Z-INSPECTION® CLAIM, ARGUMENTS AND EVIDENCE

MAGNUS WESTERLUND Z-INSPECTION® INITIATIVE <u>HTTP://Z-INSPECTION.ORG</u>

ARCADA UNIVERSITY OF APPLIED SCIENCES, HELSINKI, FINLAND KRISTIANIA UNIVERSITY COLLEGE, OSLO, NORWAY

Z-INSPECTION® IS A REGISTERED TRADEMARK.

THE CONTENT OF THIS WORK IS OPEN ACCESS DISTRIBUTED UNDER THE TERMS AND CONDITIONS OF THE CREATIVE COMMONS (ATTRIBUTION-NONCOMMERCIAL-SHAREALIKE CC BY-NC-SA) LICENSE (HTTPS://CREATIVECOMMONS.ORG/LICENSES/BY-NC-SA/4.0/)

Credits: Prof. Dr. Roberto V. Zicari 33



CONCEPT BUILDING

"AN IMPORTANT OBSTACLE TO PROGRESS ON THE ETHICAL AND SOCIETAL ISSUES RAISED BY AI-BASED SYSTEMS IS THE AMBIGUITY OF MANY CENTRAL CONCEPTS CURRENTLY USED TO IDENTIFY SALIENT ISSUES.,,

- TERMINOLOGICAL OVERLAPS
- DIFFERENCES BETWEEN DISCIPLINES
- DIFFERENCES ACROSS CULTURES AND PUBLICS
- CONCEPTUAL COMPLEXITY

Source: Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), London. Nuffield Foundation.



CONCEPT BUILDING

 MAPPING AND CLARIFYING AMBIGUITIES
BRIDGING DISCIPLINES, SECTORS, PUBLICS AND CULTURES

3. Building consensus and managing disagreements

Source: Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S (2019), London. Nuffield Foundation.

WE DEVELOP AN EVIDENCE BASE

This is an iterative process among experts with different skills and background with goal to:

- Understand technological capabilities and limitations
- BUILD A STRONGER EVIDENCE BASE TO SUPPORT CLAIMS AND IDENTIFY TENSIONS (DOMAIN SPECIFIC)
- UNDERSTAND THE PERSPECTIVE OF DIFFERENT MEMBERS OF SOCIETY

SOURCE: WHITTLESTONE, J ET AL (2019)

<u>**CLAIMS**</u>, ARGUMENTS AND EVIDENCE (CAE)

CLAIMS – "ASSERTIONS PUT FORWARD FOR GENERAL ACCEPTANCE. THEY ARE TYPICALLY STATEMENTS ABOUT A PROPERTY OF THE SYSTEM OR SOME SUBSYSTEM.

CLAIMS THAT ARE ASSERTED AS TRUE WITHOUT JUSTIFICATION BECOME **ASSUMPTIONS** AND CLAIMS SUPPORTING AN ARGUMENT ARE CALLED SUB CLAIMS. "

Source: – Brundage et al. (2020) – Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims.

CLAIMS, ARGUMENTS AND <u>EVIDENCE</u> (CAE)

EVIDENCE "THAT IS USED AS THE BASIS OF THE JUSTIFICATION OF THE CLAIM.

Sources of evidence may include the design, the development process, prior field experience, testing, source code analysis or formal analysis", peer-reviewed journals articles, peer-reviewed clinical trials, etc.

Source: – Brundage et al. (2020) – Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims.

CLAIMS, <u>ARGUMENTS</u> AND EVIDENCE (CAE)

ARGUMENTS LINK THE EVIDENCE TO THE CLAIM.

They are defined as Toulmin's warrants and are the "statements indicating the general ways of arguing being applied in a particular case and implicitly relied on and whose trustworthiness is well established", together with the validation for the scientific and engineering laws used.

Source: – Brundage et al. (2020) – Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims.

DEVELOP AN **Evidence base**

 TECHNOLOGY IS GENERALLY DESIGNED FOR A HIGHLY SPECIFIC PURPOSE A CONTEXT. HOWEVER, IT IS NOT ALWAYS CLEAR WHAT THE TECHNOLOGY'S UNINTENDED HARM MIGHT BE.

