

Self-Explaining Agents

Andrea Omicini
andrea.omicini@unibo.it

Dipartimento di Informatica – Scienza e Ingegneria (DISI)
ALMA MATER STUDIORUM – Università di Bologna

ERC CompuLaw Final Conference
Florence, Italy, 14 July 2025



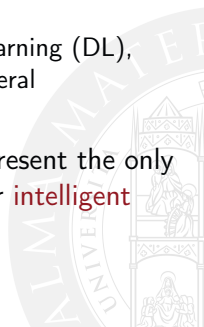
Next in Line...

- 1 Agents & Intelligent Systems
- 2 Intelligence, Rationality, Sharing
- 3 Explanators & Explainees
- 4 Back to Agents
- 5 Conclusion



Agents

- *agents* and *multi-agent systems* (MAS) play the role of the **sources** of abstractions for the design and development of intelligent systems
- the notion of agent is typically adopted to frame every single AI technique within a coherent *architecture*
- MAS provide the conceptual and technical ground for *assembling* agents within working *intelligent systems*
 - this includes recent AI advancements, including deep learning (DL), large-language models (LLM), and generative AI in general
- so that *agent-oriented software engineering* (AOSE) represent the only viable foundation of any current and future discipline for **intelligent systems engineering**



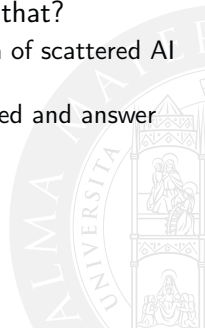
Humans

- with respect to traditional software engineering (SE), intelligent systems often need **humans** to play an *active part* in the systems, rather than merely be a client to be served
 - so, humans are still expected to bring their own *goals* to the table
 - but they are now supposed to also bring their own knowledge, skills, understanding of the world, . . .
 - which the overall system relies upon for its intelligent behaviour
- as a result, nowadays humans work as legitimate *components* of intelligent systems in the same way as software and physical agents do
 - intelligent **socio-technical systems** (STS)
- **agent-oriented software engineering** perfectly fits the scope of intelligent STS
 - where the *agent abstraction* accounts for activity, knowledge, intelligence, goals, learning. . .
 - from *both* human and artificial agents



Which Agents for Intelligent STS?

- *intelligent STS* as the general **model** for AI systems
- *AOSE as the general discipline* for AI systems **engineering**
- ? general question: how do we **shape agents** according to that?
 - ! currently, a lot of confused ideas around, and a plethora of scattered AI technologies emerging worldwide
 - ... maybe we need some *further questions* before we proceed and answer



Who Needs Explanations?

(Is it Just) People (Who) need to understand(?)

- intelligent STS ask both humans and intelligent agents to work together to generate *intelligent behaviour*
 - e.g. *decision support systems* exploit intelligent systems in order to promote rational human decisions—in health care, urban traffic management, legal procedure, ...
- ! information and actions by intelligent agents need to be **understandable** by humans to be accepted and *trusted*
 - as well as for *responsibility* and *accountability* reasons
- humans need *explanations*
 - which is where **explainable artificial intelligence** (XAI) comes from^[Gunning, 2016]
- ... just humans?

Agents & Neuro-symbolic Integration

- the technical XAI problem
 - *symbolic* approaches are *transparent* yet **slow**
 - *sub-symbolic* approaches are *fast* yet **opaque**
- so, **neuro-symbolic integration** is the most promising way out
 - and, everyone is already doing that^[Calegari et al., 2020]
- yet: what about agents?
 - integration how?
 - and, mostly, based on what **integration model**?
 - which *conceptual foundation* for neuro-symbolic integration techniques within a coherent **agent model** / **architecture**?



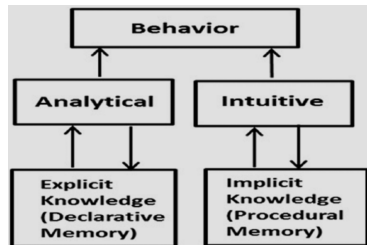
Next in Line...

- 1 Agents & Intelligent Systems
- 2 Intelligence, Rationality, Sharing**
- 3 Explanators & Explainees
- 4 Back to Agents
- 5 Conclusion



Rationality vs. Intuition

- ! human cognition processes are our main reference for modelling and shaping intelligent processes
- two sorts of cognitive processes in humans
 - *esprit de finesse* vs. *esprit de géométrie*—rationality has limits [Pascal, 1669]
 - *cognitivism* against *behaviourism* in psychology [Skinner, 1985]
 - *fast* vs. *slow thinking* [Kahneman, 2011]
- generally speaking, cognitive processes in humans are widely recognised to be either **rational** or **intuitive**
 - and, of course, human cognition can be mostly understood as a basically non-linear combination of the two



Social Intelligence: Humans Share Knowledge

- ! human cognition is *not* just an individual process^[Kirsh, 1999]
- 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]



Sharing is Rational

- we already know that there is *intelligence without representation*^[Brooks, 1991b] and *without reason*^[Brooks, 1991a]
 - yet, human cumulative culture is based on *representation* tools—language, writing, books, the Web
- *repeatable, systematic* sharing requires **rational representation**
 - even (or, maybe, in particular) when we are sharing *intuitive, implicit knowledge*
- and, *sharing implicit knowledge* typically calls for **rational explanation**



Next in Line...

