

TOWARDS AI GOVERNANCE: REFLECTIONS ON THE DEEP DIVES

Andreas Theodorou

 andreas.theodorou@umu.se

 [@recklesscoding](https://twitter.com/recklesscoding)



UMEÅ UNIVERSITY

INCIDENTS



"Alexa, Can I Trust You?"

Hyunji Chung, Michaela Iorga, and Jeffrey Voas, NIST
Sangjin Lee, Korea University

Several recent incidents highlight significant security and privacy risks associated with intelligent virtual assistants (IVAs). Better diagnostic testing of IVA ecosystems can

For ex
6-year-old
love of dc
the famil
prompter
her paper
Krafe Ka

RESEARCH ARTICLE

Even good bots fight: The case of Wikipedia

Milena Tsvetkova¹, Ruth Garcia-Gavilanes¹, Luciano Floridi^{1,2}, Taha Yasseri^{1,2*}

¹ Oxford Internet Institute, University of Oxford, Oxford, United Kingdom, ² Alan Turing Institute, London, United Kingdom

* taha.yasseri@oii.ox.ac.uk

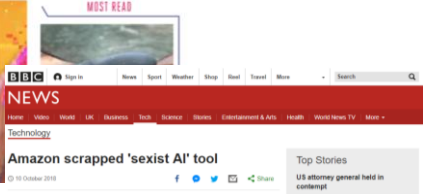
Abstract

In recent years, there has been a huge increase in the number of bots online. Web crawlers for search engines, to chatbots for online customer service, social media, and content-editing bots in online collaboration communities. T

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

By JAMES VOORHEES | Nov 24, 2016, 6:45am EST

f t+ e+ share



NEWS

Home Video World UK Business Tech Science Health Entertainment & Arts World News TV More

Technology

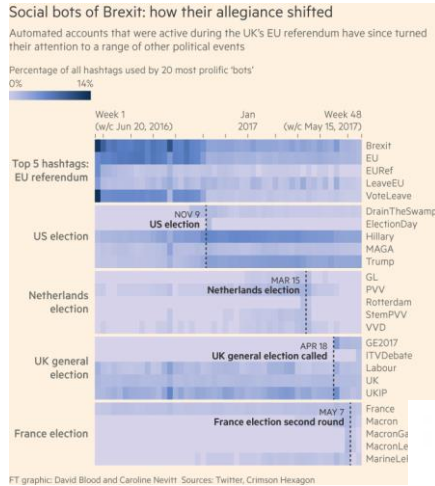
Amazon scrapped 'sexist AI' tool

© 19 October 2016

Top Stories
US attorney general held in contempt



...AND MORE INCIDENTS



Bots, #StrongerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum

COMPROP RESEARCH NOTE 2016.4

Bots and Automation over Twitter during the U.S. Election

COMPROP DATA MEMO 2016.4 / 17 NOV 2016

Bence Kollanyi
Corvinus University
kollanyi@gmail.com
@bencekollanyi

Philip N. Howard
Oxford University
philip.howard@oii.ox.ac.uk
@pnhoward

Samuel C. Woolley
University of Washington
samwool@uw.edu
@samuelwoolley

ABSTRACT
Bots are social with other user
Brexit covers automated scrip and then intera accounts that ar

ABSTRACT

Bots are social media accounts that automate interaction with other users, and political bots have been particularly active on public policy issues, political crises, and elections. We collected data on bot activity using the major hashtags posted on the 11th U.S. Presidential Election. We find that these automated bot activities increased an all-time high for the over time, but the ga the first debate to 5:1 the election, most cl content production d after Election Day.

DISINFORMATION AND SOCIAL BOT OPERATIONS IN THE RUN UP TO THE 2017 FRENCH PRESIDENTIAL ELECTION

EMILIO FERRARA
UNIVERSITY OF SOUTHERN CALIFORNIA, INFORMATION SCIENCES INSTITUTE

ABSTRACT

Recent accounts from researchers, journalists, as well as federal investigators, reached a unanimous conclusion: social media are systematically exploited to manipulate and alter public opinion. Some disinformation campaigns have been coordinated by means of bots: social media accounts controlled by



Cambridge Analytica



UMEÅ UNIVERSITY

2017 EUROBAROMETER

- **61%** of respondents have a **positive view** of robots
- **84%** of respondents agree that **robots can do jobs** that are too **hard/dangerous** for people
- **68%** agree that robots are a **good thing for society** because they help people
- **88%** of respondents consider robotics a technology that **requires careful management**
- **72%** of respondents think robots **steal people's jobs**



UMEÅ UNIVERSITY

LIKE THE ELEVATORS



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

WE NEED TO BUILD TRUST FOR OUR SYSTEMS

- To **perform as we expect them to.**
 - The implications from their development and deployment fall within:
 - **Ethical**
 - **Legal**
 - **Social**
 - **Economic**
 - **Cultural**
- (**ESLEC**) specifications and values we want to protect.



