenberg University Mainz (2011 - 2016)
- Dr. med. in Medicine, Charité (2012) - Grade: summa cum laude

Trustworthy AI for Healthcare Laboratory at Tampere University

Pedro A. Moreno Sánchez, Ph.D

In the field of healthcare, Artificial Intelligence (AI) is currently promising support for healthcare professionals in their decision-making process of disease pattern detection or predicting risk situations for patients. However, when these AI systems' outputs affect the patient's life, their adoption in clinical routine encounters barriers related to the trustworthiness of the outputs. The trustworthiness aspect of AI solutions has become a relevant value to be considered by any stakeholders related to AI solutions lifecycle. According to this importance, the European Commission delivered the Ethics guidelines for trustworthy AI where different requirements are stated. In the context of tackling the Trustworthy AI requirements in AI applications, the Z-inspection® initiative (http://z-inspection.org) proposes a framework to assess AI solutions in different phases of their lifecycle based on the Ethics Guidelines for Trustworthy AI by the European Commission High-Level Expert Group on Artificial Intelligence.

Who we are

The Trustworthy AI for Healthcare Laboratory at Tampere University, affiliated with the Z-inspection ® initiative, is composed of a group of researchers that pursue to leverage the importance of Trustworthy AI in academic and civil society to enhance the AI solutions aimed at healthcare and improve their uptake by the different healthcare stakeholders. Tampere University is the second largest university in Finland with more than 21 000 students and 4000 staff members. The Lab is sited in the BioMedTech Unit at the Faculty of Medicine and Health Technology. Specifically, the Lab is implemented by researchers from the Decision Support for Healthcare (DSH) research group

aimed at helping professionals and citizens through the extraction of knowledge from data to improve decision-making for diagnosis, prognosis or treatments. DSH experience relies on different technologies such as AI, signal processing, and machine learning approaches that deal with highly diverse real-world data, which has challenges such as imperfection, individual variability, missingness, etc. These technologies are utilized in a disease-agnostic approach through different research projects, but DSH has special experience in neurological diseases and cardiac and diabetes.

Goals of the Lab

The main goal of the Trustworthy AI for Healthcare Laboratory at Tampere University is to leverage the importance of Trustworthy AI in academic and civil society to enhance the AI solutions aimed at healthcare and improve their uptake by the different healthcare stakeholders. As specific objectives, the lab aims to:

- Design, develop and evaluate decision-support models for different healthcare fields that foster Trustworthy AI principles, making emphasis on explainable AI and the risk of bias.

- Raise the awareness of Trustworthy AI in the health data science community of TUNI.

- Support research activities and project proposals of TUNI researchers in the area of health data science who are interested in complying with Trustworthy AI requirements.

- Disseminate the Z-inspection process in TUNI's education programs as well as in different research activities to achieve design, development, and validation of AI solutions in healthcare aligned with the principles of Trustworthy AI.

- Build a research community around the topic of Trustworthy AI in Healthcare where researchers from Tampere, Finland, and abroad can share their expertise and effort toward different research and education activities.

Personnel

The main researchers associated with the Lab are Pedro A. Moreno-Sanchez, PhD (Lead researcher) and Mark van Gils, Prof. (researcher). In addition, several BSc and MSc students collaborate on the Lab's goal by working on their thesis that tackles different aspects of Trustworthy AI. As an external advisor, we count on the support of Roberto Zicari, Prof.

Activities of the Lab

Members of the lab are involved in different activities aligned with the goals of the Lab, a (non-exhaustive) list as follows:

- Moreno-Sánchez, Pedro A. "Data-Driven Early Diagnosis of Chronic Kidney Disease: Development and Evaluation of an Explainable AI Model." *IEEE Access* 11 (2023): 38359–69. https://doi.org/10.1109/ACCESS.2023.3264270.

- Lenatti, Marta, -Sánchez Pedro A. Moreno, Edoardo M. Polo, Maximiliano Mollura, Riccardo Barbieri, and Alessia Paglialonga. "Evaluation of Machine Learning Algorithms and Explainability Techniques to Detect Hearing Loss From a Speech-in-Noise Screening Test." *American Journal of Audiology* 31, no. 3S (September 21, 2022): 961–79. https://doi.org/10.1044/2022_AJA-21-00194.

- Allahabadi, Himanshi, Julia Amann, Isabelle Balot, Andrea Beretta, Charles Binkley, Jonas Bozenhard, Frédérick Bruneault, et al. "Assessing Trustworthy AI in Times of COVID-19. Deep Learning for Predicting a Multi-Regional Score Conveying the Degree of Lung Compromise in COVID-19 Patients." I*EEE Transactions on Technology and Society*, 2022, 1–1. https://doi.org/10.1109/TTS.2022.3195114.

- Zicari, Roberto V., Sheraz Ahmed, Julia Amann, Stephan Alexander Braun, John Brodersen, Frédérick Bruneault, James Brusseau, et al. "Co-Design of a Trustworthy AI System in Healthcare: Deep Learning Based Skin Lesion Classifier." *Frontiers in Human Dynamics* 3 (2021). https://www.frontiersin.org/article/10.3389/fhumd.2021.688152.

- Explainable AI lecture in DSH course

- PerCard project: PerCard project | Tampere Universities (tuni.fi)

The Lab is seeking to implement a new use case where the Z-inspection can be applied to assess the Trustworthiness of some of the projects in which DSH is participating. Therefore, the PerCard project provides the opportunity to implement the process during its final validation stage, where explicitly different ethical, legal and societal aspects are tackled. This point will be discussed with the project partners in the next consortium meeting.

Contact

To know more about the lab, do not hesitate to contact us: *pedro.morenosanchez@tuni.fi*

Pedro A. Moreno Sánchez, Ph.D
Faculty of Medicine and Health Technology, Tampere University (Finland)

Pedro A. Moreno-Sánchez works as Postdoctoral Research Fellow at Tampere University. He has RDI experience in the digital health field since 2007 working in Spanish and European projects focused on eHealth for supporting older adults and funded by several programs like Horizon 2020, EIT-Health, AAL. He has recently worked as an expert evaluator for the Horizon Europe research and innovation program.

He worked for 9 years as a digital health researcher and project manager at the Bioengineering and Telemedicine Group of the Polytechnic University of Madrid, and 3 years as a senior researcher at the Biomedical Research Foundation of the University Hospital of Getafe-Madrid (Spain). From 2020 to 2022, he worked as an RDI expert and AI researcher in digital health and wellbeing technologies at Seinäjoki University of Applied Sciences (Finland). Since 2022, he works as a Postdoctoral Research fellow in the Decision Support for Healthcare research group at Tampere University focusing on developing explainable and trustworthy AI models for healthcare applications. He also had educational experience working as a lecturer at the Polytechnic University of Madrid (Spain), Seinäjoki University of Applied Sciences (Finland), and Tampere University (Finland). He holds a Ph.D. in Telecommunication Engineering (Biomedical Engineering specialization) from Polytechnic University of Madrid (Spain). As well, he holds a master's degree in "Bioengineering and Telemedicine" and a bachelor's degree in Telecommunication Engineering by the Polytechnic University of Madrid (Spain). Obtained in 2018 the Project Management Professional –PMP® certification by the Project Management Institute®.
His current research focus is on eXplainable Artificial Intelligence (XAI) applied to the healthcare domain and disease prediction models in different diseases like cardiovascular, fibromyalgia, chronic kidney disease, or emergency services.

Pedro was awarded a Z-inspection® Teaching Certificate.

Trustworthy AI in Practice: Best practices

Hanna Sormunen

Key findings
- Explain the Z-inspection® process to the people involved at every stage

- Assigning leads for dedicated working groups

- Communicate the expectations of the outcomes clearly

- Have deadlines, enough time, and guidelines for working together

- Encourage involvement and open discussion

- Create a safe environment to share thoughts

- Celebrate milestones and have fun together

The Z-inspection® process can feel a bit overwhelming at first.
Having an experienced leader to assist through the process is essential.

https://fi.linkedin.com/in/hannasormunen

Hanna Sormunen *is a Data Scientist focusing on creating sustainable AI solutions and software systems. She leads the corporate social responsibility assessment team for AI in Finnish Tax Administration.*

She holds a master's degree in Big Data Analytics and B.Sc. in Automation Technology. She has been developing and designing software systems and AI solutions for the automation industry and for the government for nearly 25 years.

Hanna finds it essential to empower people through inclusion and co-design to make the world a better place for all.

Hanna is a Certified Z-Inspection® Teaching Expert and joined the Z-inspection® initiative in April 2021

Assessing Trustworthy AI in times of COVID-19.

Deep Learning for predicting a multi-regional score conveying the degree of lung compromise in COVID-19 patients.

Alberto Signoroni , Davide Farina , Mattia Savardi

What Did We Learn?

The Z-Inspection® process provided a valuable learning experience for the team, as it required to consideration of many different aspects beyond just the technical performance of the BrixIA-Net system.
Firstly, it gave our first professional experience of an inspection process of this kind. The rigor and scrutiny that AI systems must undergo before being deployed in a clinical setting should be high.
The Z-Inspection® process also highlighted the threats, drawbacks, but also advantages of AI in a clinical routine, such as the tension between speed and accuracy, and the need for interpretability and human oversight. This increased awareness among the team of the ethical and regulatory implications of AI for disease prevention. This is an area we are interested in continuing to work on.
Furthermore, the process required a strong emphasis on teamwork and multidisciplinarity, as approximately 50 experts from different fields contributed to the evaluation of the system. This collaboration enabled the creation of an internal working document of about 150 pages, which contained details of the activities carried out and the findings.

How Was It Different Than Usual Research Work?

The Z-Inspection® process differed from usual research work in several ways. Firstly, it involved a much larger group of experts from different disciplines working together on a specific task. This required a structured approach and small subgroups to ensure efficient communi-

cation and evaluation.
Besides the usual important performance-related goals (the system is capable of segmenting and aligning the lungs in the input CXR to provide a robust estimation of the Brixia score), to open a trustable channel of communication between the AI and the radiologist, the system produces saliency maps that provide interpretability and validation/human oversight.
Finally, the Z-Inspection® process inspired subsequent works by the team, who expressed interest in continuing to work together, particularly on the ethical and regulatory implications of AI for disease prevention. During the process, several discussions emerged regarding the emergency-driven nature of the work, the ethical and regulatory implications of AI for disease prevention, and the role of Z-inspection during the research and deployment activities.

