Measuring Trustworthiness in Neuro-Symbolic Integration

Andrea Omicini   Andrea Agiollo

Dipartimento di Informatica – Scienza e Ingegneria (DISI)
Alma Mater Studiorum—Università di Bologna
andrea.omicini@unibo.it, andrea.agiollo@unibo.it

Keynote @ FedCSIS 2023
Warsaw, Poland
19 September 2023

Next in Line. . .

Natural vs. Artificial Systems

- the match between natural systems and artificial ones is increasingly getting more and more articulated, even intricate
  - on the one hand, we understand more and more the computational aspects of natural systems—e.g., biological ones
  - on the other hand, we keep getting inspiration from natural systems for our computational models—e.g., nature-inspired computing (NIC)
- *multi-*, *inter-*, *trans-disciplinary* studies are nowadays increasingly common among computer scientists and engineers
  - even though most of them cannot tell the difference among the three sorts

Neuro-symbolic Integration Systems as Nature-Inspired

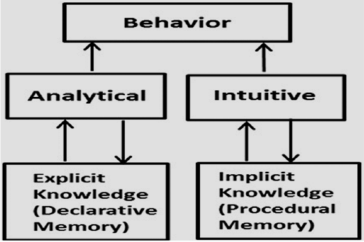
- neuro-symbolic integration systems (NeSy) integrate neural (subsymbolic) and symbolic AI approaches
  - blending the subsymbolic perspective of ML and DL agents with symbolic AI solutions focusing on high-level symbolic (human-readable) representations of problems, logic, and search
- given that
  - neurons in our brain clearly provide inspiration for neural components
  - and inspiration of symbolic techniques can be traced back at least to Aristotle's logic[De Rijk, 2002]—studying how humans reason, understand the world, and plan their course of action

⇒ NeSy are easy to deem as *nature-inspired systems*

Humans as NeSy I

Rationality vs. intuition

- two sorts of cognitive processes
  - esprit de finesse *vs.* esprit de géométrie—rationality has limits[Pascal, 1669]
  - *cognitivism* against *behaviourism* in psychology[Skinner, 1985]



- concepts and distinctions *not* born in the CS / AI fields
- yet, they roughly match the two main families of AI techniques
  - symbolic vs. sub-/non-symbolic

⇒ *humans as NeSy*

Humans Share Knowledge

- it is not brain size (or whatever like that) that separates humans from other intelligent animals like primates
  - instead, it is mostly our will to *share knowledge*[Dean et al., 2012]
- in general, knowledge sharing is a peculiar trait of humanity
  - it is how we do understand each other
  - it is how we learn
  - it is the foundation of human society
  - where human culture is a *cumulative* one

e.g. human science is a shared *social construct*
  - scientific artefacts are required to be *understandable* for the community
  - so as to enable *reproducibility* and *refutability* in the scientific process[Popper, 2002]

Human Systems as NeSy

We never think alone

- we are *hyper-social animals*: "We never think alone" [Sloman and Fernbach, 2018]
- reasoning evolved *after* our ability to interact socially
- along with *language*, as a *symbolic artefact*[Nardi, 1996, Clark, 1996]

We never *read* alone

- as we *share* knowledge through representational artefact
  - books, the Web, ...
- and work within shared *knowledge-intensive environments*
  - where both knowledge and cognition processes are *distributed* among humans and artefacts[Kirsh, 1999]

Interaction in Intelligent Systems

- *symbolic* approaches are particularly relevant within intelligent systems
- in the *shared representation* of interaction between intelligent components
  - e.g., explanation as a rational act for human and explaining agents[Omicini, 2020]
- for instance, symbolic approaches are critical when dealing with systems features such as
  - explainability
  - understandability
  - accountability
  - trustworthiness
- so, when focussing on NeSy, we better put some extra care on the *interaction* aspect of symbolic/subsymbolic integration

Next in Line. . .