UMEÅ UNIVERSITY

AI GOVERNANCE



UMEÅ UNIVERSITY



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

EPSRC PRINCIPLES OF ROBOTICS

- 1. Robots are multi-use tools.** Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security.
- 2. Humans, not robots, are responsible agents.** Robots should be designed; operated as far as is practicable to comply with existing laws & fundamental rights & freedoms, including privacy.
- 3. Robots are products.** They should be designed using processes which assure their safety and security.
- 4. Robots are manufactured artefacts.** They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent.
- 5. The person with legal responsibility for a robot should be attributed.**



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

European Union Background on AI

EU STRATEGY ON ARTIFICIAL INTELLIGENCE

published in April 2018

Boost AI uptake

Tackle socio-economic changes

Ensure adequate ethical & legal framework



In this context: appointment of Independent High-Level Expert Group on Artificial Intelligence (AI HLEG) in June 2018



UMEÅ UNIVERSITY



European
Commission

Ethics Guidelines for AI – Requirements



Human agency and oversight



Diversity, non-discrimination and fairness



Technical Robustness and safety



Societal & environmental well-being



Privacy and data governance



Accountability



Transparency

To be continuously implemented & evaluated throughout AI system's life cycle



UMEÅ UNIVERSITY



European
Commission

HIGH-LEVEL GUIDELINES



TOP 10 PRINCIPLES
FOR ETHICAL ARTIFICIAL
INTELLIGENCE



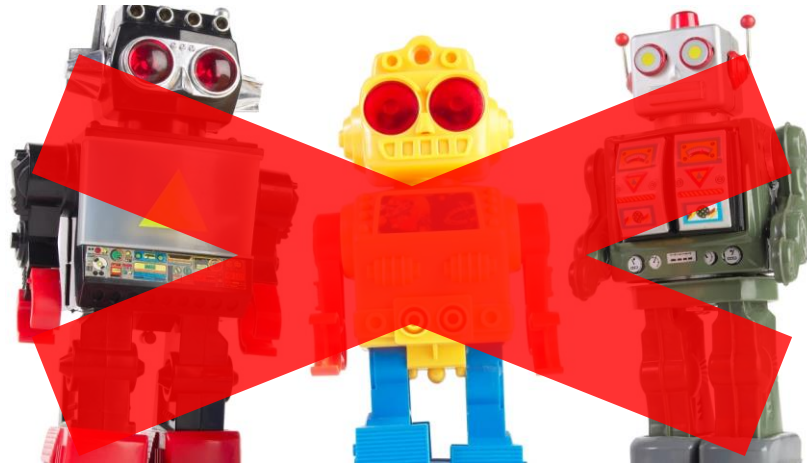
UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

EU HLEG	OECD	IEEE EAD
<ul style="list-style-type: none"> • Human agency and oversight • Technical robustness and safety • Privacy and data governance • Transparency • Diversity, non-discrimination and fairness • Societal and environmental well-being • Accountability 	<ul style="list-style-type: none"> • benefit people and the planet • respects the rule of law, human rights, democratic values and diversity, include appropriate safeguards (e.g. human intervention) to ensure a fair and just society. • transparency and responsible disclosure • robust, secure and safe • Hold organisations and individuals accountable for proper functioning of AI 	<ul style="list-style-type: none"> • How can we ensure that A/IS do not infringe human rights? • Traditional metrics of prosperity do not take into account the full effect of A/IS technologies on human well-being. • How can we assure that designers, manufacturers, owners and operators of A/IS are responsible and accountable? • How can we ensure that A/IS are transparent? • How can we extend the benefits and minimize the risks of AI/AS technology being misused?



THEY DON'T ARE NOT ADDRESSING THESE:



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

BUT ARE THEY *ACTIONABLE?*



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

**WE CHECKED THAT WITH THE
INDUSTRY.**



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING



WP5 Promoting European ethical, legal, cultural and socio-economic values for AI

Dr. Andreas Theodorou
Umeå University

Deep Dive Interviews Results



Ethics Guidelines for AI – Requirements



Human agency and oversight



Diversity, non-discrimination and fairness



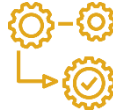
Technical Robustness and safety



Societal & environmental well-being



Privacy and data governance



Accountability



Transparency

To be continuously implemented & evaluated throughout AI system's life cycle

Ethics Guidelines for AI – Assessment List



Assessment list to operationalise the seven key requirements

- Practical questions for each requirement – 131 in total
- Test through piloting process to collect feedback from all stakeholders (public & private sector)
- The interviews are part of this feedback process (“qualitative analysis”)

