We Were Deep in NeSy When LLM Happened

Andrea Omicini   Andrea Agiollo   Giovanni Ciatto   Matteo Magnini

Dipartimento di Informatica – Scienza e Ingegneria (DISI)
Alma Mater Studiorum – Università di Bologna
andrea.omicini@unibo.it   andrea.agiollo@studio.it
giovanni.ciatto@studio.it   matto.magnini@unibo.it

HyperModeLex Meeting
Bologna, Italy – 27 September 2024

Next in Line. . .

Neuro-symbolic Integration Systems as Nature-Inspired

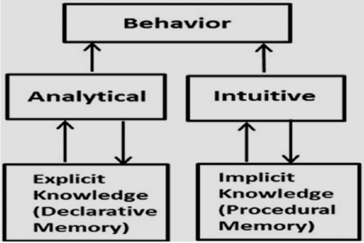
- neuro-symbolic integration systems (NeSy) integrate neural (subsymbolic) and symbolic AI approaches
  - blending the *subsymbolic* perspective of ML and DL agents with *symbolic* AI solutions focusing on high-level symbolic (human-readable) representations of problems, logic, and search
- given that
  - 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[Liu and Tsui, 2006]

Humans as NeSy I

Originally *not* from the CS / AI fields

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



- *rationality vs. intuition* roughly matches 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 Social 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 artificial agents[Omicini, 2020]
- for instance, symbolic approaches are critical when dealing with systems features such as
  - explainability
  - understandability
  - accountability
  - trustworthiness
- so, we focus on NeSy by putting some extra care on the *interaction* aspect of symbolic/subsymbolic integration

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 fundamental question for us becomes
  - ? how are we going to shape the interaction between heterogeneous intelligent components within *intelligent socio-technical systems*?

?? e.g., is NLP the answer? Or, LLM?

Rational AI is *not* the Handmaid of ML

Old story. . .

- we know that symbolic techniques can help sub-symbolic ones in making intelligent STS understandable
- yet, this does *not* makes rational AI a sort of *ancilla* of non-rational AI
- as in

  "we cannot really do anything meaningful with that, but maybe we could somehowhelp the actually-working AI"

! whereas we comfortably stand on the side of

  "yeah, no, we need it anyway"

for intelligent systems in general

Next in Line. . .

Relevant Questions for XAI

1. What are we trying to explain?
   In general, AI-based systems
2. Who is in charge of producing explanations?
   The AI system itself? Human experts? Ordinary users?
3. To whom are explanations addressed?
   Humans (developers, end users)? Other AI systems?
4. How are we going to create explanations?
   This is the actual core of XAI research
5. Which are the most adequate sorts of explanation?
   This depends on the answers to the questions above
6. When should explanations be presented to the user?
   This, too, depends on the answers to the questions above

Current Practice of XAI

1. What are we trying to explain?
   - mostly data-driven, ML-powered systems
2. Who is in charge of producing explanations?
   - AI experts, data scientists, ML engineers
3. To whom are explanations addressed?
   - people having a certain degree of expertise in AI/ML
4. How are we going to create explanations?
   - via task-, model-, and data-specific algorithms
5. Which are the most adequate sorts of explanation?
   - depends on task, model, data, and consumer at hand
   - other than on the available XAI algorithms
6. When should explanations be presented to the user?
   - mostly in the training phase; possibly in inference phase

The Future of XAI

1. What are we trying to explain?
   - any system including computational agents with some degree of autonomy
2. Who is in charge of producing explanations?
   - the system, i.e., the agents themselves
3. To whom are explanations addressed?
   - people with diverse levels of expertise
   - other computational agents
4. How are we going to create explanations?
   - via task-, model-, and data-specific algorithms
   - plus consumer-specific presentation strategies
5. Which are the most adequate sorts of explanation?
   - the ones which better adapt to the needs of the user
6. When should explanations be presented to the user?
   - upon request—i.e., as part of a dialogue