Human Rights & AI Systems

- socio-political pressure for human rights guaranteed by artificial systems
  - e.g., EU via GDPR[Voigt and von dem Bussche, 2017] recognises "the citizens' right to explanation"[Goodman and Flaxman, 2017]
- as obvious, this mostly stem from the foreseeable impact of AI systems on current / future European citizen's life

Trustworthiness of AI Systems I

- in its AI strategy, European Commission defines the *guidelines* to promote trustworthy AI[High-Level Expert Group on Artificial Intelligence, 2019]
- AI should be *lawful, respectful, robust*
- following *7 key requirements* that AI systems should meet in order to be deemed *trustworthy*

Trustworthiness of AI Systems II

Problem

- how do we know we did it?
- how can *intelligent systems engineers* ensure that their systems actually comply with the trustworthiness requirements?
  - is it just a matter of following the guidelines?

? are the guidelines precise / detailed / complete enough to actually drive the whole engineering process, leading to the desired outcome?

! of course they are not—they are not meant to be

- yet, this is not the (whole) point here
- so, we have a lot to discuss here

UN 2030 Agenda for Sustainable Development

- on September 2015, the UN General Assembly adopted the 2030 Agenda for Sustainable Development, addressing 17 Sustainable Development Goals (SDGs) by "an urgent call for action by all countries" https://sdgs.un.org/goals

? is it working?

At the global level, . . . not a single SDG is currently projected to be met by 2030 [Sachs et al., 2023]

! definitely not.

? and the problem is?

The Problem of Measuring Things

- goals come without a comprehensive approach allowing for *quantitative* evaluation
- when only qualitative definitions of goals are provided, the *assessment* (of the levels) of achievement – the measure of success – is simply not possible
- goals, guidelines, targets, features—without suitable *quantitative* evaluation frameworks and *measuring tools*, they are likely to be *ineffective*

Measure as Symbolic?

- as scientists, we may tend towards a notion of measure that is mostly a symbolic one
- yet, this is not strictly necessary
- e.g., our brain keeps measuring time at any scale using a wide range of different neural circuits
- so – *disclaimer* – that is not the point here

Next in Line. . .

Key Questions Here

- how can we ensure that our NeSy will match EU requirements for trustworthy AI?
- are the guidelines defined by EU enough for that?
- is the general notions of AI and intelligent system (implicitly) adopted there enough when NeSy are concerned?
- do they have enough focus on the critical issues of NeSy?
- are we equipped with the ability of measuring the compliance of any specific NeSy to the key criteria for trustworthiness?

Our Motivation Here

- definition of trustworthiness requirements are not enough without metrics
- to some extent, EU trustworthiness requirements seems to focus more on data-driven solutions
  - at their best, on rule-based systems
- in any case, NeSy have *specific* features and issues
  - so that NeSy require more detailed notions of trustworthiness and related metrics to be deemed as trustworthy,

Contribution

- we start discussing how AI trustworthiness requirements should translate when applied to NeSy realm
- we analyse some available metrics for each novel NeSy trustworthiness requirement
- we suggest novel metrics to measure specific NeSy elements
- in particular, we focus on some specific NeSy sorts, based on
  - *symbolic knowledge injection* (SKI)
  - *symbolic knowledge extraction* (SKE)

Next in Line...

EU Definition of AI Trustworthiness

In its AI strategy[High-Level Expert Group on Artificial Intelligence, 2019] EU defines *7 key criteria for trustworthiness*

1. *human agency and oversight* → control over AI's actions
2. *robustness and safety* → reliable/predictable actions
3. *privacy and data governance* → data access, quality, integrity
4. *transparency* → what is AI doing/thinking?
5. *diversity, non-discrimination, and fairness* → non-biased actions
6. *environmental and societal well-being* → focus on future generations
7. *accountability* → actions responsibility

Human Agency and Oversight I

AI version

Need for oversight mechanisms enabling the informed interaction between AI agent(s) and human(s)

$\Downarrow$

Questions to answer

- which are interaction mechanism exists? and, which are the key ones?
- how much can human user interact with or affect AI?
- what is overall extent of the interaction mechanisms?
- ...

Human Agency and Oversight II

NeSy perspective

Symbolic components can play a key role in the interaction between human and system

- human-in-the-loop (HITL), human-on-the-loop (HOTL), human-in-command (HIC), ...

and some of them are typically already in place when NeSy are concerned

$$\Downarrow$$

NeSy version

Need for assessing how much *symbolic components* already in place as well as symbolic/subsymbolic *interaction improve* informed human-AI interaction and oversight

Human Agency and Oversight III

New questions to answer

- are there NeSy components impacting human oversight and control?
- is oversight extension and quality improved or worsened by them?
- ...