Challenges we faced

During the Z-Inspection® process, the team faced challenges related to the emergency-driven nature of the work, the tension between accuracy and speed, interpretability and detail, and the type and size of data. Additionally, the experts mostly carried out the ex-post assessment, raising questions about how and when Z-inspection should be included during research and deployment activities.
One last consideration was about the nature of the relationship that was almost always „virtual“ (except for the recent lovely Venice meetup). This „no site visit“ approach has a clear advantage, yet adds some challenges in the first part in which trust between peers must be established.
However, these challenges also led to increased awareness of the importance of ethical and regulatory considerations in AI-assisted medical diagnosis and the need for continued collaboration and communication between experts from different fields.

Outlook

The BrixIA-Net system has been shown to be highly accurate and validated on a dataset from multiple centres worldwide and has the potential to improve the consistency and accuracy of radiologists‘ diagnoses, ultimately leading to better patient outcomes.
Moving forward, the team is interested in continuing to work on the ethical and regulatory implications of AI for disease prevention and improving the tension between explainability and accuracy in AI systems. This requires continued collaboration between experts from different fields to achieve success in AI-assisted medical diagnosis. Overall, the Z-Inspection® process was an essential step towards ensuring the quality and effectiveness of the BrixIA-Net system and has provided valuable lessons for the team moving forward.

Summary and key findings

The Z-Inspection® process taught the team valuable lessons beyond just the technical performance of the BrixIA-Net system, with an holistic approach. The inspection process also emphasized the importance of interpretability and human oversight in AI, multidisciplinary teamwork, and ethical and regulatory considerations. The process differed from usual research work by involving a larger group of experts from different fields and inspiring subsequent works by the team. Challenges faced included the emergency-driven nature of the work, the tension between accuracy and speed, interpretability and detail, and the virtual nature of the relationship between experts.
The BrixIA-Net system was highly accurate and validated on a dataset from multiple centres worldwide, with the potential to improve the consistency and accuracy of radiologists‘ work. The team is interested in continuing to work on the ethical and regulatory implications of AI for disease prevention and improving the tension between explainability and accuracy in AI systems.

Key Findings:

1. Interpretability and human oversight are crucial in AI, and multidisciplinary teamwork is necessary for effective evaluation.
2. The virtual nature of the relationship between experts was a challenge during the process. Still, it enabled this kind of approach.
3. Colud happens that some tension cannot

be resolved, especially in case of conflicting perspectives. Yet, especially the highlighting of those tensions ensures a higher level of awareness and oversight.
4. The team is interested in continuing to work on the ethical and regulatory implications of AI for disease prevention and improving explainability in AI systems.

Assessing Trustworthy AI in times of COVID-19.
Deep Learning for predicting a multi-regional score conveying the degree of lung compromise in COVID-19 patients.

IEEE Transactions on Technology and Society
DECEMBER 2022 * Volume: 3, Issue: 4
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9845195

Alberto Signoroni
Department of Information Engineering, University of Brescia, Italy

Alberto Signoroni is an assistant professor of Signal Processing and Communications at the University of Brescia, Italy. He currently teaches courses of Advanced Methods for Information Representation, Remote Sensing Data Analysis (with Machine Learning and Deep Learning contents), Law and Regulation for ICT.

His research interests revolve around computer vision, image and multidimensional visual data analysis and understanding, machine learning, deep learning, biomedical image analysis, geometry processing, 3D data processing (mesh and point clouds), computer graphics applications (biomedicine, cultural heritage, computer aided design), hyperspectral imaging, compressive sensing, image and multidimensional data compression. He leads the research activities of several funded research projects and university-industry collaborations.

Davide Farina
Department of Medical and Surgical Specialties, Radiological Sciences, and Public Health, University of Brescia, Brescia, Italy

Davide Farina is Associate Professor of Radiology at Università degli Studi di Brescia (Italy). His clinical and scientific interests are Imaging of the Head and Neck and Cardiovascular Imaging.

Mattia Savardi
Department of Information Engineering, University of Brescia, Italy

Mattia Savardi received his M.Sc. in Communication Technologies and Multimedia (cum laude) and obtained his PhD with merit in Technology for Health at UNIBS with a thesis on Deep Learning techniques applied to Medical Image Analysis.

He was the recipient of the GTTI PhD award in 2020 for the best thesis. During his research he worked with brain functional MR, CXR, Hyper-spectral and RGB biomedical images, in collaboration with many international universities. He is currently doing a Postdoc at UNIBS.

Integrating The Fundamental Rights and Algorithm Impact Assessment (FRAIA) into the Z-Inspection® Process

Roberto V. Zicari

I can only second what Willy Tadema says: „Integrating The Fundamental Rights and Algorithm Impact Assessment (FRAIA) into the Z-Inspection® method contributed to great conversations about human rights, both in the pilot and during the conference“.

Not only that, but we applied a novel approach to the assessment of fundamental human rights for AI which is based on Claims, Arguments and Evidence. This led to a truly original approach within the context of the Z-Inspection® process. We will soon publish some of the results.
The Pilot Project: „Assessment for Responsible Artificial Intelligence“ was conducted by a team of experts of the Z-Inspection® initiative together with Rijks ICT Gilde part of the Ministry of the Interior and Kingdom Relations (BZK)- and the province of Fryslân (The Netherlands).

During this six-month pilot, the practical application of a deep learning algorithm from the province of Fryslân was investigated and assessed. The AI maps heathland grassland by means of satellite images for monitoring nature reserves.
The Trustworthiness of this AI was assessed using the Z-Inspection® process.

The Assessment for Responsible AI pilot took place from May 2022 through January 2023.

Project Members:

Sara M. Beery,
Marjolein Boonstra,
Frédérick Bruneault,
Subrata Chakraborty,
Tjitske Faber,
Alessio Gallucci,
Eleanore Hickman,
Gerard Kema,
Heejin Kim,
Jaap Kooiker,
Ruth Koole,
Elisabeth Hildt,
Annegret Lamadé,
Emilie Wiinblad Mathez
Florian Möslein,
Genien Pathuis,
Rosa Maria Roman-Cuesta,
Marijke Steege,
Alice Stocco,
Willy Tadema,
Jarno Tuimala,
Isabel van Vledder,
Dennis Vetter,
Jana Vetter,
Elise Wendt,
Magnus Westerlund,
Roberto V. Zicari

The Trustworthy AI Laboratory at the University of Copenhagen

Staff and Research overview of Trustworthy AI Laboratory, UCPH, Copenhagen

Boris Düdder

https://di.ku.dk/english/research/groups/trustworthyai/

Key findings

- Trustworthy AI is not a product but a process.

- A holistic approach requires cross-faculty collaboration.

- Neither certification nor assessment can prevent misuse.

The Trustworthy AI Lab is dedicated to advancing research, education, policy, and best practices for the ethical, responsible, mindful, sustainable, and trustworthy use of Artificial Intelligence (AI) and AI-based applications.
As AI poses both opportunities and threats to society, the Lab's mission is to help shape the future of AI in a way that benefits society as a whole.
The Lab is highly multidisciplinary, bringing together researchers from various fields such as computer science, philosophy, law, medicine, communication research, sociology, psychology, and engineering. This multidisciplinary approach ensures that the Lab's research and recommendations are comprehensive, robust, and inclusive of various perspectives and considerations.
The Lab serves as a peer forum for academic and industrial research and applications, facilitating collaboration and knowledge-sharing between researchers, policymakers, and industry professionals. This approach enables the Lab to stay abreast of the latest developments in AI and related fields, while also contributing to creating new knowledge and best practices.
In addition to its research activities, the Lab offers workshops and continuous learning opportunities for a wide range of stakeholders, including citizens, local and regional private entities, healthcare professionals, engineers, developers, and students. By providing accessible and engaging education on the ethical and trustworthy use of AI, the Lab aims to empower individuals and organizations to make informed decisions about AI and its applications.
Overall, the Trustworthy AI Lab is committed to advancing the responsible and ethical use of AI while promoting the development of AI-based applications that serve the common good. Through its research, education, and policy activities, the Lab seeks to ensure that AI is developed and used to benefit society and foster trust between individuals, organizations, and technology.

Dr. Boris Düdder *is an associate professor at the Department of Computer Science (DIKU) at the University of Copenhagen (UCPH), Denmark. He is widely recognized as a leading authority in the field of software engineering and formal methods. Dr. Düdder leads the research group on Software Engineering & Formal Methods at DIKU, where he leads innovative research projects on trustworthy distributed systems. Dr. Düdder's primary research interests lie in the areas of formal methods and programming languages in software engineering of trustworthy distributed systems, with a focus on automated program generation for adaptive systems with high-reliability guarantees. His work also involves the computational foundations of reliable and secure Big*

Data ecosystems. Dr. Düdder's long-term research vision is to develop dependable, adaptive, and software-defined technical systems based on program synthesis for manufacturing and logistics. These systems will enable safe and secure vertical and horizontal, cross-industry integration and interaction among untrusted parties. In addition, Dr. Düdder is a guest professor at the Copenhagen Business School, where he teaches courses in emerging technologies and digital transformation. He is involved in multiple projects focused on innovative and dependable industrial IT infrastructure for enterprises, manufacturing industries, and national healthcare IT, including Data Ecosystems, Smart Factories, and Industry 4.0. Dr. Düdder's research and teaching are widely respected, and his work has been published in numerous high-impact journals and conference proceedings.

Trustworthy AI Lab at the Imaging Lab, University of Pisa (Pisa, Italy)

Roberto Francischello, Prof. Emanuele Neri and the ImagingLab

The deployment of AI solutions into the healthcare setting requires a stringent analysis of the trustworthiness of the model and a deep assessment of the deployment's context. The results of such characterization depend on the researcher's background and training. Indeed, the researcher carriers lay the foundation for the operator's bias and contribute to the definition of the operator's "cultural tool" used to characterize the AI solution.
The Imaging Lab is a multidisciplinary laboratory dedicated to frontier research in the study of biomedical images. The Lab is coordinated by Prof. Emanuele Neri, Chair of the Unit of Academic Radiology, Chair of the Post-graduate School of Radiology, Head of Medical School, and Faculty of the Department of Translational Research of the University of Pisa. The Lab has a multidisciplinary staff of 6 radiologists, 2 physicists, one specialist in nuclear medicine, and one neuropsychologist. The group's main research focus is the construction of oncological imaging biobanks. We are partners in 5 European projects: PRIMAGE, EuCanImage, CHAIMELEON, PROCANCER, and EUCAIM. We are also partners of the Navigator project financed by the Tuscany region.
The EuCanImage project will build a highly secure, federated, and large-scale European cancer (breast, liver, and colorectal) imaging platform, with capabilities that will significantly improve the capabilities of artificial intelligence (AI) in oncology. During the first phase of the project, clinical and imaging data will be collected to create the biobank, then AI solutions will be developed by the consortium partner, and in the last phase, the biobank will be accessible to external users to develop new AI solutions.
I, Roberto Francischello, am a physicist with a Ph.D. in Chemistry and I'm currently working as a postdoc for the EuCanImage project at

the ImageingLab of Pisa University. I started my research activity in the field of material science and slowly drifted into the life science field, first during my Ph.D. when I worked on small animal magnetic resonance imaging, and then in my postdoc at the Imaging Lab where I started my experience with clinical research. Due to my heterogeneous background, I didn‘t receive any formal training in ethical matters. Due to my involvement with the EuCanImage project, I decided to expand my knowledge on the ethical issue related to the use of AI in healthcare and Prof. Emanuele Neri suggested Z-inspection.