 THEREFORE, AN IMPORTANT PART OF OUR ASSESSMENT PROCESS IS TO BUILD AN EVIDENCE BASE THROUGH THE SOCIO-TECHNICAL SCENARIOS TO IDENTIFY TENSIONS AS POTENTIAL ETHICAL ISSUES.

IDENTIFY CLAIMS

 "AI DEVELOPERS REGULARLY MAKE CLAIMS REGARDING THE PROPERTIES OF AI SYSTEMS THEY DEVELOP AS WELL AS THEIR ASSOCIATED SOCIETAL CONSEQUENCES. "

SOURCE: TOWARD TRUSTWORTHY AI DEVELOPMENT: MECHANISMS FOR SUPPORTING VERIFIABLE CLAIMS

HTTPS://ARXIV.ORG/PDF/2004.07213.PDF

- CLAIMS FOR TECHNOLOGICAL CAPABILITY (FOR EXAMPLE AIM, PERFORMANCE, ARCHITECTURE, OR FUNCTIONALITY, ETC.) SERVE AS AN IMPORTANT INPUT IN DEVELOPING THE EVIDENCE BASE.
- This is an iterative process among experts of the assessment team with different skills and backgrounds with a goal to understand technological capabilities and limitations

VERIFIABLE CLAIMS

- "VERIFIABLE CLAIMS ARE STATEMENTS FOR WHICH EVIDENCE AND ARGUMENTS CAN BE BROUGHT TO BEAR ON THE LIKELIHOOD OF THOSE CLAIMS BEING TRUE.
- The degree of attainable certainty in such claims will vary across contexts. ,,

• SOURCE: TOWARD TRUSTWORTHY AI DEVELOPMENT: MECHANISMS FOR SUPPORTING VERIFIABLE CLAIMS HTTPS://ARXIV.ORG/PDF/2004.07213.PDF

EXAMPLES OF CLAIMS

We will adhere to the data usage protocols we have specified

The cloud services on which our AI systems run are secure

We will evaluate risks and benefits of publishing AI systems in partnership with appropriately qualified third parties

The AI system is very accurate...

The AI system is more accurate then....

The AI system is 98% accurate...

The AI predicts with high quality

Using the AI system results in saving XXX dollars

We will not create or sell AI systems that are intended to cause harm

We will assess and report any harmful societal impacts of AI systems that we build

Broadly, we will act in a way that aligns with society's interests.

"KEEP YOUR AI CLAIMS IN CHECK" US FTC

ARE YOU EXAGGERATING WHAT YOUR AI PRODUCT CAN DO?

- OR EVEN CLAIMING IT CAN DO SOMETHING BEYOND THE CURRENT CAPABILITY OF ANY AI OR AUTOMATED TECHNOLOGY?
- FOR EXAMPLE, WE'RE NOT YET LIVING IN THE REALM OF SCIENCE FICTION, WHERE COMPUTERS CAN GENERALLY MAKE TRUSTWORTHY PREDICTIONS OF HUMAN BEHAVIOR.
- Your performance claims would be deceptive if they lack scientific support or if they apply only to certain types of users or under certain conditions.

ARE YOU PROMISING THAT YOUR AI PRODUCT DOES SOMETHING BETTER THAN A NON-AI PRODUCT?

 IT'S NOT UNCOMMON FOR ADVERTISERS TO SAY THAT SOME NEW-FANGLED TECHNOLOGY MAKES THEIR PRODUCT BETTER — PERHAPS TO JUSTIFY A HIGHER PRICE OR INFLUENCE LABOR DECISIONS. YOU NEED ADEQUATE PROOF FOR THAT KIND OF COMPARATIVE CLAIM, TOO, AND IF SUCH PROOF IS IMPOSSIBLE TO GET, THEN DON'T MAKE THE CLAIM.

> Source: Keep your AI claims in check, By Michael Atleson, Attorney, US Federal Trade Commision Division of Advertising Practices February 27, 2023

"KEEP YOUR AI CLAIMS IN CHECK" US FTC

ARE YOU AWARE OF THE RISKS?

You need to know about the reasonably foreseeable risks and impact of your AI product before putting it on the market. If something goes wrong – maybe it fails or yields biased results – you can't just blame a third-party developer of the technology. And you can't say you're not responsible because that technology is a "black box" you can't understand or didn't know how to test.