Explain What? I

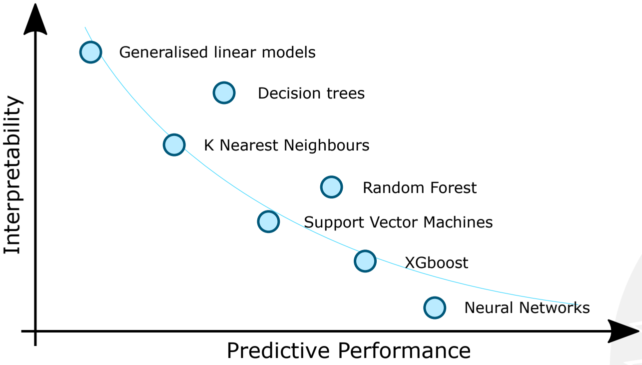
Most efforts are devoted to *supervised* ML, and in particular

- specific sorts of tasks, e.g. classification and regression
- specific sorts of data, e.g. images, text, or tables
- specific sorts of predictors, e.g. neural networks, SVM

  i.e. essentially, functions of the form $f : \mathcal{X} \subseteq \mathbb{R}^n \rightarrow \mathcal{Y} \subseteq \mathbb{R}^m$

Explain What? II

Trade-off between *interpretability* and *predictivity*



Explain What? III

Conventionally...

- ...linear models, or decision trees/rules are considered as interpretable
- ...other kinds of predictors are considered poorly interpretable
  - hence in need of explanation

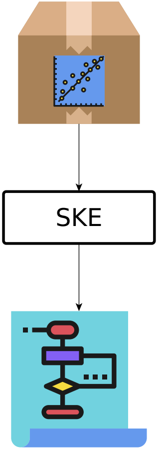
Explain What? IV

Our focus is on *supervised ML*, but XAI is wider than that

- explainable *unsupervised* learning
  - e.g., clustering [Sabbatini and Calegari, 2022]
- explainable reinforcement learning (XRL) [Milani et al., 2022]
- explainable planning (XAIP) [Hoffmann and Magazzeni, 2019]
- explainable agents and robots (XMAS) [Ciatto et al., 2019, Anjomshoae et al., 2019]
- ...

Next in Line. . .

Overview I



Main idea behind SKE is...

- to demystify neural networks by extracting symbolic knowledge out of them
- to use the extracted knowledge to generate explanations

Insight

- search of a surrogate interpretable model...
- ...consisting of symbolic knowledge

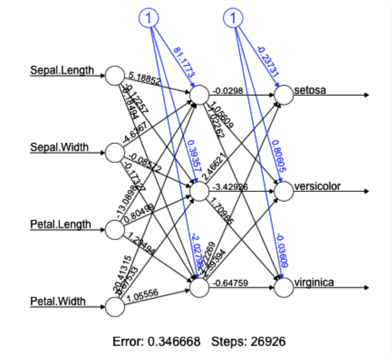
Overview II

Symbolic Knowledge Extraction (SKE) [Ciatto et al., 2024b]

Any algorithmic procedure accepting trained sub-symbolic predictors as input and producing symbolic knowledge as output, in such a way that the extracted knowledge reflects the behaviour of the predictor with high fidelity

Overview III

Example



$\rightarrow$

$$\mathit{Class} = \texttt{setosa} \leftarrow \mathit{PetalWidth} \leq 1.0.$$

$$\mathit{Class} = \texttt{versicolor} \leftarrow \mathit{PetalLength} > 4.9 \wedge \mathit{SepalWidth} \in [2.9,\ 3.2].$$

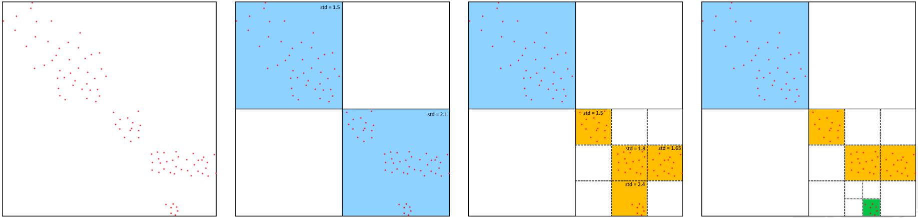
$$\mathit{Class} = \texttt{versicolor} \leftarrow \mathit{PetalWidth} > 1.6.$$

$$\mathit{Class} = \texttt{virginica} \leftarrow \mathit{SepalWidth} \leq 2.9.$$

$$\mathit{Class} = \texttt{virginica} \leftarrow \mathit{SepalLength} \in [5.4,\ 6.3].$$

$$\mathit{Class} = \texttt{virginica} \leftarrow \mathit{PetalWidth} \in [1.0,\ 1.6].$$

SKE with GridEx I



(a) Surrounding cube

(b) Iteration 1 ($p_1 = 2$)

(c) Iteration 2 ($p_2 = 3$).