At first, I was curious and surprised by the Z-inspection procedure for the trustworthiness assessment of AI. I was surprised by the step-by-step nature of the method and its open-endedness; the Z-inspection procedure is not just a checklist but a guided procedure to identify the ethical tension depending on the characteristic of the AI solution assessed. In addition, the ethical assessment procedure prompted me to abandon a metrics-based evaluation for a consequence-based evaluation. Instead of asking, „What is the AUC of my model?“ I began to ask, „What impact does the AUC of my model have on stakeholders?“ Deepening my experience with Z-inspection I noticed that there were relationships between some ethical tension such as “Personalization vs solidarity”, “Accuracy vs explainability”, and “Accuracy vs fairness”, and the classic Digital Signal Processing trade-off like “Bias vs Variance”, “Outlier Handling”, and “Performance vs physical meaning” that I learned how to handle in my previous research activities. Therefore, my previous research experience is still valuable during the ethical tension assessment.

Prof. Emanuele Neri participate in the ethical assessment of a deep learning method for the prediction of the Brixia score described in the paper “Assessing Trustworthy AI in Times of COVID-19: Deep Learning for Predicting a Multiregional Score Conveying the Degree of Lung Compromise in COVID-19 Patients” published in IEEE TRANSACTIONS ON TECHNOLOGY AND SOCIETY, VOL. 3, NO. 4, DECEMBER 2022. It was an insightful experience that highlighted the importance of collegiality and multidisciplinarity of the Z-inspection procedure.

Our future plan for the Trustworthy AI Lab at the Imaging Lab of UNIPI includes both retrospective and prospective evaluation. We are involved in 6 projects for the development of multiple AI solutions with 10s of the potential models to retrospectively assess. The ImagingLab is also expanding the research activity into the field of interpretable machine learning and the practice of a trustworthiness evaluation could be easily incorporated into our research workflow.Co-Design: Think Holistically

The Trustworthy AI in Healthcare Lab in Berlin

Vision and goals of the Z-Inspection® Trustworthy AI in Healthcare Lab in Berlin at the QUEST Centre for Responsible Research at the Berlin Institute of Health of Charité Berlin

Vince I. Madai

https://www.bihealth.org/en/quest/service/service/trustworthy-ai-in-healthcare-lab
https://www.linkedin.com/in/vince-madai/

Key findings

- AI in healthcare raises ethical and societal concerns that need to be addressed.

- The newly founded Trustworthy AI in Healthcare Lab uses the Z-Inspection® process to assess the ethical implications of AI systems.

- The lab is located in Berlin‘s vibrant research and startup ecosystem

- The lab is uniquely suited to lead and conduct comprehensive and high-quality assessments of AI systems in healthcare.

- The lab‘s vision is to promote ethical AI in healthcare through collaboration and international standards.

Artificial intelligence (AI) is rapidly transforming healthcare research and the healthcare industry. The ability of AI to process large amounts of data and recognize patterns that are not easily detectable by humans has the potential to significantly advance, for example, medical diagnosis, treatment, and drug development. AI has the potential to save lives, improve patient outcomes, and increase the efficiency of healthcare services. However, the use of AI in healthcare also raises significant ethical and societal concerns.

To address these concerns, Z-inspection® and the Project Team Responsible Algorithms at the QUEST Centre for Responsible Research have established the Z-inspection® Trustworthy AI in Healthcare Lab based in Berlin, Germany. The lab is a multidisciplinary initiative that aims to bring together experts from various fields, including healthcare, computer science, engineering, ethics, social sciences, and law, to conduct ethical assessments and co-creation designs of AI systems specifically designed for healthcare. The lab‘s location in the vibrant research and startup ecosystem of Berlin will foster the lab‘s growth and innovation by providing access to a diverse range of experts, cutting-edge research facilities, and a thriving entrepreneurial community. The varied grants of the Project Team responsible algorithms, including Horizon Europe projects and projects funded by the German Ministry for Education and Research, will also support the Z-Inspection® Trustworthy AI lab by providing expertise to lead and conduct comprehensive and high-quality assessments of AI systems in healthcare. Finally, the leader of the lab, Dr. Vince Madai, holds an official Z-Inspection® teacher certificate and has participated in three best practice use cases on cardiac arrest, skin lesion classification, and a COVID-19 radiology tool.

The assessments conducted by the lab are based on the international expertise developed under the Z-Inspection® initiative, which is a process for assessing the trustworthiness of AI systems. The Z-Inspection process uses the definition of trustworthy AI given by the high-level European Commission‘s expert group on AI. It is a general process that can be applied to a variety of domains where AI systems are used, including business, the public sector, and, importantly, healthcare.
The Z-inspection® Trustworthy AI in Healthcare Lab aims to achieve several goals. First, it seeks to promote the development of AI systems that are designed with ethical considerations in mind. Second, the lab aims to provide healthcare researchers, developers, providers and policymakers with public knowledge on the ethical implications of AI systems used in healthcare. Third, the lab hopes to foster collaboration between experts from different labs, institutions, and disciplines to ensure that AI systems in healthcare are designed and used in a manner that is consistent with ethical and societal values. Finally, the lab wants to contribute to the development of international stan-

dards for the ethical use of AI in healthcare.
In conclusion, the Trustworthy AI in Healthcare Lab is an initiative that seeks to ensure that AI systems used in healthcare are designed and used in a manner that is consistent with our ethical and societal values. The lab's use of the Z-Inspection® process, which is based on applied ethics, ensures that assessments are conducted in a systematic and transparent manner.
The lab's vision to promote the development of ethical AI systems and to contribute to the development of international standards for the ethical use of AI in healthcare is a critical step towards building a future where AI is used to improve healthcare outcomes while respecting human values.

Trustworthy AI Lab at Goethe University Frankfurt

A lab to advocate Mindful use of AI

Karsten Tolle and Gemma Roig

http://www.cvai.cs.uni-frankfurt.de/trustAI.html

We present the Trustworthy AI Lab at Goethe University Frankfurt. In our lab, we advocate for the development and use of responsible and trustworthy AI from a holistic perspective taking into account all stakeholders. Our core values are inspired by the EU Ethics Guidelines for Trustworthy AI, and we have been collaborating in the development of the Z-Inspection® process for assessing Trustworthy AI. With this, we want to establish a *„Mindful use of AI"* (#MUAI).

We are interested in actively developing new AI methods that can help evaluate and quantify such principles.

Our team is formed by computer scientists from Goethe University Frankfurt, and we are in constant dialogue with many collaborators from different domains.

We are collaborating in a plethora of projects including the following EU funded ones: i) eXplainable Artificial Intelligence (xAIM - https://xaim.eu/) in Healthcare Management, which is for establishing an interdisciplinary Master's Program at the intersection of AI and health care; and ii) Pan-European Response to the Impacts of COVID-19 and future Pandemics and Epidemics (PERISCOPE - https://eupreVent.eu/periscope/), in which we are investigating the broad socio-economic and behavioral impacts of the COVID-19 pandemic, to make Europe more resilient and prepared for future large-scale risks.

In domains like archeology the definition of Trustworthy can differ, especially compared to domains like healthcare. However, there are also many parallels.

In the area of teaching we have been involved in some lectures of Roberto Zicari and plan to integrate Trustworthy AI in our lectures. In addition we offer student projects at bachelor and master level at Goethe University in our mission to create awareness of the importance of Trustworthy AI.

Karsten Tolle i*s senior researcher and Frankfurt Big Data Lab director at the Goethe University Frankfurt. His research interests are the application of machine learning approaches (NLP, deep learning / image recognition) and Linked Open Data (LOD) within the area of archeology. One of the open questions here is how deep learning systems (with limited explainability) can be accepted and trusted by other domains like archeology. In addition Trustworthy AI will be included in his teaching modules.*

Gemma Roig *is a professor at the computer science department in the Goethe University Frankfurt and a member of The Hessian Center for Artificial Intelligence (hessian.ai). Her research aim is to build computational models of human vision to understand its underlying principles, and to use those models to build applications of artificial intelligence ,as well as to promote and design Trustworthy AI systems.*

Ethical Handling of Data, Algorithms & AI at a Multinational Corporation

How Merck Established and Operationalized Digital Ethics Principles to strengthen trust in new digital technologies and business models.

Dr. Jean Enno Charton, Director Digital Ethics & Bioethics, Merck KGaA

Key findings

- Merck, a science and technology leader, has long been closely aligning research and development with the adherence to ethical principles.

- Since more than 10 years the company has been integrating ethical principles into all its business activities and became an industry thought leader in applied bioethics

- driven by arising questions on Ethical Handling of Data, Algorithms and AI, establishment of an external Digital Ethics Advisory Panel and the development of the Code of Digital Ethics to strengthen trust in new digital technologies and business models.

People, machines, data, and processes are becoming ever more closely interconnected, and important key technologies such as data analytics are rapidly driving forward the digital transformation. Therefore, for Merck, handling data and algorithms (including AI) strategically is key for its future success. For this reason, Merck is expanding its data and analytics capabilities with a sophisticated data strategy. The data and analytics teams enable quicker and better-informed decision-making, which in turn makes it possible to offer customers, healthcare systems, and especially patients, innovative solutions. These new technologies and capabilities however do trigger with new ethical questions companies such as Merck: how to ethically handle data and algorithms. The new field of digital ethics demonstrates how digital transformation can be shaped in accordance with ethical principles. From Merck's perspective, it is a matter of making these ethical principles an integral part of a value-based corporate governance. This is what strengthens the trust in our business models and thus lays the foundation for broad societal acceptance. Merck intends to achieve this with the guiding principles of the Code of Digital Ethics (CoDE). This framework defines ethical principles that provide orientation when it comes to handling data and algorithms as well as introducing new technologies at Merck. It is the result of a scientific analysis of existing digital-ethical structures and frameworks in a collaboration with experts from the University of Witten/Herdecke. The CoDE is based on five core principles: autonomy, justice, beneficence, non-maleficence, and transparency. Each core principle is defined according to three sub-principles, for example autonomy through explainability, privacy, and digital skills. The principles of the CoDE [1] provide practical orientation for how the company can act as a responsible user in a hyperconnected world. By applying the principles of the CoDE, Merck ensures that its various business sectors and individual employees working in new fields and innovations act in a values-based manner. The CoDE thus serves as a tool for ethical risk assessment in existing business fields and when implementing "ethics checkpoints" for newly established digital solutions.