Robustness and Safety I

AI version

Need for accuracy, reliability, predictability, resilience, and security of AI

$\Downarrow$

Questions to answer

- is the system robust against perturbation?
- how does AI behave for out-of-distribution samples?
- are system prediction reliable and safe?
- is the agent secure against malevolent usage?
- ...

Robustness and Safety II

NeSy perspective

Symbolic components generally verifiable and stable, yet subsymbolic ones (still) lack strong mathematical modelling of their behaviour

- symbolic components could help harnessing subsymbolic elements
- subsymbolic components produce imperfect and not-so-reliable symbolic knowledge

$\Downarrow$

NeSy version

Need to assess the impact of symbolic (verifiable) and subsymbolic (not verifiable) *interaction* on system *stability*

Robustness and Safety III

New questions to answer

- does NeSy improve system stability at all?
- symbolic verifiable elements correctly integrated in NeSy?
- what happens if the symbolic component is somehow altered/corrupted?
- is the system stable over symbolic representation variation?
- ...

Privacy and Data Governance I

AI version

Need for ensuring legitimate access to data, taking into account data quality and integrity

⇓

Questions to answer

- should used data be publicly available or private?
- who can access the data? Model leaking data information?
- is data collection process reliable?
- are there any missing information or misleading/bugged data?
- ...

Privacy and Data Governance II

NeSy perspective

- symbolic component relies on knowledge-bases, ontologies, etc.
- one should ensure such components are qualitatively sound
- possible issues in knowledge-bases impact NeSy performance negatively and are difficult to spot during integration

$\Downarrow$

NeSy version

Need for ensuring the *quality* of both *data* and *symbolic knowledge* of a NeSy system, along with its proper accessibility

Privacy and Data Governance III

New questions to answer

- compatibility/overlap between data and symbolic knowledge?
- bugs or conflicting information in the symbolic knowledge used?
- is the human-centred building process of symbolic knowledge impacting on its quality/reliability?
- does NeSy leak information about its symbolic component?
- ...

Transparency I

AI version

Need for providing human users with explanations of the AI decision process

$\Downarrow$

Questions to answer

- are explanations for the AI decision process available in some form?
- how much are explanations understandable?
- what is the level of fidelity between AI and its explanations?
- ...

Transparency II

NeSy perspective

- symbolic component makes most NeSy systems more transparent solutions by design.
- complexity of explanation extraction process is reduced, explanations understandability is increased due to symbolic component somehow understandable by humans

$$\Downarrow$$

NeSy version

Need for assessing the *gain* in terms of *transparency* obtained by a NeSy system with respect to its pure subsymbolic components/counterparts

Transparency III

New questions to answer

- what is the quality of system's explanations before and after symbolic and subsymbolic integration?
- is the gain measurable, and how?
- how does explanation change with NeSy?
- are automatically-measurable quantities enough for explanations?
- ...

Fairness I

AI version

Need for avoiding unfair bias while enabling everyone's access to AI

$\Downarrow$

Questions to answer

- is the outcome for the AI's decision making process *equal* for everyone?
- what are the groups affected by bias in predictions?
- what are the features affecting agent's bias?
- ...

Fairness II

NeSy perspective

- biases of subsymbolic models and NeSy counterparts differ in their root causes
- bias can rise in NeSy as consequence of
  - any unexpected behaviour of their subsymbolic components
  - interaction of their symbolic and subsymbolic elements
- bias/fairness of symbolic components is verifiable and provable, its interaction with subsymbolic is not

$$\Downarrow$$

NeSy version

Need for *measuring* biased/discriminative behaviour of NeSy rooted in *interaction* between symbolic and subsymbolic components

Fairness III

New questions to answer

- does NeSy integration increases or decreases bias?
- is bias increment/decrement due to symbolic/subsymbolic component or their interaction?
- is it possible to measure only the impact of symbolic and subsymbolic integration upon fairness?
- ...

Resource Efficiency I

AI version

Need for sustainability of AI and transition to their environmentally-friendly development

$\Downarrow$

Questions to answer

- how much energy is required by the system to be optimised?
- is the AI system scalable?
- what is the amount of data – along with collection complexity – required by the AI?
- ...