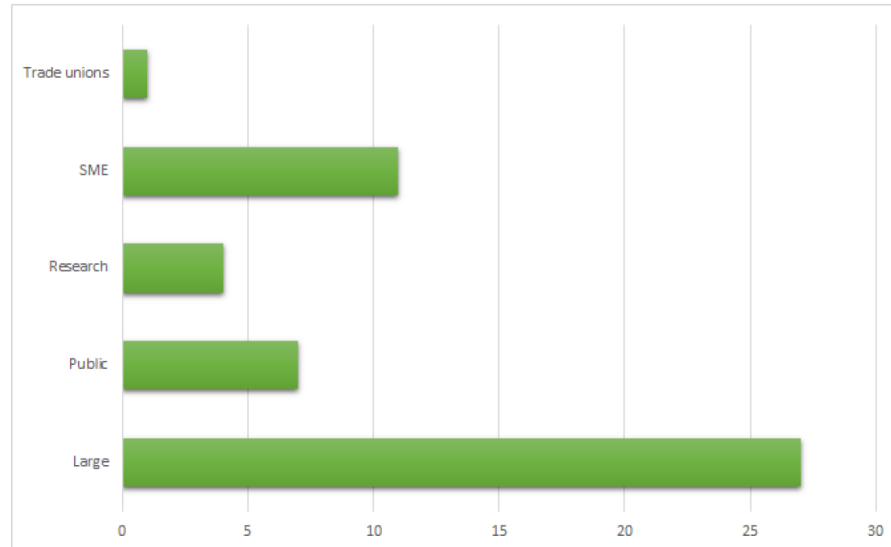
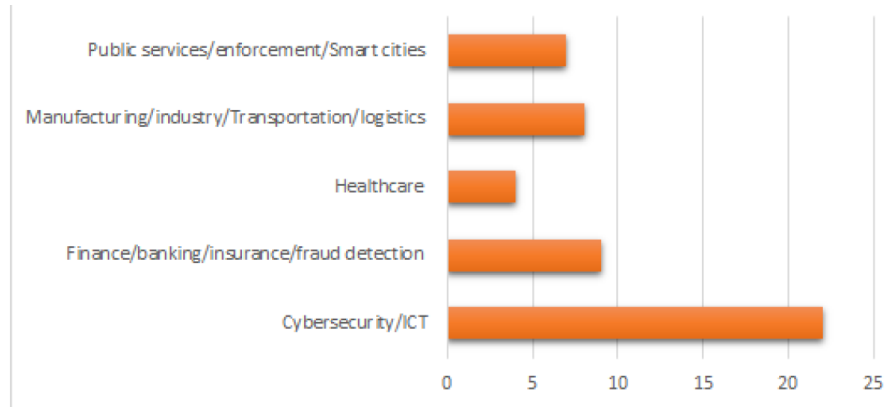
Ethics Guidelines for AI – Assessment List