(d) Iteration 3 ($p_3 = 2$).

Figure: Example of GridEx's hyper-cube partitioning *(merging step not reported)*

SKE with GridEx II

Using GridEx for SKE [Sabbatini et al., 2021]

1. partition the input space into p_1^n hypercubes
   - evenly splitting the n dimensions into p_1 bins
2. partition each non empty-region into p_2^n hypercubes
   - evenly splitting the n dimensions into p_2 bins
3. repeat the splitting arbitrarily
4. assign a prediction with each non-empty partition (e.g. average value)
5. write an if-then rule for each non-empty partition:

   if expressions delimiting the partition
   then prediction of that partition

Next in Line...

Why SKI?

Benefits

- prevent the predictor to become a black-box
- reduce learning time
- reduce the data size needed for training
- improve predictor's accuracy
- build a predictor that behave as a logic engine

Symbolic Knowledge Injection I

Key insights

- a priori altering ML predictors. . .
- . . . to make they comply to user-provided knowledge. . .
- . . . which is represented in symbolic form

Main idea behind SKI

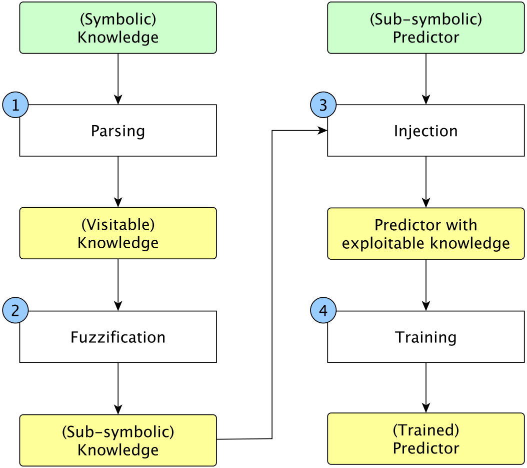
- *forcing* NN to comply with some *intensional symbolic knowledge* while learning, hence making their operation more controllable, and their outcomes more predictable

Symbolic Knowledge Injection II

Symbolic Knowledge Injection (SKI) [Ciatto et al., 2024b]

Any algorithmic procedure affecting how sub-symbolic predictors draw their inferences in such a way that predictions are either computed as a function of, or made consistent with, some given symbolic knowledge

Symbolic Knowledge Injection– General Workflow I



SKI with KINS I

Knowledge Injection via Network Structuring (KINS)[Magnini et al., 2023]

A general SKI algorithm not bound to any specific sort of sub-symbolic predictor

- aim → enrich
- predictor → neural network
- how → structuring
- logic → stratified Datalog with negation

Public implementation on PSyKI[Magnini et al., 2022]

Next in Line...

What about LLM, SKE& SKI?

Fundamental questions

- ? are SKE and SKI still relevant in the LLM era?
  - ? after all, LLM "speak the language", so do we really need explanations of any sort?
  - ! still, the generative mechanism is obscure, so we may get the "what" but not the "why"
- ? in case they are, how do we deal with them?
  - LLM are neural networks, yet they are so huge that
    - extracting the whole knowledge of a LLM is challenging
    - injecting symbolic knowledge into the LLM seems to be technologically problematic, or, even not feasible

About SKE & LLM

Yep, still useful

- for instance, when we need to extract *structured knowledge* from LLM
- maybe not necessarily useful for TAI / XAI, but quite important when the knowledge acquired from the LLM must be "brought back" in symbolic form
  - idea: bridging LLM with "ordinary" software technologies, which expect some clean data schemas as inputs

SKE & LLM: An Example I

Recommendation systems

- providing hints to users based on their preferences, and on the profiles of similar users
- chicken-and-egg problem, namely the *cold start*
  - the system needs data to operate effectively
  - the system acquires data from users during operation
  - no data, and no users at the beginning
  - how do we get out of this?