DOES THE PRODUCT ACTUALLY USE AI AT ALL?

IF YOU THINK YOU CAN GET AWAY WITH BASELESS CLAIMS THAT YOUR PRODUCT IS AI-ENABLED, THINK AGAIN. IN AN INVESTIGATION, FTC TECHNOLOGISTS AND OTHERS CAN LOOK UNDER THE HOOD AND ANALYZE OTHER MATERIALS TO SEE IF WHAT'S INSIDE MATCHES UP WITH YOUR CLAIMS. BEFORE LABELING YOUR PRODUCT AS AI-POWERED, NOTE ALSO THAT MERELY USING AN AI TOOL IN THE DEVELOPMENT PROCESS IS NOT THE SAME AS A PRODUCT HAVING AI IN IT.

> Source: Keep your Al claims in check, By Michael Atleson, Attorney, US Federal Trade Commission Division of Advertising Practices February 27, 2023

ENSURING THE VERIFIABILITY OF CLAIMS IN AI DEVELOPMENT IS HIGHLY DESIRABLE

- "FIRST, THOSE POTENTIALLY AFFECTED BY AI DEVELOPMENT—AS WELL AS THOSE SEEKING TO REPRESENT THOSE PARTIES' INTERESTS VIA GOVERNMENT OR CIVIL SOCIETY—DESERVE TO BE ABLE TO SCRUTINIZE THE CLAIMS MADE BY AI DEVELOPERS IN ORDER TO REDUCE RISK OF HARM OR FOREGONE BENEFIT."
- "Second, to the extent that claims become verifiable, various actors such as civil society, policymakers, and users can raise their standards for what constitutes responsible AI development.
- This, in turn, can improve societal outcomes associated with the field as a whole."
- "THIRD, A LACK OF VERIFIABLE CLAIMS IN AI DEVELOPMENT COULD FOSTER OR WORSEN A "RACE TO THE BOTTOM" IN AI DEVELOPMENT, WHEREBY DEVELOPERS SEEK TO GAIN A COMPETITIVE EDGE EVEN WHEN THIS TRADES OFF AGAINST IMPORTANT SOCIETAL VALUES SUCH AS SAFETY, SECURITY, PRIVACY, OR FAIRNESS "

WHAT IS EVIDENCE?

- WHO IS "QUALIFIED" TO GIVE STRONG EVIDENCE? WE COULD INTRODUCE DIFFERENT LEVELS OF WHAT CONSTITUTES "EVIDENCE".
- Strong evidence is when testing is possible. However, testing is not always possible. We look at peer-reviewed journal articles supporting a claim. This is also evidence.
- WHEN DOMAIN EXPERTS HAVE DIFFERENT VIEWPOINTS, THEN WE LIST SUCH DIFFERENT VIEWPOINTS AND RELATED SUPPORTING EVIDENCE AS TENSIONS.

DIFFERENT VIEWPOINTS

- EXPERTS IN DIFFERENT FIELDS WILL SEE THE AI SYSTEM QUITE DIFFERENTLY. WHAT MAY BE CONSIDERED A LACK OF KNOWLEDGE CAN JUST BE A DIFFERENT LENS.
 IT'S CRUCIAL THE TEAM UNDERSTANDS THAT THERE WILL BE VERY DIFFERENT PERSPECTIVES BASED ON THE SPECIFIC ROLE OR SUBDOMAIN DIFFERENT EXPERTS REPRESENT.
- MANAGING DIFFERENT VIEWPOINTS BETWEEN EXPERTS COMPOSING THE ASSESSMENT TEAM IS AN ESSENTIAL PART OF THE PROCESS.
- One of the key lessons learned is that there may be tensions when considering what the relevant existing evidence to support a claim is.

IDENTIFYING "ISSUES"

 A CLAIM WITH NO EVIDENCE BECOMES AN ASSUMPTION, AND THIS COULD BE A POTENTIAL RISK.

WE CALL THEM "ISSUES".

- How to describe "issues"?
- Use free text and an open vocabulary

EXAMPLE: TENSIONS IN EVIDENCE BASE

CASE OF A SKIN CANCER DETECTION AI TOOL (*):

THERE WERE *TENSIONS* BETWEEN THE VARIOUS ARGUMENTS LINKING EVIDENCE TO SUPPORT THE CHOICE OF A DESIGN DECISION DERIVED FROM THE DIFFERENT VIEWPOINTS EXPRESSED BY DOMAIN EXPERTS.