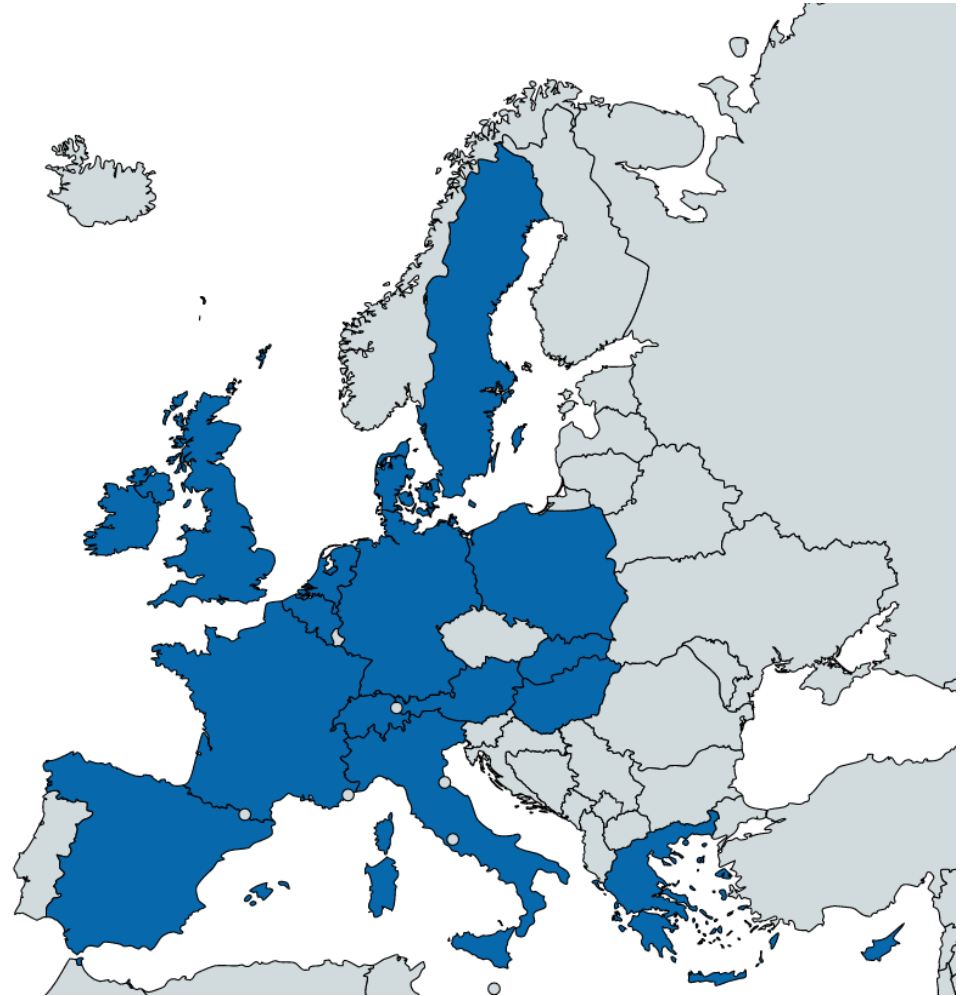
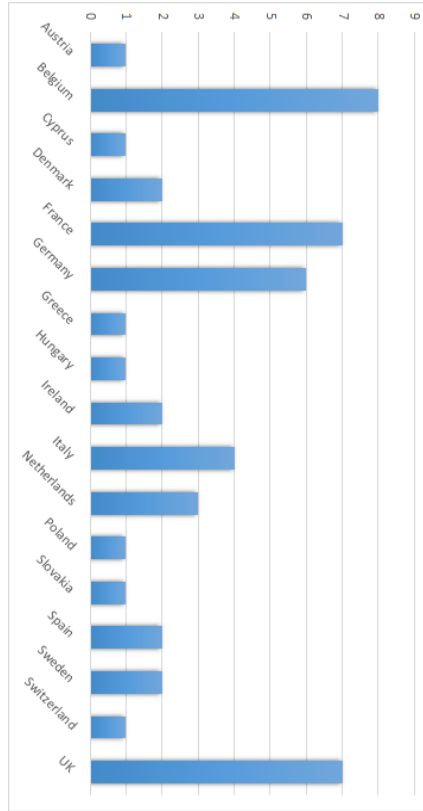
- The Ethics Guidelines for Trustworthy AI provide an assessment list that operationalises the seven key requirements and offers guidance to implement them in practice.
- In order to test the assessment list and provide practical feedback on how it can be improved, in-depth interviews (“deep dives”) with a number of representative organisations are conducted, gathering detailed qualitative feedback.
- How can the Trustworthy AI assessment list be implemented and operationalised in each organization? What is missing? What is unclear?
 - E.g. incorporating the assessment list into existing governance mechanisms; or,
 - implementing new processes.

Methology

- Selection:
- AI Alliance registrations indicating interest in participation
- Max. 50 organisations
- Cross-sector
- Cross organization type
- European spread
- Invitations July 2019 followed by reminder and extra selection to make up numbers