After a successful pilot, Merck is implementing an evaluation framework embedded in its data analytics platforms, which project leaders can use as a basis to assess whether a project raises ethical issues. Although other companies have already researched the strong societal

impacts of the digital transformation and the handling of science and health data, a well-founded approach for enabling ethical conduct related to this in business operations is still lacking. Merck is aiming to close this gap by introducing the CoDE, which has been published in the international journal AI & Society [2], thus setting a standard for the industry. The CoDE was created in collaboration with - and meant to be used by – it's Digital Ethics Advisory Panel, which Merck established in 2021. The panel includes renowned scientific and industrial experts from Europe and the United States who advise Merck on the topics of digital ethics, legal regulations and regulatory requirements, Big Data technologies, and digital health as well as medical ethics and data governance. The CoDE serves as a guideline for the panel to evaluate digital-ethical questions. The panel is tasked with applying its expertise on ethical questions concerning the use of data, algorithms, and new digital technologies for the benefit of the company.

With the panel and the CoDE in place, Merck ensures that the company is developing digital innovations in a responsible manner and taking potential ethical questions into account in all its business sectors, striving towards best practice in the new field of digital ethics.

Dr. Jean Enno Charton *is Director Digital Ethics & Bioethics at Merck KGaA. In this role, he is responsible for all digital ethics and bioethics questions arising from company's activities worldwide, reporting directly into Board of Directors.*
He joined Merck in 2014 and held several roles in Research and Development Healthcare, most recently as Chief of Staff to the Chief Medical Officer, who was responsible for the Medical Affairs and Pharmacovigilance
Prior to joining Merck, Jean Enno Charton worked at Smith and Nephew Inc. after receiving his PhD in Immuno-Oncology from the University of Lausanne, Switzerland. He holds a degree in biochemistry from the University of Tübingen, Germany, and has research experience at Harvard Medical School (primary immunodeficiency, PID) and the Canadian Science Centre for Human and Animal Health (Ebola), among others.

https://www.linkedin.com/in/jean-enno-charton-71210269/

[1] https://www.merckgroup.com/company/responsibility/us/products-businesses/CoDE-Code_of_Digital_Ethics.pdf

[2] SJ Becker et al. AI & Society. 2022

Inspiring Trust in AI for Customers

Striking the balance is key: Promoting AI-enabled innovation and ensuring trustworthy AI technology

Lisa Bechtold

At Zurich, we are aspiring to use the full potential of AI with innovative digital offerings to improve the customer journey wherever possible. For this, we have established an AI governance framework to ensure our AI operations adhere to high-quality standards with a focus on customer benefit and trust.

The successful operation of an AI governance framework requires an integrated perspective and tracking of high-pace technological progress (e.g., Chat GPT) and an increasingly complex regulatory landscape (data / privacy / AI), evolving at a rapid pace (e.g., EU Digital Strategy). A balanced and risk-based approach is critical as governance should enable (and not impede) AI-driven innovation …

Applying high quality and ethics standards to AI translates into a win-win situation for both customers and companies (insurers). When assessing AI solutions with respect to ethical values like fairness and transparency (including explainability) it is important to generate tangible customer value. For example, what form and what level of transparency is actually considered helpful for a particular stakeholder or customer group?

In the overall AI & Ethics debate, there is a risk of mixing the mitigation of AI risks (technology level) with a general discussion on fairness, etc. (policy level). It is important to acknowledge that AI may be the trigger to revisit specific policy topics but it is industry-specific – and NOT AI regulation – that is to address such policy questions.

In the regulatory and public policy debate, there is a great focus on the risks that might be triggered by AI, while all the convenience and seemingly endless opportunities AI can bring about are not equally highlighted. It's key to draw the attention of customers and the general public to all such benefits to provide for a positive and balanced view on digital innovation and the values that technology can unlock for society at large.

At Zurich, we are aspiring to use the full potential of AI with innovative digital offerings to improve the customer journey wherever possible. As we are putting the customer at the heart of everything we do, the responsible use of data and AI is paramount. We need to make sure that our use of AI is, first and foremost, responsible, ethically sound and complies with applicable laws and regulations. In addition, the deployment of AI needs to be underpinned by a risk-based governance and assurance framework that provides for appropriate risk and compliance assessments, effective monitoring and implementation (end-to-end). Our guiding principles are fairness (to mitigate bias), transparency, accountability, complemented by robustness and security. As a responsible organization, we need a strong commitment to align, foster, and scale values-led decision-making which builds trust and inspires confidence with both internal and external stakeholders.

At Zurich, we have established such an AI governance framework to ensure our AI operations adhere to high-quality standards with a focus on customer benefit and trust. We firmly believe that applying high quality and ethics standards to AI translates into a win-win situation for both customers and companies (insurers). In order to unlock the full potential of data and AI, we are committed to inspire confidence in a digital society.

Dr. Lisa Bechtold, LL.M. (Berkeley)
Global Lead AI Assurance & Data Governance at Zurich Insurance Group Switzerland
As Global Lead of Data Governance & Oversight within Zurich's Data & Business Intelligence function, Lisa drives the Group-wide implementation of AI and data governance with a focus on customer value generation through innovation and digital ethics. By applying an integrated perspective on strategy, risk management, governance and sustainability considerations, she represents Zurich in the public policy discourse on privacy, data and AI governance.

Lisa started her career with Zurich in 2009 in Group Legal where she held various roles in the areas of Corporate Finance, M&A, and Corporate Governance. In 2019, she transitioned to Group Risk Management as Head of Digital & Resilience Risk Governance, managing a broad spectrum of data, AI and related digital risks. Prior to joining Zurich, Lisa worked as an attorney-at-law with both in a boutique law firm in Germany and an international law firm in New York.
Lisa obtained a law degree and a Ph.D. in international law from the University of Cologne, Germany, an LL.M. degree from the University of California at Berkeley and completed an executive education program at MIT Sloan School of Management.

AI in Healthcare

Bryn Roberts

Dr Bryn Roberts
Global Head of Data & Analytics
Roche Information Solutions
F.Hoffmann-La Roche Ltd
Diagnostics Division
Basel
SWITZERLAND

AI models are powerful research tools that are used extensively in the lifesciences and healthcare. In addition, they are increasingly being deployed in healthcare practice as medical devices, under the SaMD (Software as a Medical Device) framework.
When AIs are influencing healthcare decisions and outcomes it is critical that these models are reliable and can be trusted in their area of application. Clearly, as with all medical interventions and devices, robust and relevant evidence needs to be generated to prove that the models perform in line with their intended use. Bias needs to be minimised during model development, through the use of representative datasets and best-practices in model development and optimisation.
To ensure that the models continue to perform as expected, on-market surveillance and monitoring are required. As the addressable populations, indications and input data evolve over time, transfer learning and retraining of the models ensures that these AIs remain fair and trustworthy, continuing to deliver value to individuals and healthcare systems overall.

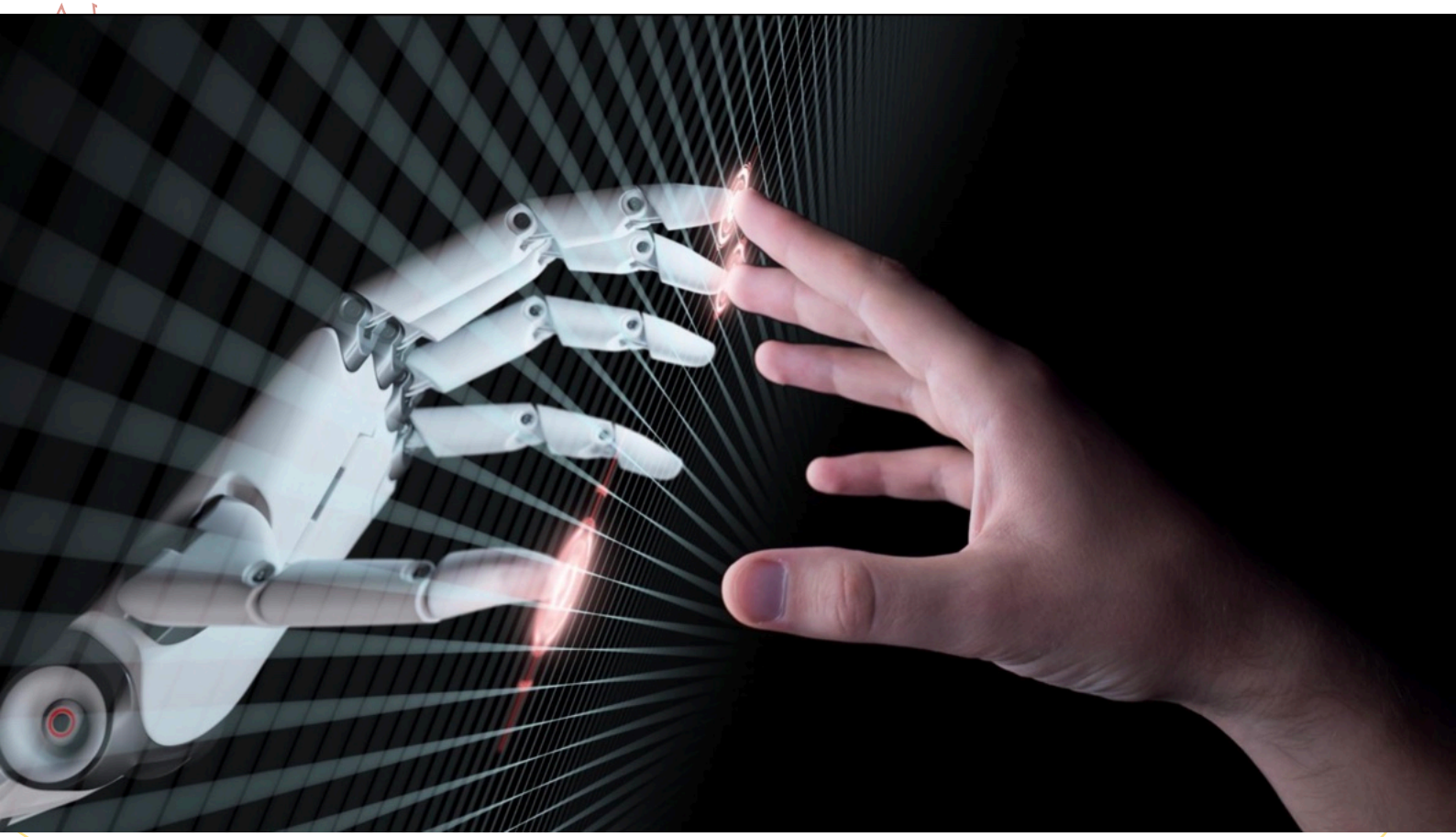
Bryn Roberts has a PhD in Pharmacology and a background in Data Science. He joined Roche in Basel in 2006 and, in his current role as Global Head of Data & Analytics within Roche Diagnostics, Bryn's accountabilities include data strategy, architecture and governance, data engineering, and data science. Beyond Roche, Bryn is a Visiting Fellow at the University of Oxford, with interests in AI and machine learning, systems biology, and scientific software development. He is an Associated Faculty member with the University of Frankfurt Big Data Lab and lectures in medical informatics at the University of Applied Sciences, NW Switzerland. He is a member of several advisory boards including the Pistoia Alliance, University of Oxford Dept of Statistics and SABS Centre for Doctoral Training, the Microsoft Research/University of Trento Center for Computational and Systems Biology, and RoX Health, and is a non-executive director with Deepmatter.