Resource Efficiency II

NeSy perspective

- in optimal NeSy interaction, symbolic component lifts part of learning burden from subsymbolic elements, reducing resources required for optimisation
- overcomplicated NeSy interaction incurs in resource waste given by translation/interface overhead

$$\Downarrow$$

NeSy version

Need for assessing the *gain/loss* in terms of *sustainability* of NeSy systems with respect to their pure subsymbolic components

Resource Efficiency III

New questions to answer

- how much energy/time/memory can NeSy integration save/waste w.r.t. pure subsymbolic AI agents?
- is the complex interaction between symbolic and subsymbolic components introducing resource overhead?
- can NeSy systems learn using less data?

Accountability I

AI version

Need for ensuring responsibility and accountability for behaviour and outcomes of AI systems

⇓

Questions to answer

- is it possible to justify AI behaviour?
- is the AI informative enough for human users?
- who is to blame when an AI system fails?
- ...

Accountability II

NeSy perspective

- accountability tightly linked with transparency
- symbolic component makes most NeSy more accountable by design
- explanations simpler to extract with increased understandability—transparency effect of symbolic component in NeSy

$\Downarrow$

NeSy version

Need for assessing *gain* in *answerability* obtained by NeSy with respect to subsymbolic components/counterparts

Accountability III

New questions to answer

- are general AI accountability criteria straightforwardly applicable to NeSy?
- how does NeSy interaction affect extracted explanations?
- are obtained explanations robust against input variability?
- ...

Next in Line. . .

On the Need for Metrics

- most requirements easy to understand conceptually
- yet, requirements are not binary
  - broad spectrum of transparency level $\rightarrow$ grey-box models
  - weak vs. strong bias
  - oversight on whole model or single "tweakable" component
  - ...
- whether or how much a requirement is satisfied is difficult to determine

Trustworthiness metrics

Definition of trustworthiness metrics rather than requirements makes it possible to

- measure system properties
- analyse grey areas
- define satisfiability thresholds
- simply compare different solutions

Background: SKI and SKE

Symbolic Knowledge Injection (SKI)

NeSy characterised by explicit procedures affecting how subsymbolic components draw inference for them to be (made) consistent with symbolic knowledge

Symbolic Knowledge Extraction (SKE)

NeSy accepting *subsymbolic predictors* as input and producing symbolic knowledge as output, distilling knowledge that a subsymbolic predictor grasped from data into symbolic form

Human Agency and Oversight

Available metrics

- measuring how explanations guide people to respond/predict AI behaviour [de Graaf and Malle, 2017]
- subjectively rating system predictability, likability, and the like, based on user feedback[Huang and Mutlu, 2012]
- *focus on general AI, assuming they can transfer to NeSy*

Missing metrics

- assessment of human influence/control on AI system
- amount of injected knowledge effectively absorbed by NeSy (SKI) model
- portion of symbolic knowledge extracted in SKE
- amount of knowledge extracted, refined and injected back in NeSy (SKE+SKI) being correctly assimilated

Robustness and Safety

Available metrics

- performance in out-of-distribution [Li et al., 2022, Liu et al., 2023]
- prediction coherence and consistency [Nye et al., 2021]
- subsymbolic verification via NeSy [Xie et al., 2022]
- robustness input perturbations [Yang and Chaudhuri, 2022]
- robustness against adversarial attacks [Vilamala et al., 2023]
- *qualitative vs. quantitative*

Missing metrics

- assessment of preservation of stability and verifiability of symbolic components in NeSy
- portion of symbolic elements correctly integrated in SKI
- stability of SKI when injected knowledge is altered (imperfect automation process)
- stability of NeSy over symbolic representation variability

Data & Knowledge Quality

Available metrics

- class overlap[Denil and Trappenberg, 2010]
- boundary complexity[Lorena et al., 2019]
- label noise[Northcutt et al., 2021]
- class imbalance[Lu et al., 2020]
- missing value analysis[Corrales et al., 2018]
- *data component only*

Missing metrics

- level of compatibility/overlap between data and symbolic knowledge in SKI
- measure of incomplete knowledge bases
- measure of bugged knowledge bases

Transparency

Available metrics

- explanations attributes [Hoffman et al., 2018]
- metrics for simplicity, broadness, and fidelity of explanations [Nguyen and Martínez, 2020]
- causability scale [Holzinger et al., 2020]
- unambiguity and interactivity in SKE [Lakkaraju et al., 2017]