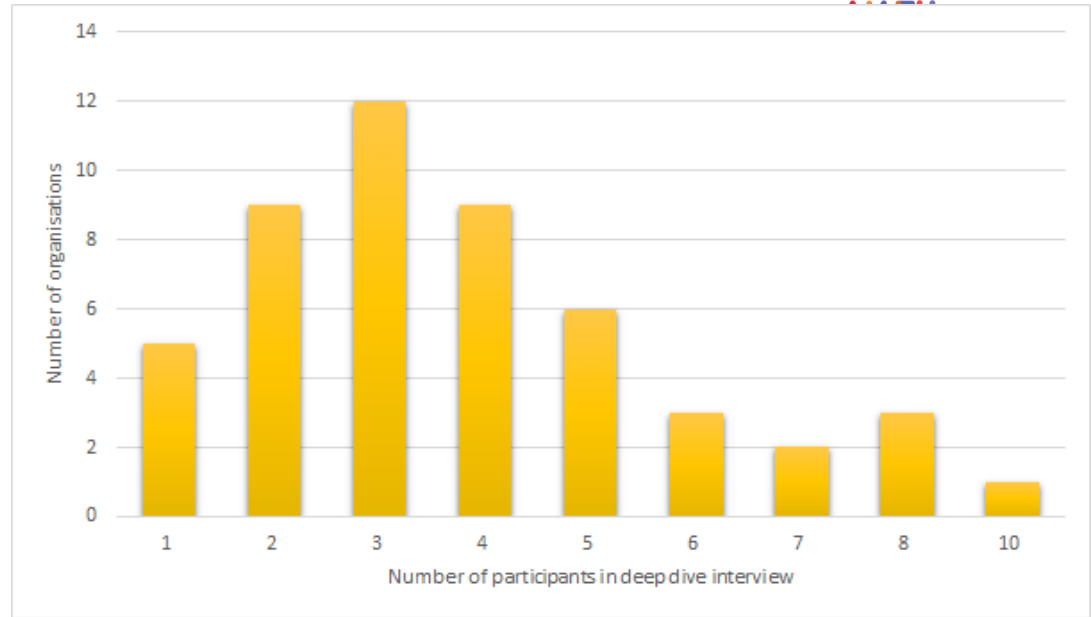


Participants



Interviews

- End September to begin November 2019
- Interviewers Team:
 1. Giulia Carra, Wavestone
 2. Atia Cortés, BSC
 3. Begum Genc, UCC
 4. Gabriel Gonzalez-Castañé, UCC
 5. Cédric Goubard, Wavestone
 6. Juan Carlos Nieves, Umeå University
 7. Yann Alan Pilatte, Sorbonne University
 8. Teresa Scantamburlo, University of Venice
 9. Marie Schacht, EC contractor
 - 10. Andreas Theodorou, Umeå University**
 11. Risto Uuk, ECcontractor
 12. Andrea Visentin, UCC
 13. Florian Zimmermann, Fraunhofer
- Coordination:
 - Virginia Dignum
 - Raja Chatila



- Average 3,6 participants per interview
- Roles: CEO, CTO, DPO, data analyst, business developer, head AI/Research, legal expert, (senior) project/product manager, strategy expert, ...

“ Evaluation - Feasibility

- **Effort:** The list is too long.
- **Clarity:** Too elaborate; very generic/abstract terms; not proactive (suggestions).
- **Focus:** Many redundancies; different levels of abstractions
- **Target:** Full list is not applicable / useful to most.
- **Examples/Use cases:** examples on AI projects, industrial cases, scenarios and possible remedies would help to clarify how to interpret requirements. Borderline examples or trade-off examples should be given.
- **Lacking definitions:** not self-contained.
- **Actors:** no distinction between corporate level questions and specific AI projects.

“ Evaluation - content

- **Not AI specific:** many aspects relate to IT products in general (e.g. safety, robustness) or they are already covered in specific domains (e.g. data governance and privacy in finance / insurance / health).
- **Not sector specific:** lack of coverage for B2B
- **Overlaps/relevance:** (e.g. questions in Bias and Robustness)

“ Evaluation – Alignment with existing practice

- Large companies: harmonize the assessment list with existing policies by EU supervising agencies (e.g. EIPOA)
- Requirements already covered by policies, standards, or existing practices: privacy / safety / robustness / data governance, etc.
 - Frameworks for project management (e.g. Agile methodology)
 - Platforms for software documentation
 - GDPR and other tools like the data protection impact assessment
 - ISO standards
 - Regular security tests (Confidentiality / Integrity / Availability).
- Some organizations already have internal guidelines for trustworthy AI, which are already well aligned to the guidelines of the HLEG. Others are in the process of developing them or are part of panels / round tables (at national level) on AI and society.

“ Recommendations – Impact of assessment

- **Competitiveness:** Too complex assessment can slow down business.
- **Innovation:** transparency and explainability requirements, may require disclose the innovative aspects of their product.
- **Overload:** require to produce much more documentation.
- **Market impact:** actionable guidelines must focus on different markets and specific to country regulations.
- **Awareness:** answering questions was a good exercise to raise awareness and trigger reflection on responsible AI
- **Education:** guidelines can be used as a learning instrument
- **Design principles:** use guidelines to inform the AI governance frameworks and design principles

“ Recommendations – missing issues

- Means to distinguish maturity levels in application
- Means to enforce
- Specification of what is are obligations for organisations, under which circumstances

“ Suggestions on structure

- **Intention:** clarifying the intention and expectations
- **Risk-based:** organizing questions based on risk impact and relevance
- **Principle-based:** providing examples of problems and possible solutions. The assessment should be easy to be remembered
- **Hierarchical/Layered presentation** to differentiate more general questions from those that are more specific. E.g. define pathways, interactive environment.
- **Abstraction level:** separate corporate-level and more specific project-level questions.
- **Interlinks** among sections that are strongly connected (e.g. fairness and accountability)
- **Actors:** specify who (stakeholder role) should answer the question (e.g. trade union, legal department, DPO, data scientist, etc.)

“ Needs

- **SMEs:**
 - Tools are required to assess their solutions.
 - Resources are needed to improve the time required to readapt their business.
 - Guidance is required for assessing them.
- **Large Companies:**
 - Trained people are required for performing the assessment.

“ Needs

- **Culture / education:**
 - Narrative on AI assessment suffers from a problem of miscommunication.
 - Every organisation interpreted the Principles differently.
- **Incentives:**
 - fines / certification could be a good incentive for the application of the assessment list.
 - But certification could arise lobby issues

“ Needs: Tools / services (External or EC)

- 'EU approved' external partners for auditing processes with respect to ethics and accountability
- Certification/recommendations or references and tools
- building up Trustworthy AI competencies and providing mid-career training
- re-skilling people
- establishing a network to engage in on a practical level, share best practices, exchange about development tools, etc.