Speakers:

Gerard Kema
Innovation Manager
Province of Fryslân
g.w.kema@fryslan.frl

Willy Tadema
AI Ethics Lead
Ministry of the Interior and
Kingdom Relations
w.tadema@rijksoverheid.nl

Marijke ter Steege
Sr. Consultant Data & Strategy
Ministry of the Interior and
Kingdom Relations
marijke.steege@rijksoverheid.nl




HELP!

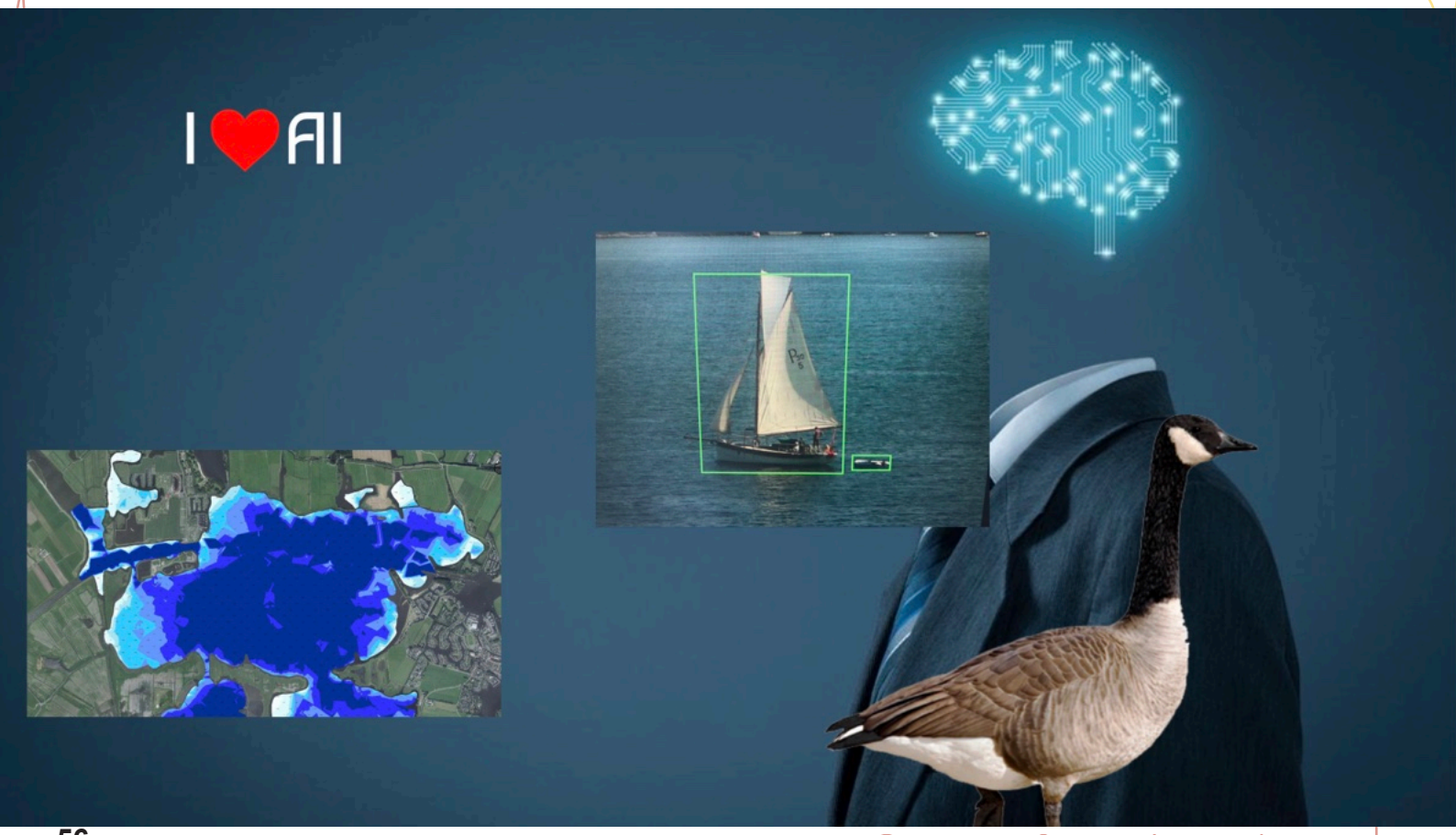
Kick off
May 2022

Assessment
June – December 2022

Training
- Ethical framework
- Z-Inspection method

May 2022

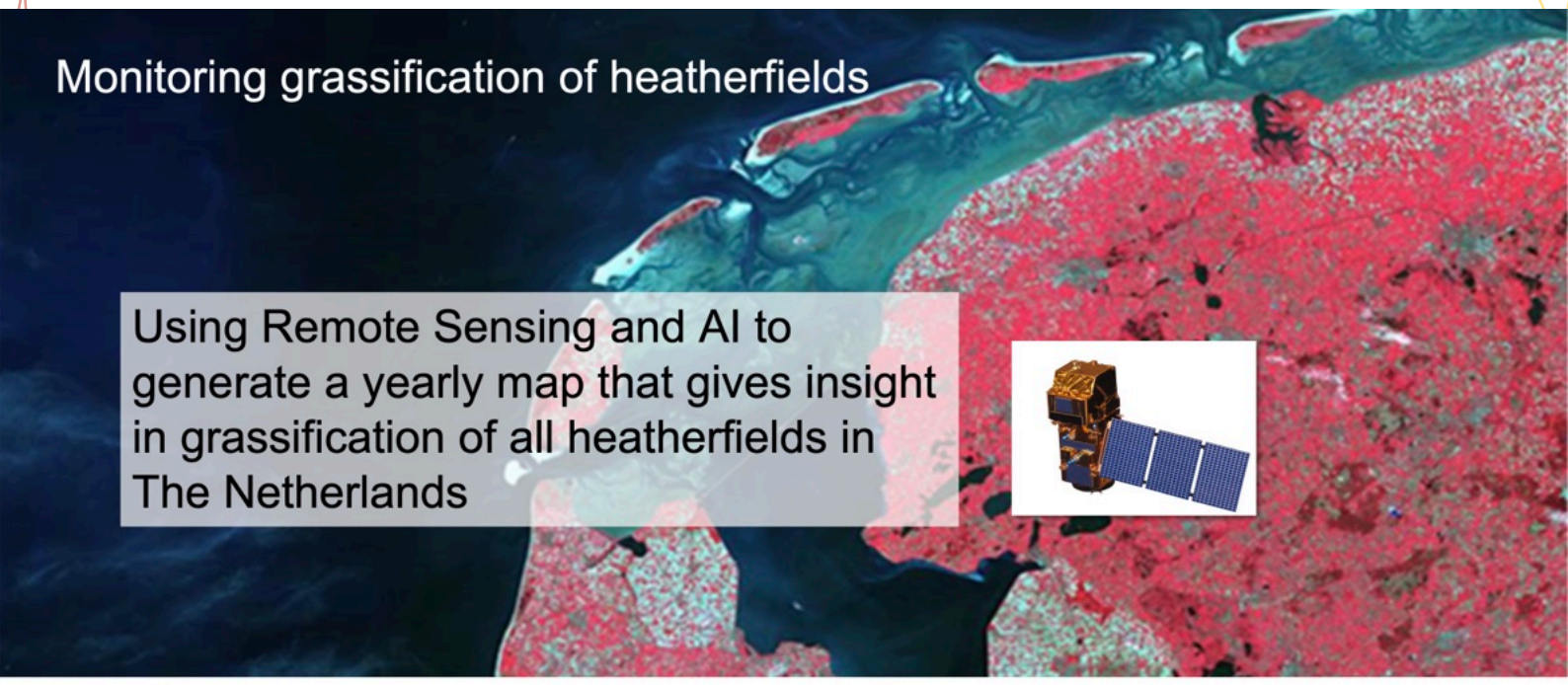
Wrap up
Presenting and publishing results
January - March



Natura 2000

"Natura 2000 is a European network of protected natural areas. In these Natura 2000 areas, certain animals, plants and their natural habitats are protected in order to preserve biodiversity (diversity)."