Missing metrics

- *gain* in transparency: before vs. after NeSy
- measure of human-oriented specifications
- measure complexity of explanations extraction process in SKE

Fairness

Available metrics

- observational vs. causal fairness [Calegari et al., 2023]
- *independence vs. separation vs. sufficiency metrics*
- fairness through SKI[Gao et al., 2022]
- fairness through SKE and continual learning[Wagner and d'Avila Garcez, 2021]

Missing metrics

- differential of observational fairness between SKI/SKE and ML/DL counterpart
- measure fairness over set of symbolic knowledge bases (representing fairness goals)
- measure fairness over set of subsymbolic components

Resource Efficiency

Available metrics

- qualitative data efficiency of NeSy
  [Mao et al., 2019, Zhang et al., 2021, Škrlj et al., 20
- SKI resource efficiency improvements
  [Agiollo et al., 2023]
  - energy
  - latency
  - time
  - data

Missing metrics

- carbon footprint measurement
- SKE measures, subsymbolic vs. symbolic emulation resource usage:
  - energy
  - latency

Accountability

Available metrics

- transparency metrics

Missing metrics

- mix of transparency metrics and robustness metrics
- explainability over set of perturbations

Metrics – Summing Up

Findings

1. several metrics already available for AI systems
2. *not so many* metrics specifically tailored to NeSy
3. several metrics available for easily-measurable requirements
    - resource efficiency
    - robustness
    - data quality
4. very few metrics available for not-so-easily-measurable requirements
    - transparency
    - human oversight
    - accountability
    - ...

Next in Line...

Overall

Summing up

- trustworthiness EU requirements are a new pillar for AI
- yet they mostly focus on data-driven approaches
  - at best, on rule-based AI
- they are not straightforwardly applicable to NeSy
- trustworthiness measurability is required

Future work

- rigorous definition of NeSy trustworthy metrics
- implementation and analysis of NeSy trustworthy metrics
- comparison between NeSy systems in the trustworthiness perspective

Next in Line...

Intelligent Socio-Technical Systems

- in the realm of intelligent systems, nowadays, humans are legitimate components in the same way as software and physical agents
- where both *human* and *software agents* accounts for activity, knowledge, intelligence, goals, learning, ...
  - as legitimate components of intelligent socio-technical systems
- so that now the general fundamental question becomes
  - ? how are we going to shape the interaction between heterogeneous intelligent components within *intelligent socio-technical systems*?

?? e.g., is (generative) NLP the answer?

Are We Focussing on the Real Problem?

- we crave trustworthiness, understandability, accountability, . . .
- we try and find them in AI, as if can we already had them before AI
- ? do actually humans trust, understand, . . . , each others?
- so, are we preserving features of human interaction that could be changed and harmed by AI, or, are we just looking for surrogates?
- when most or all of human processes are going to rely on intelligent socio-technical systems, this is not going to be an idle question to answer

Explanation?

- I worked as a professor and a researcher all of my adult life
- I am supposed to know *exactly* what an explanation is
- it turned out *I did not*.
- when I started working on XAI, I suddenly became aware of that
- and, I had to work on that—I still have ot work on that

Explanation as Representation & Transformation

- contribution from *math teaching* [D'Amore, 2005]
  - being math the most difficult subject to explain & teach
- a semiotic representation is required whenever the object of an explanation is inaccessible to perception

  noetics — *conceptual acquisition* of an object

  semiotics — acquisition of a *representation built out of signs*
- explaining a concept via different *semiotic representations*

  transformation of treatment — changing representation within the same register of semiotics

  transformation of conversion — changing register of semiotics for the representation
- *explanation* as
  - first, *generation of semiotic representation*
  - then, transformation of semiotic register
  - finally, sharing of the transformed representation

! explainers *share* their cognitive process with explainees as explanation

Measuring Humans

So, finally

- once artificial intelligent agents become effective components of processes within human organisation, our flawed understanding and imprecise definitions of the essential properties of human behaviour become a liability
- pervasiveness of AI is finally a chance to force us to precisely define what we mean when we talk about *understanding each other*, *trusting each other*, ...
- and, to measure *how much* we do that