SKE & LLM: An Example II

Our example: virtual nutrition coach

- as part of the CHIST-ERA IV project Expectation, we needed to
  - design a virtual coach for nutrition, providing users with personalised advices on *what to eat and when*
  - feed the system with data about
    - food e.g. recipes, their ingredients, their nutritional values, etc.
    - users e.g. their preferences and their habits, goals, medical issues, etc.
  - generate the data schema
  - find some data matching the schema

SKE & LLM: An Example III

Task: filling the CHIST-ERA ontology

- the ontology schema was designed as one of the results of the project (*TBox*)
- how to populate the ontology?
  - by assigning individuals to concepts or roles (*ABox*)?
- typically done manually, and so very expensively
- called *ontology learning*, when done (semi)automatically

! idea: replacing domain experts with LLM, using them as oracles for automating ontology population

SKE & LLM: An Example IV

Our algorithm: KGFILLER [Ciatto et al., 2024a]

Our algorithm for ontology population through LLM, stepping through 4 phases

population phase each concept is populated with a set of individuals

relation phase each property is populated with a set of role assertions

- as a by-product, some concepts may be populated even further

redistribution phase some individuals are reassigned to more adequate concepts

merge phase similar individuals are merged into a single one

- mostly, duplicate removal

SKE, LLM & MAS

- LLM increasingly push intelligent agents towards sub-symbolic models for NLP tasks in human-agent interaction
- lack of transparency, however, hinders LLM applicability there
- many approaches around focusing on local post-hoc explanations (LPE) by the XAI community in NLP realm
- LPE represents a popular solution to explain the reasoning process by highlighting how different portions of the input sample impact differently the produced output, by assigning a *relevance score* to each input component
- yet, most local post-hoc explainers are loosely correlated—agents *quarrel* about explanations [Agiollo et al., 2023]

SKE, LLM & MAS: GELPE I

Global Explanations from Local Post-hoc Explainers [Agiollo et al., 2024]

- aggregation of the local explanations obtained by each local post-hoc explainer into a set of *global impact scores*
- extraction of *global explanations* from the output of LPE agents, enabling the extraction of logic rules from LLM

What about SKI & LLM?

Some space for research

- SKI, too, may make sense for LLM
  - possibly not as a means for TAI
  - instead, as a way to teach / force LLM to *reason* somehow
- research question: can SKI work for LLM
  - via *prompt engineering*?
  - via *fine tuning*?
- main challenge
  - training LLM is technologically impractical
  - so SKI should work with *no need to re-train*

We Were Deep in NeSy When LLM Happened

Andrea Omicini Andrea Agiollo Giovanni Ciatto Matteo Magnini

Dipartimento di Informatica – Scienza e Ingegneria (DISI)
Alma Mater Studiorum – Università di Bologna
andrea.omicini@unibo.it andrea.agiollo@studio.it
giovanni.ciatto@studio.it matto.magnini@unibo.it

HyperModeLex Meeting
Bologna, Italy – 27 September 2024

References I

[Agiollo et al., 2023] Agiollo, A., Siebert, L. C., Murukannaiah, P. K., and Omicini, A. (2023).
The quarrel of local post-hoc explainers for moral values classification in natural language processing.
In Calvaresi, D., Najjar, A., Omicini, A., Aydoğan, R., Carli, R., Ciatto, G., Mualla, Y., and Främling, K., editors, *Explainable and Transparent AI and Multi-Agent Systems*, volume 14127 of *Lecture Notes in Computer Science*, chapter 6, pages 97–115. Springer
(APICe) DOI:10.1007/978-3-031-40878-6_6

[Agiollo et al., 2024] Agiollo, A., Siebert, L. C., Murukannaiah, P. K., and Omicini, A. (2024).
From large language models to small logic programs: Building global explanations from disagreeing local post-hoc explainers.
Autonomous Agents and Multi-Agent Systems, 38:1–33.
Special Issue on Multi-Agent Systems and Explainable AI
DOI:10.1007/s10458-024-09663-8

[Anjomshoae et al., 2019] Anjomshoae, S., Najjar, A., Calvaresi, D., and Främling, K. (2019).
Explainable agents and robots: Results from a systematic literature review.
In Elkind, E., Veloso, M., Agmon, N., and Taylor, M. E., editors, *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*, pages 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems
http://dl.acm.org/citation.cfm?id=3331806

[Ciatto et al., 2024a] Ciatto, G., Agiollo, A., Magnini, M., and Omicini, A. (2024a).
Large language models as oracles for instantiating ontologies with domain-specific knowledge.
Submitted
(APICe)