- Designate
- Monitor
- Maintain



Ministerie van Binnenlandse Zaken en Koninkrijksrelaties

provincie Drenthe

Provincie Noord-Brabant

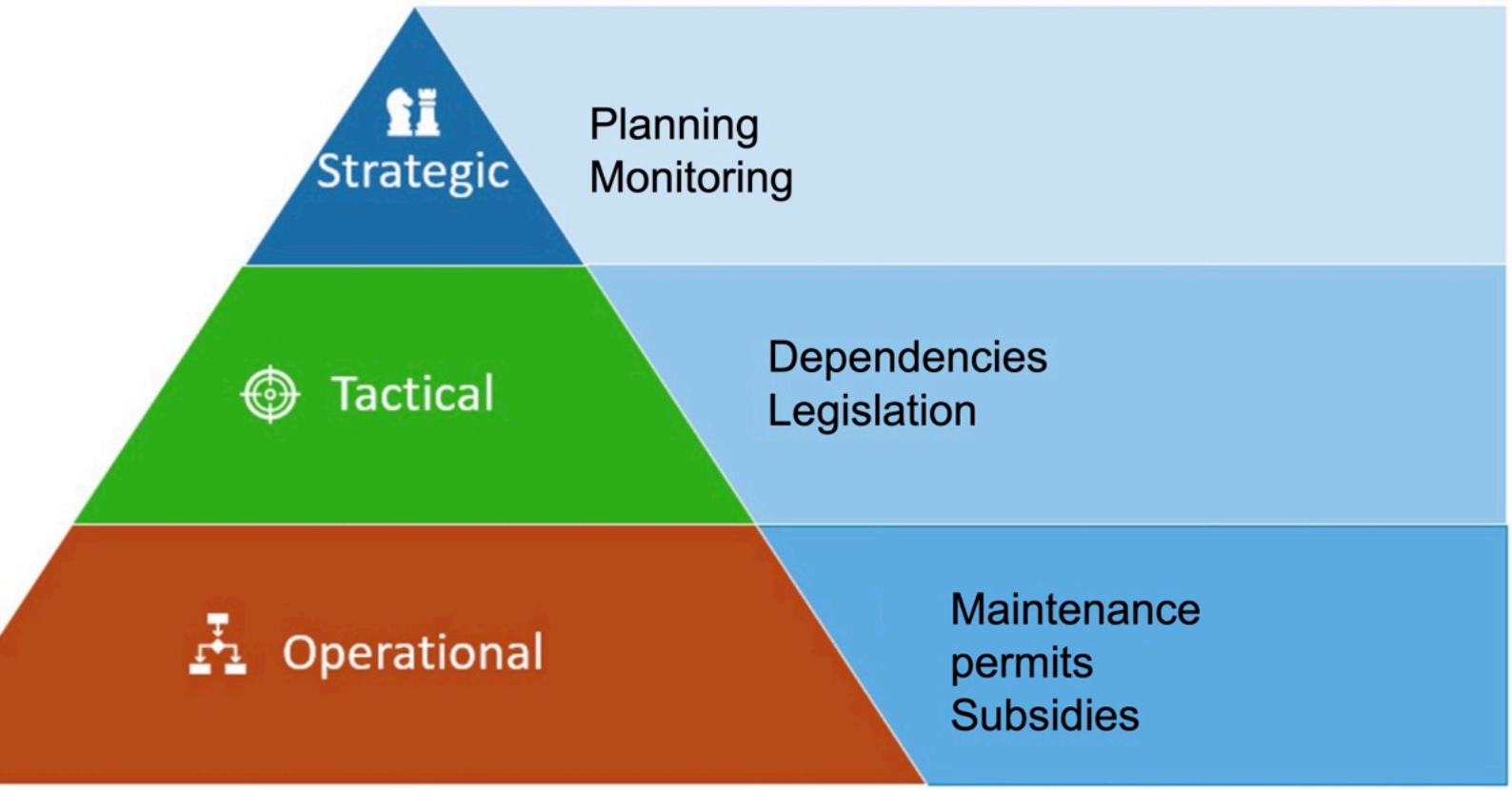
provinsje fryslân
provincie fryslân

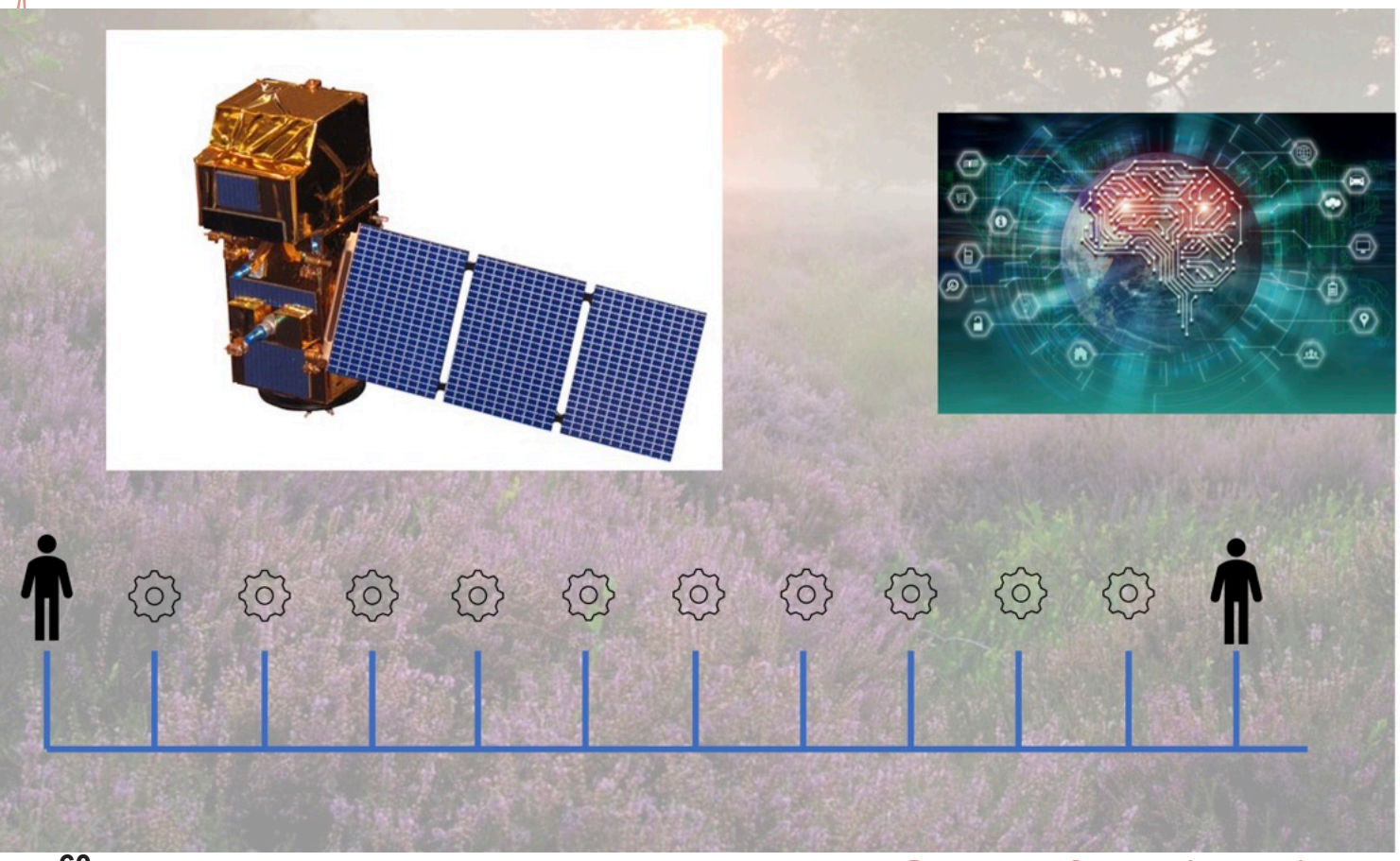
BJ12 Werkt voor provincies

provincie groningen

provincie :: Utrecht

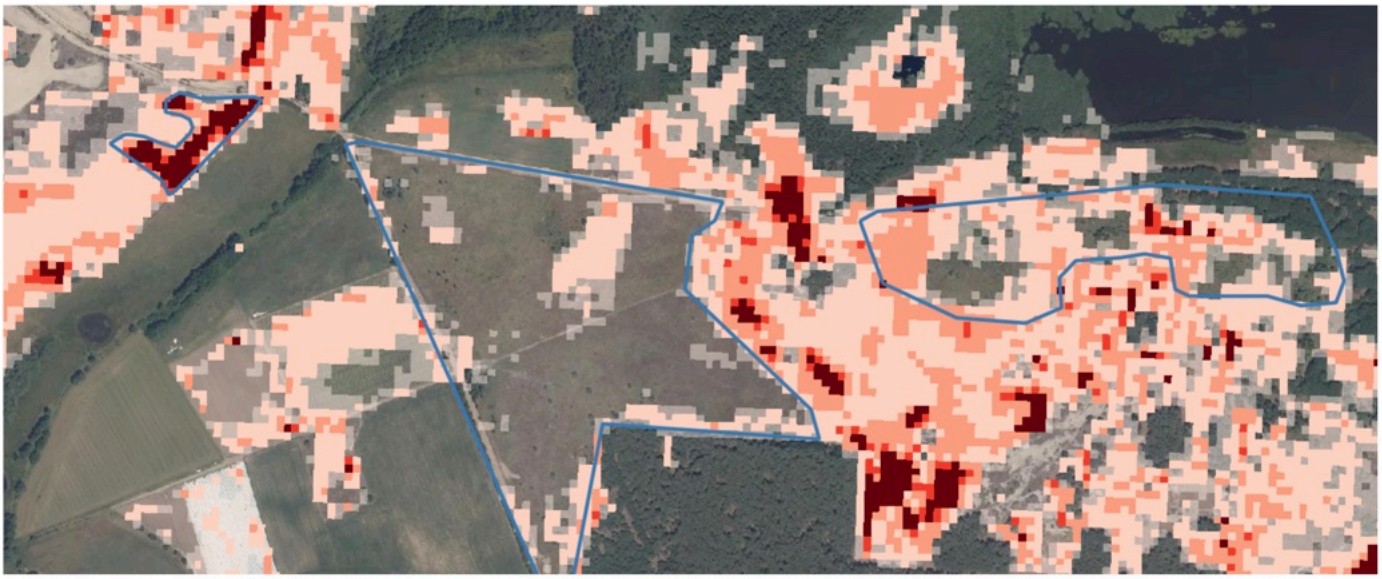

Molinia
Caerulea
Avenella
Flexuosa


Strategic
Planning
Monitoring
Tactical
Dependencies
Legislation
Operational
Maintenance
permits
Subsidies

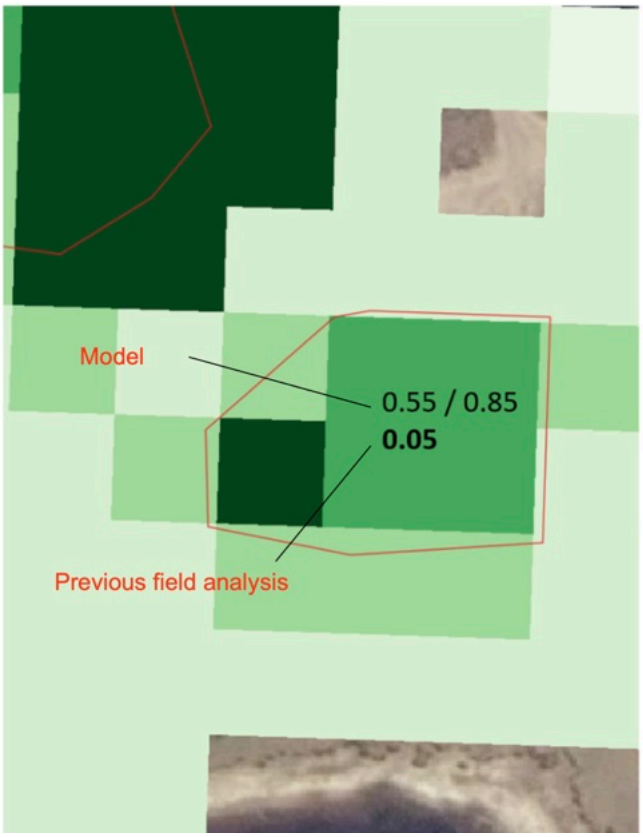


Prediction

Transparrent: < 25%.
Pink: > 25%, < 50%.
Lightred: > 50%, < 75%.
Darkred: > 75%.