Acknowledgment

This work was partially supported by

- PNRR – M4C2 – Investimento 1.3, Partenariato Esteso PE00000013 – "FAIR—Future Artificial Intelligence Research" – Spoke 8 "Pervasive AI", funded by the European Commission under the NextGenerationEU programme
- CHIST-ERA IV project "Expectation" – CHIST-ERA-19-XAI-005 –, co-funded by EU and the Italian MUR (Ministry for University and Research)

Measuring Trustworthiness in Neuro-Symbolic Integration

Andrea Omicini   Andrea Agiollo

Dipartimento di Informatica – Scienza e Ingegneria (DISI)
Alma Mater Studiorum—Università di Bologna
andrea.omicini@unibo.it, andrea.agiollo@unibo.it

Keynote @ FedCSIS 2023
Warsaw, Poland
19 September 2023

References I

[Agiollo et al., 2023] Agiollo, A., Rafanelli, A., Magnini, M., Ciatto, G., and Omicini, A. (2023).
Symbolic knowledge injection meets intelligent agents: QoS metrics and experiments.
Autonomous Agents and Multi-Agent Systems, 37(2):27:1–27:30
https://link.springer.com/10.1007/s10458-023-09609-6.

[Calegari et al., 2023] Calegari, R., Castañé, G. G., Milano, M., and O'Sullivan, B. (2023).
Assessing and enforcing fairness in the AI lifecycle.
In *32nd International Joint Conference on Artificial Intelligence (IJCAI 2023)*, Macau, China. IJCAI
.

[Clark, 1996] Clark, H. H. (1996).
Using Language.
Cambridge University Press, Cambridge, UK
DOI:10.1017/CBO9780511620539.

[Corrales et al., 2018] Corrales, D. C., Corrales, J. C., and Ledezma, A. (2018).
How to address the data quality issues in regression models: A guided process for data cleaning.
Symmetry, 10(4):99
https://www.mdpi.com/2073-8994/10/4/99.

[D'Amore, 2005] D'Amore, B. (2005).
Noetica e semiotica nell'apprendimento della matematica.
In Laura, A. R., Eleonora, F., Antonella, M., and Rosa, P., editors, *Insegnare la matematica nella scuola di tutti e di ciascuno*, Milano, Italy. Ghisetti & Corvi Editore
http://www.dm.unibo.it/rsddm/it/articoli/damore/676noeticaesemioticaBari.pdf.

References II

[de Graaf and Malle, 2017] de Graaf, M. M. A. and Malle, B. F. (2017).
How people explain action (and autonomous intelligent systems should too).
In *2017 AAAI Fall Symposia, Arlington, Virginia, USA, November 9-11, 2017*, pages 19–26. AAAI Press
https://aaai.org/ocs/index.php/FSS/FSS17/paper/view/16009.

[De Rijk, 2002] De Rijk, L. M. (2002).
Aristotle: Semantics and Ontology. Volume I: General Introduction. The Works on Logic, volume 91 of *Philosophia Antiqua*.
Brill Academic Publishers
https://brill.com/view/title/7491.

[Dean et al., 2012] Dean, L. G., Kendal, R. L., Schapiro, S. J., Thierry, B., and Laland, K. N. (2012).
Identification of the social and cognitive processes underlying human cumulative culture.
Science, 335(6072):1114–1118
DOI:10.1126/science.1213969.

[Denil and Trappenberg, 2010] Denil, M. and Trappenberg, T. P. (2010).
Overlap versus imbalance.
In Farzindar, A. and Keselj, V., editors, *Advances in Artificial Intelligence*, volume 6085 of *Lecture Notes in Computer Science*, pages 220–231. Springer
https://link.springer.com/10.1007/978-3-642-13059-5_22.

[Gao et al., 2022] Gao, X., Zhai, J., Ma, S., Shen, C., Chen, Y., and Wang, Q. (2022).
FairNeuron: improving deep neural network fairness with adversary games on selective neurons.
In *44th International Conference on Software Engineering, ICSE 2022*, pages 921–933. ACM
https://dl.acm.org/doi/10.1145/3510003.3510087.

References III

[Goodman and Flaxman, 2017] Goodman, B. and Flaxman, S. (2017).
European Union regulations on algorithmic decision-making and a “right to explanation”.
AI Magazine, 38(3):50–57
DOI:10.1609/aimag.v38i3.2741.