“ Conclusions

- Trustworthy AI guidelines and Assessment List
 - **Overall view is positive**, and they are well accepted.
 - comprehensive, useful, and can be used by to analyse their current AI model situation with respect to what is recommended.
 - **But lack of context and unclear purpose and focus.**
 - All requirements are relevant but not in the same degree; some already covered by other processes
 - Concerns on the level of regulatory enforcement:
 - Preference for soft regulation (as it may impact competitiveness)
 - If hard regulation then EU agency is needed



WP5 Promoting European ethical, legal, cultural and socio-economic values for AI



HOW CAN WE AGREE ON ESLEC INTERPRETATIONS WHEN WE CAN'T AGREE ON TECHNICAL TERMS?



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

ARTIFICIAL INTELLIGENCE IS...

- **A (computational) technology that is able to infer patterns and possibly draw conclusions from data** (currently AI technologies are often based on machine learning and/or neural networking based paradigms)
- **A field of scientific research** (this is the original reference and still predominant in academia); the field of AI includes the study of theories and methods for adaptability, interaction and autonomy of machines (virtual or embedded)
- **An (autonomous) entity** (e.g. when one refers to ‘an’ AI); this is the most usual reference in media and science fiction, but is however the most incorrect one. Brings with it the (dystopic) view of magic powers and a desire to conquer the world.

Theodorou, A. and Dignum V. (2020), *Towards Ethical Socio-Legal Governance in AI*. Nature Machine Intelligence.



“AI IS WHATEVER HASN'T BEEN DONE YET.”

Douglas Hofstadter; *Gödel, Escher, Bach: An
Eternal Golden Braid*



UMEÅ UNIVERSITY

LACK OF DEFINITIONS LEADS TO...

- A constant **re-writing of similar high-level policy statements.**
- **Creates loopholes to be exploited.**
- **Increases public's misconceptions; “true AI”, “superintelligence”.**

Theodorou, A. and Dignum V. (2020), Towards Ethical Socio-Legal Governance in AI. Nature Machine Intelligence.



UMEÅ UNIVERSITY

MOVING AWAY FROM HIGH-LEVEL GUIDELINES

Making Concrete Definitions



UMEÅ UNIVERSITY

IT IS A LONG & HARD PROCESS

WHEN WE SAY THAT WE PUT
ETHICS BEFORE PROFITMAKING,
IT MEANS THAT WE CAN
CONTINUE MAKING MORE PROFITS!



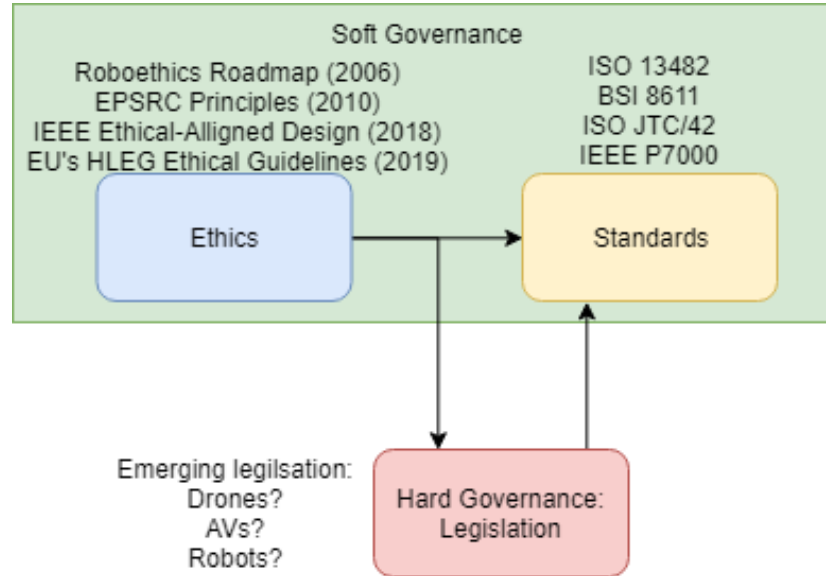
69. Moral functionalism (also instrumentalism)

The view that ethics should merely be a useful instrument for other purposes. A risk is that ethics is not seen as a value in and of itself.



UMEÅ UNIVERSITY

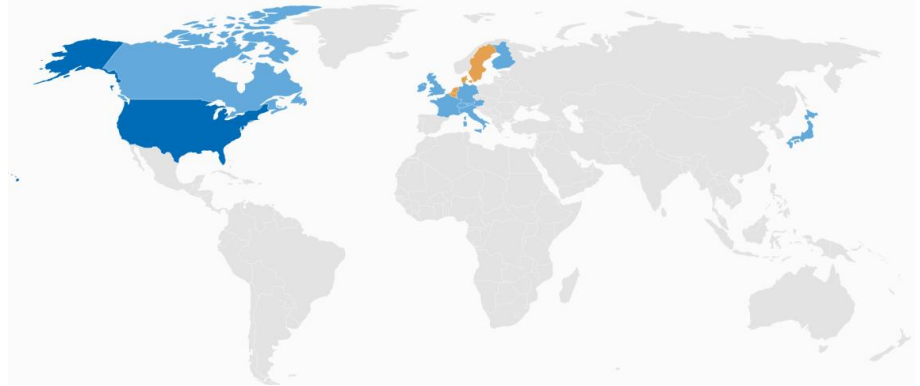
SOCIO-LEGAL SOLUTIONS





STANDARDS

- ISO/IEC/JTC 1/SC 42
- ISO/IEC JTC 1 N13468 46T
Artificial Intelligence
Concepts and Terminology.
- ISO/IEC JTC 1 N 13502
Framework for Artificial
Intelligence (AI) Systems
Using Machine Learning
(ML).