Model vs kartering



Bevinding ecoloog tijdens veldvalidatie

Validation

* In vitro:
Comparing model-predictions to unseen Vegetationmaps

	Nauwkeurigheid	F1-score	Precision gras	Precision overig	Recall gras	Recall overig
Veluwe Kootwijk	83%	83	78%	87%	84%	82%
Dwingelderveld	70%	72	52%	97%	96%	60%
Groote Peel	76%	77	95%	51%	72%	88%
Totaal	76%	76	79%	72%	77%	75%

Validation

* In vitro:
Comparing model-predictions to unseen Vegetationmaps

* In situ: Field-validating model-preditions at different locations in Drenthe and Noord-Brabant

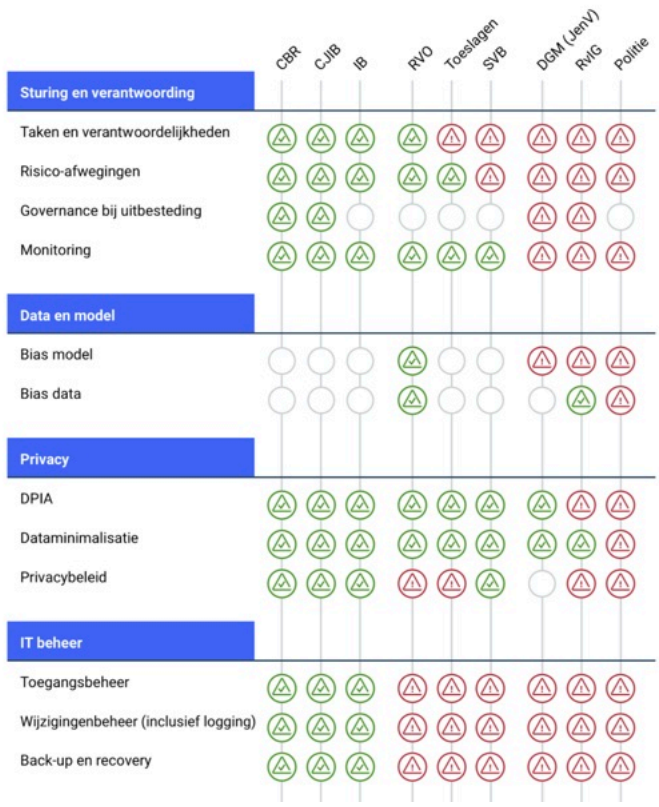
Framework for responisble AI

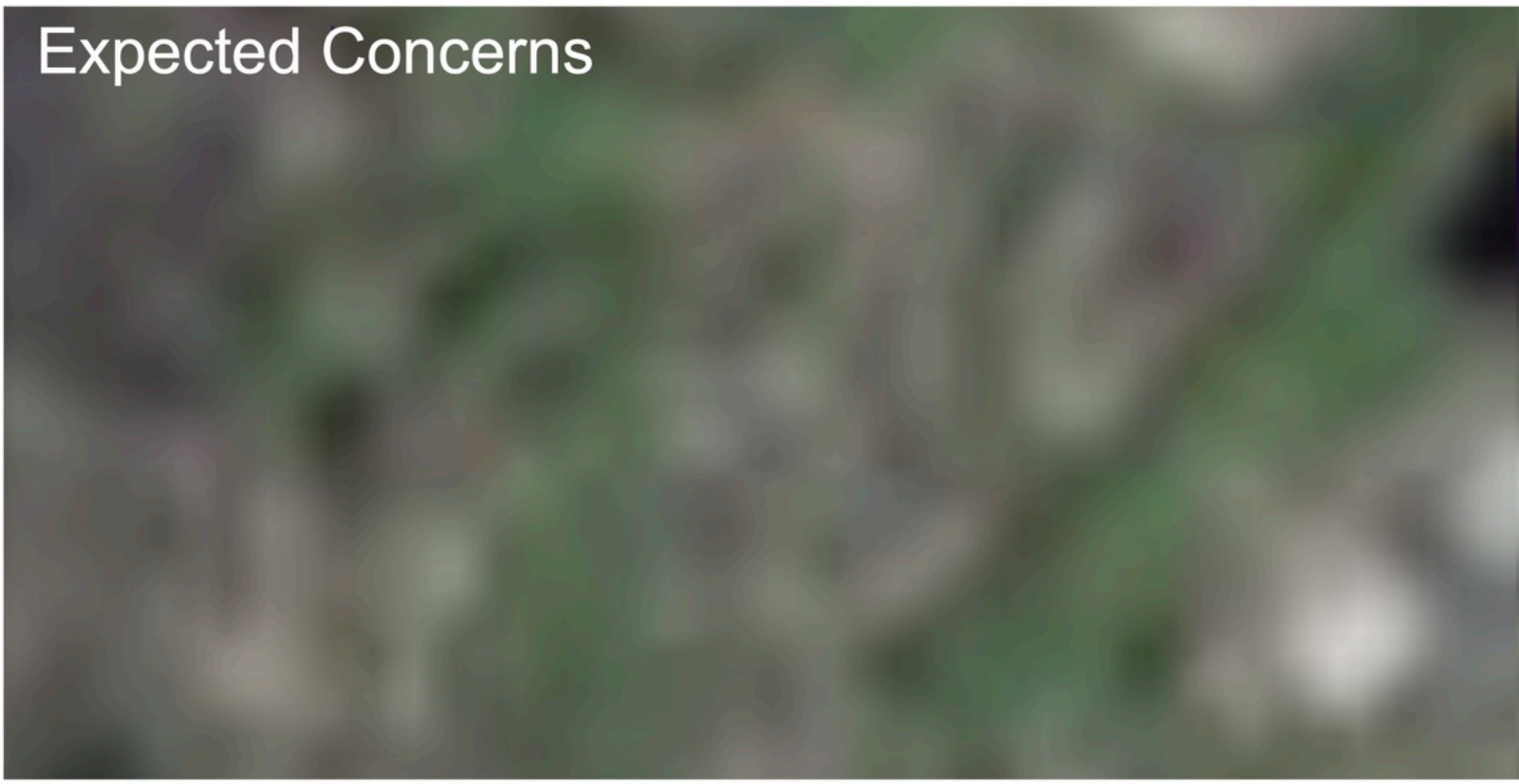
Accountability

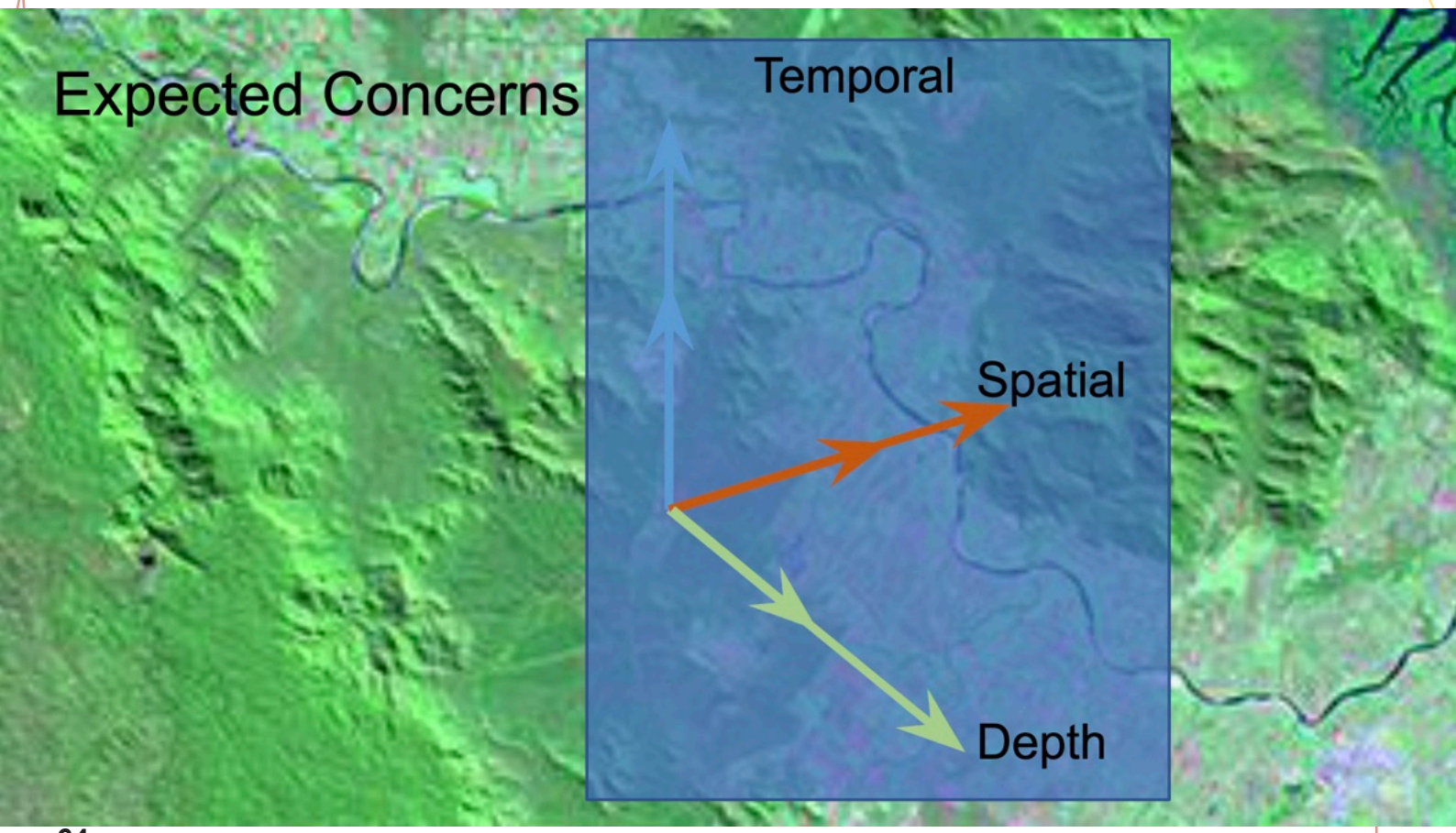
Data

Privacy

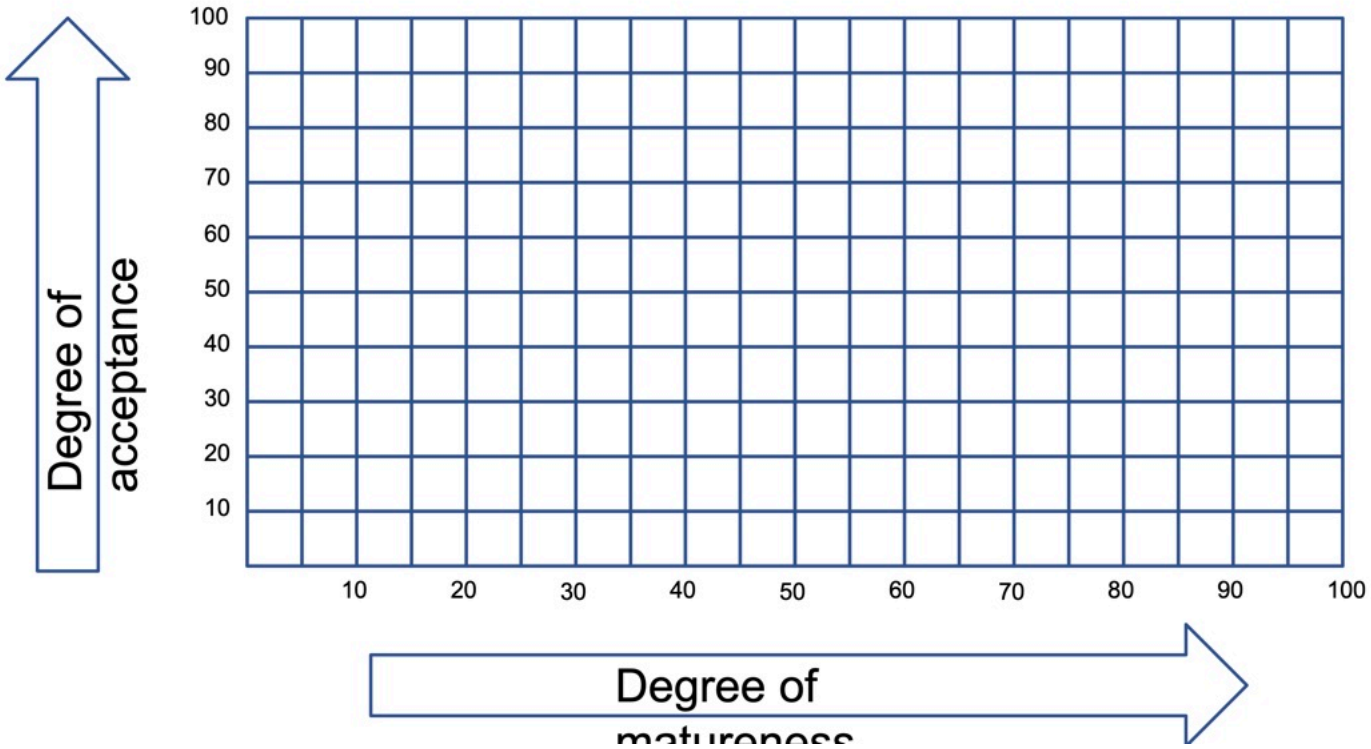
IT



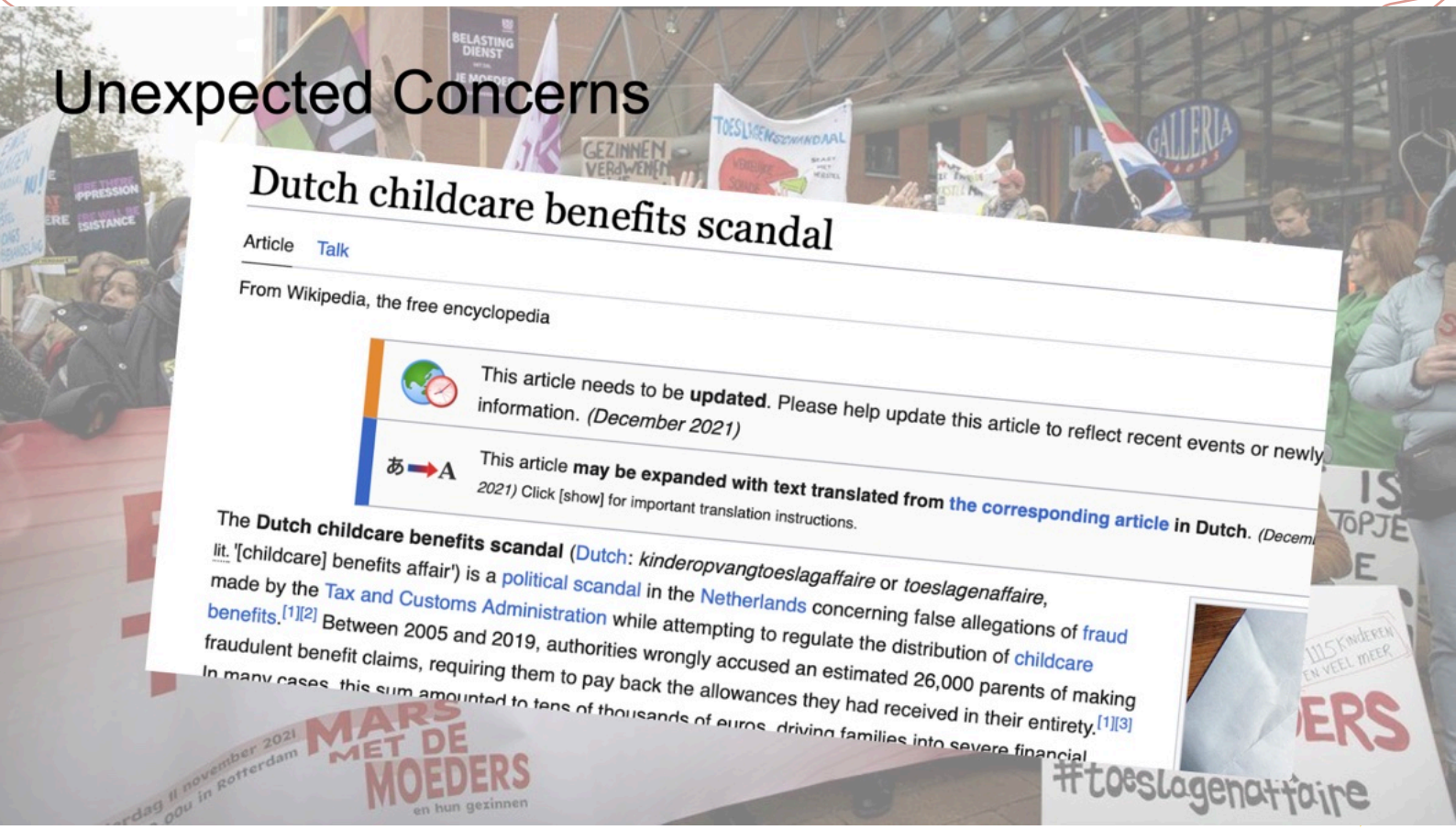

Expected Concerns


Expected Concerns
Temporal
Spatial