[High-Level Expert Group on Artificial Intelligence, 2019] High-Level Expert Group on Artificial Intelligence (2019).
Ethics guidelines for trustworthy AI.
Report, European Commission
https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[Hoffman et al., 2018] Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2018).
Metrics for explainable AI: challenges and prospects.
CoRR, abs/1812.04608
http://arxiv.org/abs/1812.04608.

[Holzinger et al., 2020] Holzinger, A., Carrington, A. M., and Müller, H. (2020).
Measuring the quality of explanations: The system causability scale (SCS).
KI - Künstliche Intelligenz, 34(2):193–198
https://link.springer.com/10.1007/s13218-020-00636-z.

[Huang and Mutlu, 2012] Huang, C. and Mutlu, B. (2012).
Robot behavior toolkit: generating effective social behaviors for robots.
In Yanco, H. A., Steinfeld, A., Evers, V., and Jenkins, O. C., editors, *International Conference on Human-Robot Interaction, HRI’12, Boston, MA, USA - March 05 - 08, 2012*, pages 25–32. ACM
https://dl.acm.org/doi/10.1145/2157689.2157694.

References IV

[Kirsh, 1999] Kirsh, D. (1999).
Distributed cognition, coordination and environment design.
In Bagnara, S., editor, *3rd European Conference on Cognitive Science (ECCS'99)*, pages 1–11, Certosa di Pontignano, Siena, Italy. Istituto di Psicologia, Consiglio Nazionale delle Ricerche
.

[Lakkaraju et al., 2017] Lakkaraju, H., Kamar, E., Caruana, R., and Leskovec, J. (2017).
Interpretable & explorable approximations of black box models.
CoRR, abs/1707.01154
http://arxiv.org/abs/1707.01154.

[Li et al., 2022] Li, Z., Wang, X., Stengel-Eskin, E., Kortylewski, A., Ma, W., Durme, B. V., and Yuille, A. L. (2022).
Super-CLEVR: A virtual benchmark to diagnose domain robustness in visual reasoning.
CoRR, abs/2212.00259
https://arxiv.org/abs/2212.00259.

[Liu et al., 2023] Liu, A., Xu, H., Van den Broeck, G., and Liang, Y. (2023).
Out-of-distribution generalization by neural-symbolic joint training.
In *AAAI Conference on Artificial Intelligence*, volume 37(10), pages 12252–12259
DOI:10.1609/aaai.v37i10.26444.

[Lorena et al., 2019] Lorena, A. C., Garcia, L. P. F., Lehmann, J., de Souto, M. C. P., and Ho, T. K. (2019).
How complex is your classification problem?: A survey on measuring classification complexity.
ACM Computing Surveys, 52(5):107:1–107:34
https://dl.acm.org/doi/10.1145/3347711.

References V

[Lu et al., 2020] Lu, Y., Cheung, Y., and Tang, Y. Y. (2020).
Bayes imbalance impact index: A measure of class imbalanced data set for classification problem.
IEEE Transactions on Neural Networks and Learning Systems, 31(9):3525–3539
https://ieeexplore.ieee.org/document/8890005.

[Mao et al., 2019] Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., and Wu, J. (2019).
The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision.
In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.
OpenReview.net
https://openreview.net/forum?id=rJgMlhRctm.

[Nardi, 1996] Nardi, B. A., editor (1996).
Context and Consciousness: Activity Theory and Human-Computer Interaction.
MIT Press
http://portal.acm.org/citation.cfm?id=223826.

[Nguyen and Martínez, 2020] Nguyen, A. and Martínez, M. R. (2020).
On quantitative aspects of model interpretability.
CoRR, abs/2007.07584
https://arxiv.org/abs/2007.07584.

[Northcutt et al., 2021] Northcutt, C. G., Jiang, L., and Chuang, I. L. (2021).
Confident learning: Estimating uncertainty in dataset labels.
Journal of Artificial Intelligence Research, 70:1373–1411
https://jair.org/index.php/jair/article/view/12125.

References VI

[Nye et al., 2021] Nye, M. I., Tessler, M. H., Tenenbaum, J. B., and Lake, B. M. (2021).
Improving coherence and consistency in neural sequence models with dual-system, neuro-symbolic reasoning.
In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pages 25192–25204
https://proceedings.neurips.cc/paper/2021/hash/d3e2e8f631bd9336ed25b8162aef8782-Abstract.html.