IEEE STANDARDS

- **IEEE P7000™** - Model Process for Addressing Ethical Concerns During System Design
- **IEEE P7001™** - Transparency of Autonomous Systems
- **IEEE P7002™** - Data Privacy Process
- **IEEE P7003™** - Algorithmic Bias Considerations
- **IEEE P7004™** - Standard on Child and Student Data Governance
- **IEEE P7005™** - Standard on Employer Data Governance
- **IEEE P7006™** - Standard on Personal Data AI Agent Working Group
- **IEEE P7007™** - Ontological Standard for Ethically driven Robotics and Automation Systems
- **IEEE P7008™** - Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems
- **IEEE P7009™** - Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems
- **IEEE P7010™** - Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems
- **IEEE P7011™** - Standard for the Process of Identifying and Rating the Trustworthiness of News Sources
- **IEEE P7012™** - Standard for Machine Readable Personal Privacy Terms



**DOES THIS MEAN THAT
EXISTING STANDARDS ARE
NO LONGER APPLICABLE?**

NO.



UMEÅ UNIVERSITY

HOW CAN WE ENFORCE THE ADOPTION OF STANDARDS?

**By giving them “teeth” through
legislation.**



UMEÅ UNIVERSITY

LEGISLATION

- Will enforce the adoption of standards.
- **Only some tuning of existing regulations is necessary.**
- Aims is to ensure right attribution of legal accountability.

Bryson J.J., Theodorou A. (2019). *How Society Can Maintain Human-Centric Artificial Intelligence*. Toivonen-Noroand M and Saari E eds. *Human-Centered Digitalization and Services*. Springer, Berlin.



CONSISTENCY!

Create a concrete ethics policy. It should include any necessary definitions and leave little to the imagination.



UMEÅ UNIVERSITY

CONTEXT MATTERS

Stakeholders, projects, societies that will be deployed to, etc should be taken into consideration through the process.



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

ETHICS

Not a binary compliance checking...



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

ETHICS IS NOT AN AFTERTHOUGHT

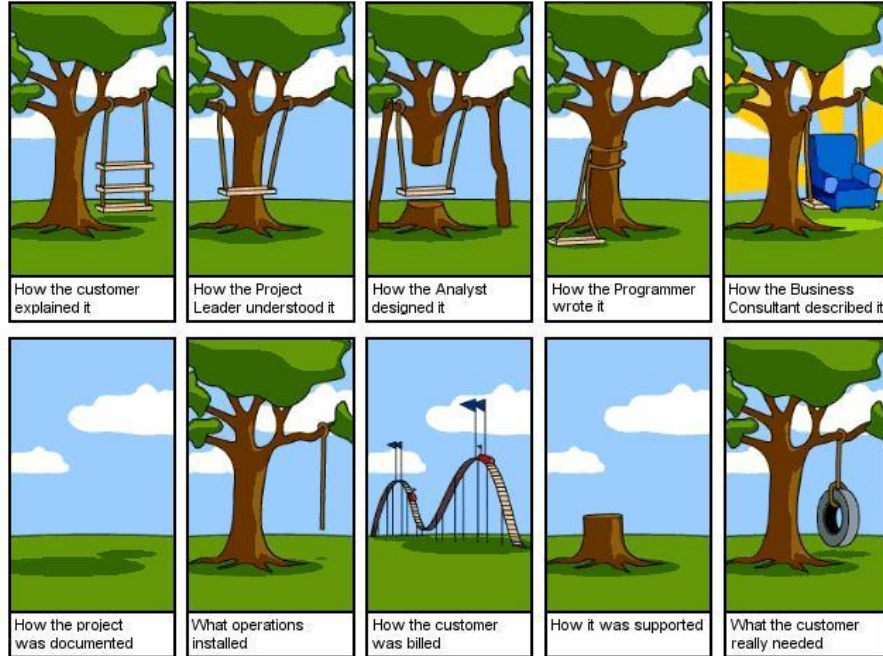
“As you set out for Ithaka
hope your road is a long one,
full of adventure, full of discovery.”



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

TECHNICAL SOLUTIONS



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

REQUIREMENTS

Think about them. Define them; check their compliance against your policy and law.



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

DESIGN

Your policy should inform (and influence) them.



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

DESIGN

**You may need to make hard decisions;
e.g. performance vs explainability,
reducing utility to avoid deception, etc.**



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

DESIGN

But also remember, the newest/fancier model is not necessary the best.