CLAIM:

This AI System helps dermatologists to early detection of malignant melanoma.

ARGUMENT:

Malignant melanoma is a very heterogeneous tumor with a clinical course that is very difficult to predict. To date, there are no reliable biomarkers that predict prognosis with certainty. Therefore, there exist subgroups of melanoma patients with different risks for metastasization, some might never metastasize and diagnosing them would be overdiagnosing.

> (*) – **Co-Design of a Trustworthy AI System in Healthcare: Deep Learning Based Skin Lesion Classifier.** Front. Hum. Dyn. |Human and Artificial Collaboration for Medical Best Practices, July 13, 2021

EXAMPLE: TENSIONS IN EVIDENCE BASE

- VIEW POINT DERMATOLOGIST: EARLY DETECTION OF MALIGNANT MELANOMA IS CRITICAL, AS THE RISK OF METASTASIS WITH WORSE PROGNOSIS INCREASES THE LONGER MELANOMA REMAINS UNTREATED.
- VIEW POINT EVIDENCE BASED MEDICINE PROFESSIONAL: THERE ARE NO RELIABLE BIOMARKERS THAT CAN PREDICT THE PROGNOSIS OF MELANOMA BEFORE EXCISION. THERE ARE PATIENTS WHO SURVIVE THEIR LOCALIZED MELANOMA WITHOUT THERAPY. THEREFORE, THE EARLY DIAGNOSIS DOES NOT NECESSARILY MEAN A BETTER PROGNOSIS; ON THE CONTRARY, THERE IS A RISK OF POOR PATIENT CARE DUE TO OVERDIAGNOSIS.

IN SUMMARY: BUILDING A SOLID KNOWLEDGE / EVIDENCE BASE

- WE SUGGEST BUILDING A SOLID KNOWLEDGE / EVIDENCE BASE AMONG ALL TEAM MEMBERS OF THE USE CASE BEFORE THE INSPECTION STARTS AND ALSO A SOLID Q&A LOG DURING THE INSPECTION PROCESS.
- EXPERTS MAY APPROACH THE USE CASE QUITE DIFFERENTLY:
- Interpretations of and expectations for the AI tool being inspected may differ
- FOCUS OF INTEREST MAY BE VERY DIFFERENT

IN SUMMARY: CAE FRAMEWORK

- THE CLAIMS, ARGUMENTS AND EVIDENCE (CAE) FRAMEWORK (*) CAN HELP WITH THE STRUCTURING OF THE USE CASE IN A CLEAR AND PRECISE FORM THAT IS SUPPORTED BY EVIDENCE.
- A CLAIM SHOULD ONLY BE ABOUT ONE SPECIFIC PROPERTY OF THE SYSTEM
- IT SHOULD BE PHRASED IN A WAY THAT IS CLEARLY VERIFIABLE OR FALSIFIABLE.
- THE CAE FRAMEWORK ALSO PROVIDES GUIDANCE ON HOW TO DISSEMINATE COMPLEX CLAIMS INTO EASIER ONES.

(*) HTTPS://CLAIMSARGUMENTSEVIDENCE.ORG

THE Z-INSPECTION® PROCESS: CO-DESIGN

N DESIGN AND DEVELOPMENT PHASES:

Z- INSPECTION[®] CAN BE USED AS A **CO-CREATION PROCESS** TO HELP AI ENGINEERS, DOMAIN EXPERTS TO ENSURE THAT THE DESIGN OF THEIR AI SYSTEM MEETS THE TRUSTWORTHY AI CRITERIA.



WHEN IN CO-DESIGN.

- CONSIDER THE AI INITIAL DESIGN AS A CLAIM THAT NEEDS TO BE VERIFIED WITH EVIDENCE.
- EXAMPLE: WHEN DESIGNING, TRAINING AND TESTING AN AI-SYSTEM (E.G. MACHINE-LEARNING ALGORITHM) WE DO "EMBED" INTO THE SYSTEM NOTIONS SUCH AS "GOOD", "BAD", "HEALTHY", "DISEASE", ETC. MOSTLY NOT IN AN EXPLICIT/TRANSPARENT WAY.