[Omicini, 2020] Omicini, A. (2020).
Not just for humans: Explanation for agent-to-agent communication.
In Vizzari, G., Palmonari, M., and Orlandini, A., editors, *AIxIA 2020 DP —– AIxIA 2020 Discussion Papers Workshop*, volume 2776 of *AI*IA Series*, pages 1–11, Aachen, Germany. Sun SITE Central Europe, RWTH Aachen University
http://ceur-ws.org/Vol-2776/paper-1.pdf.

[Pascal, 1669] Pascal, B. (1669).
Pensées.
Guillaume Desprez, Paris, France
.

[Popper, 2002] Popper, K. R. (2002).
The Logic of Scientific Discovery.
Routledge, London, UK, 2nd edition.
1st English Edition: 1959
DOI:10.4324/9780203994627.

[Sachs et al., 2023] Sachs, J. D., Lafortune, G., Fuller, G., and Drumm, E. (2023).
Implementing the SDG Stimulus. Sustainable Development Report 2023.
Dublin University Press, Dublin, Ireland
DOI:10.25546/102924.

References VII

[Skinner, 1985] Skinner, B. F. (1985).
Cognitive science and behaviourism.
British Journal of Psychology, 76(3):291–301
DOI:10.1111/j.2044-8295.1985.tb01953.x.

[Škrlj et al., 2021] Škrlj, B., Martinc, M., Lavrač, N., and Pollak, S. (2021).
autoBOT: evolving neuro-symbolic representations for explainable low resource text classification.
Machine Learning, 110(5):989–1028
https://link.springer.com/article/10.1007/s10994-021-05968-x.

[Sloman and Fernbach, 2018] Sloman, S. and Fernbach, P. (2018).
The Knowledge Illusion: Why We Never Think Alone.
Penguin Random House
https://www.penguinrandomhouse.com/books/533524/the-knowledge-illusion-by-steven-sloman-and-philip-fernbach/.

[Vilamala et al., 2023] Vilamala, M. R., Xing, T., Taylor, H., Garcia, L., Srivastava, M., Kaplan, L. M., Preece, A. D., Kimmig, A., and Cerutti, F. (2023).
DeepProbCEP: A neuro-symbolic approach for complex event processing in adversarial settings.
Expert Systems with Applications, 215:119376:1–26
https://www.sciencedirect.com/science/article/pii/S0957417422023946.

[Voigt and von dem Bussche, 2017] Voigt, P. and von dem Bussche, A. (2017).
The EU General Data Protection Regulation (GDPR). A Practical Guide.
Springer
DOI:10.1007/978-3-319-57959-7.

References VIII

[Wagner and d'Avila Garcez, 2021] Wagner, B. and d'Avila Garcez, A. (2021).
Neural-symbolic integration for fairness in AI.
In Martin, A., Hinkelmann, K., Fill, H., Gerber, A., Lenat, D., Stolle, R., and van Harmelen, F., editors, *AAAI-MAKE 2021 – Combining Machine Learning and Knowledge Engineering*, volume 2846 of *CEUR Workshop Proceedings*. CEUR-WS.org
https://ceur-ws.org/Vol-2846/paper5.pdf.

[Xie et al., 2022] Xie, X., Kersting, K., and Neider, D. (2022).
Neuro-symbolic verification of deep neural networks.
In De Raedt, L., editor, *Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022*, pages 3622–3628. ijcai.org
https://www.ijcai.org/proceedings/2022/503.

[Yang and Chaudhuri, 2022] Yang, C. and Chaudhuri, S. (2022).
Safe neurosymbolic learning with differentiable symbolic execution.
In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net
https://openreview.net/forum?id=NYBmJN4MyZ.

[Zhang et al., 2021] Zhang, Q., Wang, L., Yu, S., Wang, S., Wang, Y., Jiang, J., and Lim, E. (2021).
NOAHQA: Numerical reasoning with interpretable graph question answering dataset.
In Moens, M., Huang, X., Specia, L., and Yih, S. W., editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4147–4161. ACL
https://aclanthology.org/2021.findings-emnlp.350/.