UMEÅ UNIVERSITY

DEVELOPMENT

**DON'T HACK CODE TOGETHER. THINK
IN TERMS OF YOUR ARCHITECTURE.**



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

DEVELOPMENT

**VERSION CONTROL: SAVES THE MENTAL HEALTH
OF YOUR TEAM AND PROVIDES TRACEABILITY.**

ISO 9001..



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

TESTING

**CHECK ROBUSTNESS
REPRODUCIBILITY.**



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

TESTING

**KEEP IN MIND: YOUR SIMULATOR WILL
ONLY TEST WHAT YOU THOUGHT OF.**



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

TESTING

CYBER SECURITY GOES WITHOUT SAYING.



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

DEPLOYMENT

TALK TO YOUR STAKEHOLDERS.



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

DEPLOYMENT

Consider the effects of the system to the society; do you need to train people? Will people lose their jobs? What about the environment? Behaviour change?



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

DON'T BE AFRAID
GOING BACK TO THE DRAWING BOARD, IF
NEEDED, IS NOT A BAD IDEA.



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

DOCUMENT EVERYTHING!

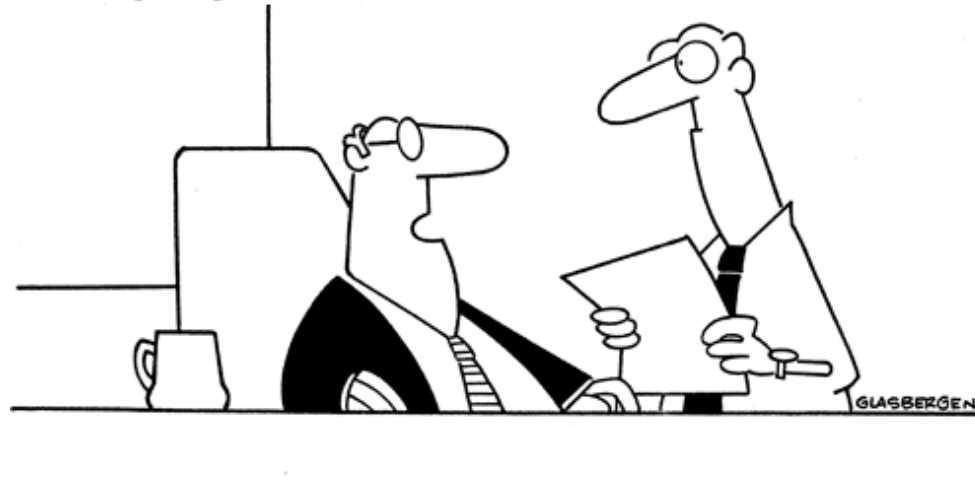
**Ensures transparency in the process.
Proves due diligence – helps with
responsibility and accountability.**



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

© Randy Glasbergen.
www.glasbergen.com



**“We’ve got to draw the line on unethical behavior.
But draw it in pencil.”**



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

OH! AND AVOID OVERSTATEMENTS.

You can't have an "unbiased" data-driven system. It simply wouldn't work.



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING

TRADE-OFF

**ALL OF THIS IS A MULTI-OBJECTIVE
OPTIMISATION PROBLEM.**



UMEÅ UNIVERSITY

ANDREAS THEODOROU | T: @RECKLESSCODING



Prof. Virginia Dignum
Professor of Social and Ethical AI
 @vdignum
 virginia@cs.umu.se



Dr. Juan Carlos Nieves
 Associate Professor
 juan.carlos.nieves@umu.se

Read our group's research:



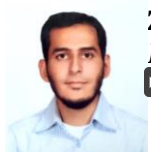

Dr. Andrea Aler Tubella
Postdoctoral Researcher
 andrea.aler@umu.se



Dr. Andreas Theodorou
Postdoctoral Researcher
 @recklesscoding
 andreas.theodorou@umu.se



Julian Mendez
PhD Student
 julian.mendez@umu.se



Zahoor Ul Islam
PhD Student
 zahoor.ul.islam@umu.se



Prof. Frank Dignum
Professor Social Aware AI
 Umeå University
 frank.dignum@umu.se



Prof. Joanna J Bryson
Professor of Technology & Ethics
 Hertie School of Governance
 @j2bryson



Dr. Teresa Scantamburlo
Postdoc – Univ. Of Venice
 AI4EU – Deep Dives



Dr. Àtia Scantamburlo
Postdoc – Univ. Of Venice
 AI4EU – Deep Dives
 @atcortesm

Dr. Loizos Michael – Open Uni. Cyprus
 Prof. Antonis Kakas – Uni. Of Cyprus
 Dr. Robert Wortham – Uni. Of Bath



Q & A

Andreas Theodorou

 **andreas.theodorou@umu.se**

 **@recklesscoding**



UMEÅ UNIVERSITY