THE Z-INSPECTION® PROCESS

IN DEPLOYMENT AND AFTER DEPLOYMENT:

Z-INSPECTION® CAN BE USED AS A VALIDATION PROCESS TO ASSESS THE TRUSTWORTHINESS OF THE AI SYSTEM BEING DEVELOPED.

Additionally, it can form part of an AI certification, audit or **monitoring process**.

THE LATTER CAN BE CONSIDERED A PART OF "ETHICAL MAINTENANCE" FOR TRUSTWORTHY AI.

WHEN THE AI IS DEPLOYED

- VERIFY CLAIMS OF THE PRODUCER OF THE AI WITH EVIDENCE
- Example: "Embedded" Ethics in AI for healthcare: Medical Diagnosis

"EMBEDDED" ETHICS IN AI FOR HEALTHCARE: MEDICAL DIAGNOSIS

"IN CASE MEDICAL DIAGNOSIS OR TREATMENT RECOMMENDATIONS ARE BEING DEFERRED TO MACHINE LEARNING ALGORITHMS, IT IS THE ALGORITHM WHO SETS THE BAR ABOUT HOW A DISEASE IS BEING DEFINED."

"The deployment of machine learning in medicine might resurge the debate between *naturalists* and *normativists*."

-- THOMAS GROTE , PHILIPP BERENS

Source: Grote T, Berens P. J Med Ethics Epub ahead of print: [please include Day Month Year]. doi:10.1136/ medethics-2019-105586

LESSONS LEARNED

THERE MAY BE TENSIONS IN BUILDING A STRONGER EVIDENCE BASE ON THE CURRENT USES AND IMPACTS (DOMAIN SPECIFIC)

DIFFERENT VIEW POINTS AMONG DOMAIN EXPERTS

• WHO IS "QUALIFIED" TO GIVE A STRONG EVIDENCE?

HOW TO HANDLE IP

- CLARIFY WHAT IS AND HOW TO HANDLE THE IP OF THE AI AND OF THE PART OF THE ENTITY/COMPANY TO BE EXAMINED.
- IDENTIFY POSSIBLE RESTRICTIONS TO THE INSPECTION PROCESS, IN THIS CASE ASSESS THE CONSEQUENCES (IF ANY)
- Define if and when Code Reviews are needed/possible. For example, check the following preconditions (*):
 - There are no risks to the security of the system
 - PRIVACY OF UNDERLYING DATA IS ENSURED
 - NO UNDERMINING OF INTELLECTUAL PROPERTY

DEFINE THE IMPLICATIONS IF ANY OF THE ABOVE CONDITIONS ARE NOT SATISFIED.

(*) Source: "Engaging Policy Shareholders on Issue in AI governance" (Google)

IMPLICATION OF IP ON TRUSTWORTHY AI

 THERE IS AN INEVITABLE TRADE OFF TO BE MADE BETWEEN DISCLOSING ALL ACTIVITIES OF A TRUSTWORTHY AI ASSESSMENT VS. DELAYING THEM TO A LATER STAGE.

Benjamin Haibe-Kains, et al. The importance of transparency and reproducibility in artificial intelligence research. (Submitted on 28 Feb 2020 (v1), last revised 7 Mar 2020 (this version, v2))

HTTPS://ARXIV.ORG/PDF/2003.00898.PDF

SUMMARY

- SOCIO-TECHNICAL SCENARIO BUILDING CAN HELP TEAMS TO THINK AHEAD AND CREATE SYSTEMS THAT A GROUNDED IN TRUSTWORTHY AI FROM THE ONSET.
- THEN CAE FRAMEWORK IS A POWERFUL TOOL TO QUESTION CLAIMS, AND TO VALIDATE THEM IN THE QUEST FOR EVIDENCE.
- THERE ARE SOME OTHER CONSIDERATIONS TO BE MADE THAT RELATES TO TRUST, SUCH AS PROTECTION OF IP AND LEGAL LIABILITY (ADDITIONAL GOVERNING LAWS) THAT MAY NEED TO BE CONSIDERED.
- Z-Inspection as a process is very versatile, and can easily accommodate Fundamental Rights assessments or technical reviews