Depth

Expected Concerns

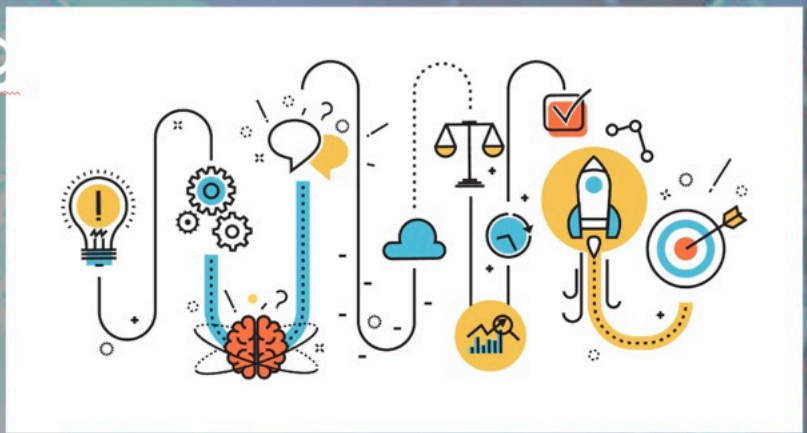


Unexpected Concerns



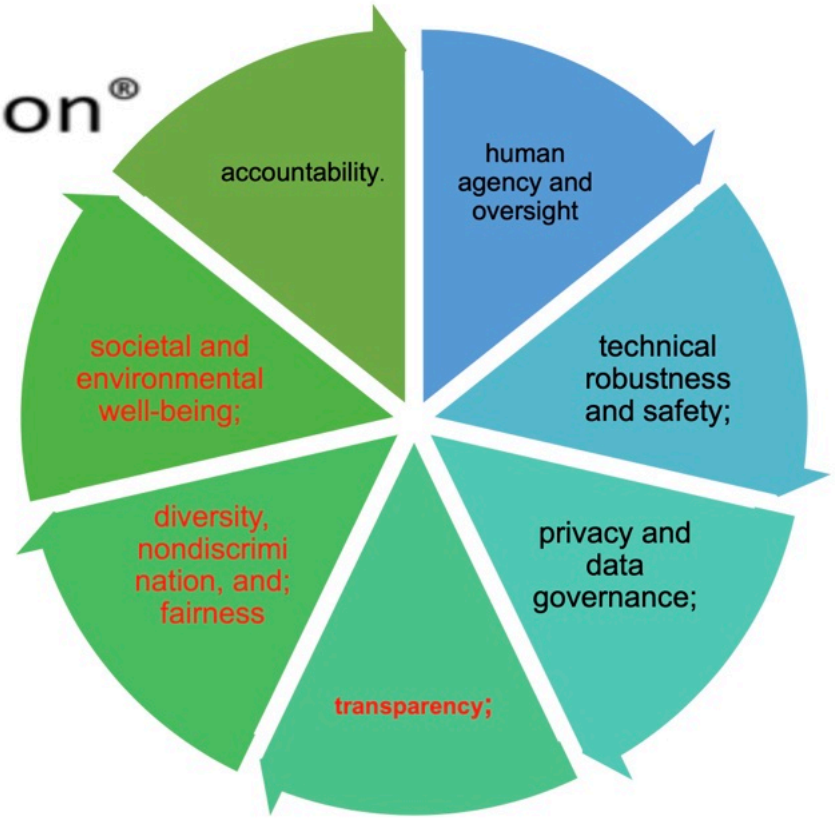
Trust

1) Assessment Heidevergrassing



2) Process & framework

Z-Inspection®



Ecology team

Technical team

Rights/ethics team

Key learningpoints

Trust extends way beyond KPI's and hard metrics

Involve stakeholders from the beginning in every part of the (design) process and determine viewpoints, needs and possible conflicts

Multi-level perspectives from an interdisciplinary team aid in determining the broadness of the problem definition

Underdstanding that the purpose of the system, subsystems and outcomes need to be translated to claims that needs to be validated.

Privacy
vs
Safety

IMPRINT

Publisher:
Z-inspection® Initiative
https://z-inspection.org/imprint/

Design: Robert Freudenreich
Pictures: Venice Urban Lab