References II

[Ciatto et al., 2019] Ciatto, G., Calegari, R., Omicini, A., and Calvaresi, D. (2019).
Towards XMAS: eXplainability through Multi-Agent Systems.
In Savaglio, C., Fortino, G., Ciatto, G., and Omicini, A., editors, *AI&IoT 2019 – Artificial Intelligence and Internet of Things 2019*, volume 2502 of *CEUR Workshop Proceedings*, pages 40–53. CEUR WS
http://ceur-ws.org/Vol-2502/paper3.pdf

[Ciatto et al., 2024b] Ciatto, G., Sabbatini, F., Agiollo, A., Magnini, M., and Omicini, A. (2024b).
Symbolic knowledge extraction and injection with sub-symbolic predictors: A systematic literature review.
ACM Computing Surveys, 56(6):161:1–161:35
DOI:10.1145/3645103

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

[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

References III

[Hoffmann and Magazzeni, 2019] Hoffmann, J. and Magazzeni, D. (2019).
Explainable AI planning (XAIP): overview and the case of contrastive explanation (extended abstract).
In Krötzsch, M. and Stepanova, D., editors, *Reasoning Web. Explainable Artificial Intelligence - 15th International Summer School 2019, Bolzano, Italy, September 20-24, 2019, Tutorial Lectures*, volume 11810 of *Lecture Notes in Computer Science*, pages 277–282. Springer
DOI:10.1007/978-3-030-31423-1_9

[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

[Liu and Tsui, 2006] Liu, J. and Tsui, K. C. (2006).
Toward nature-inspired computing.
Communications of the ACM, 49(10):59–64
(APICe) DOI:10.1145/1164394.1164395

[Magnini et al., 2022] Magnini, M., Ciatto, G., and Omicini, A. (2022).
On the design of PSyKI: a platform for symbolic knowledge injection into sub-symbolic predictors.
In Calvaresi, D., Najjar, A., Winikoff, M., and Främling, K., editors, *Explainable and Transparent AI and Multi-Agent Systems*, volume 13283 of *Lecture Notes in Computer Science*, chapter 6, pages 90–108. Springer.
4th International Workshop, EXTRAAMAS 2022, Virtual Event, May 9–10, 2022, Revised Selected Papers
DOI:10.1007/978-3-031-15565-9_6

[Magnini et al., 2023] Magnini, M., Ciatto, G., and Omicini, A. (2023).
Knowledge injection of Datalog rules via neural network structuring with KINS.
Journal of Logic and Computation, 33(8):1832–1850
DOI:10.1093/logcom/exad037

References IV

[Milani et al., 2022] Milani, S., Topin, N., Veloso, M., and Fang, F. (2022).
A survey of explainable reinforcement learning.
CoRR, abs/2202.08434
https://arxiv.org/abs/2202.08434

[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

[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

References V

[Sabbatini and Calegari, 2022] Sabbatini, F. and Calegari, R. (2022).
Clustering-based approaches for symbolic knowledge extraction.
In *XLoKR 2022 - Third Workshop on Explainable Logic-Based Knowledge Representation*, Haifa, Israel
https://arxiv.org/abs/2211.00234

[Sabbatini et al., 2021] Sabbatini, F., Ciatto, G., and Omicini, A. (2021).
GridEx: An algorithm for knowledge extraction from black-box regressors.
In Calvaresi, D., Najjar, A., Winikoff, M., and Främling, K., editors, *Explainable and Transparent AI and Multi-Agent Systems. Third International Workshop, EXTRAAMAS 2021, Virtual Event, May 3–7, 2021, Revised Selected Papers*, volume 12688 of *LNCS*, pages 18–38. Springer Nature, Basel, Switzerland
DOI:10.1007/978-3-030-82017-6_2

[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

[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/

Funding Projects

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
- the CHIST-ERA IV project "Expectation" – CHIST-ERA-19-XAI-005 –, co-funded by EU and the Italian MUR (Ministry for University and Research)
- "AEQUITAS" project funded by European Union's Horizon Europe research and innovation programme under grant number 101070363
- "ENGINES — ENGineering INtElligent Systems around intelligent agent technologies" project funded by the Italian MUR program "PRIN 2022" under grant number 20229ZXBZM