- 1 Agents & Intelligent Systems
- 2 Intelligence, Rationality, Sharing
- 3 Explanators & Explainees**
- 4 Back to Agents
- 5 Conclusion



Explanation Everywhere

- the notion of *explanation* is the core of many research efforts
 - along with accessory notions such as *interpretation* and *understandability*
- and undergone a constant flow of diverse and (sometimes) even extravagant definitions
 - e.g., even GDPR^[Voigt and von dem Bussche, 2017] recognises “the citizens’ right to explanation”^[Goodman and Flaxman, 2017]
- most of them encompassing both the **explainer**’s and the **explainee**’s acts in the same acceptance of the term ‘explanation’
 - the *dialogical* acceptance of explanation



Explanation as an Explanator's Act

- however, if we start by adopting the dialectical notion of explanation, how can we understand explanation in the most abstract way?
 - not everything in an explanation depends just on the explainee
- e.g., the same explanation could be given by an expert to a single friend, to a class of listeners in a physical room, to an undetermined audience on a social platform. . .
 - how can we model explanation independently of the explainees?
 - and possibly measure its effectiveness in some way?
- ! even though the dialectical notion of explanation is perfectly legitimate, here we stick to the notion of

explanation as an explanator's act,

so that

- we can model it in the most abstract way, and
- we can focus on the (individual) *cognitive process* of the *explanator*



Explanation in Math Teaching I

- contribution from *math teaching* [D'Amore, 2005]
 - being math the most difficult subject to explain & teach, math teaching is likely the best source of inspiration possible
- a **semiotic representation** is required whenever the object of an explanation is *inaccessible to perception*
 - semiotics** — acquisition of a *representation built out of signs*
 - noetics** — *conceptual acquisition* of an object
- e.g. a rectangle is an abstract geometrical concept, yet teachers draw some random rectangle on the blackboard when they try to make pupils grasp the abstract notion of rectangle
- how do they *explain* abstract concepts, promoting **learning as a noetic process** out of some **semiotic representation**?

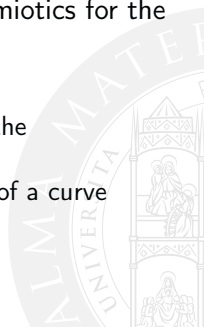
Explanation in Math Teaching II

- explanation as a (semiotic) *transformation*
 - explaining a concept by providing different *semiotic representations* of the same concept

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

transformation of conversion — changing register of semiotics for the representation

- e.g.
- converting between polar and Cartesian coordinates in the representation of a curve (*transformation of treatment*)
 - switching from the analytic to the geometric definition of a curve (*transformation of conversion*)



Explanation in Math Teaching III

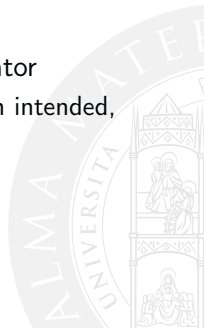
! *explanation as*

- first, **generation** of semiotic representation
- then, **transformation** of semiotic representation

by the explainer

! explanation is first of all an **individual act** of the explainer

- which either may or may not be directed explicitly to an intended, specific, pre-defined explainee



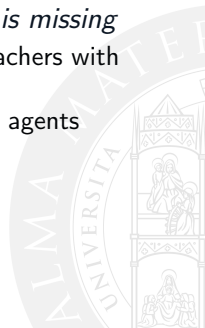
Next in Line...

- 1 Agents & Intelligent Systems
- 2 Intelligence, Rationality, Sharing
- 3 Explanators & Explainees
- 4 Back to Agents**
- 5 Conclusion



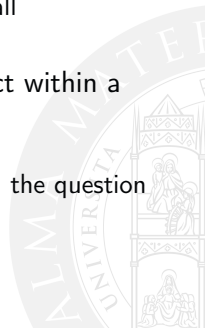
Is XAI Just Intelligent Systems Explaining to Humans?

- most of the research on explainability aims at making *intelligent systems understandable to humans*
 - explanators are *software components*—e.g., agents
 - explainees are *humans*^[Rosenfeld and Richardson, 2019, Anjomshoe et al., 2019, Guidotti et al., 2019]
- looking at STS, this means that at least *one dimension is missing*
 - we know about humans explaining to humans—e.g., teachers with pupils at schools
 - we know that humans try to be understood by software agents everyday—e.g., by Alexa & Siri, by ChatGPT & Gemini
 - ? what about **agents explaining to agents**?



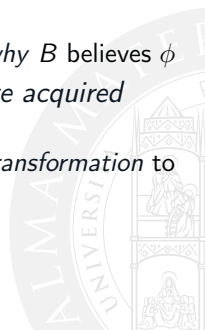
Effective Agent Interaction is Not Just Communication

- MAS are open, distributed, *heterogeneous* systems, working in *unpredictable environment* to solve *general-purpose* problems
- *intelligent MAS* are meant to **aggregate diverse sorts of intelligence**
 - each agent encapsulating different algorithms, techniques, logics
 - competing / cooperating to produce the best solution a run-time
 - either by selecting the best one, or integrating some / all
- ? how can hugely-heterogeneous, intelligent agents interact within a MAS so as to understand each other?
 - yet, based on which conceptual ground?
 - what if we replace “intelligent agents” with “humans” in the question above—as in intelligent STS?



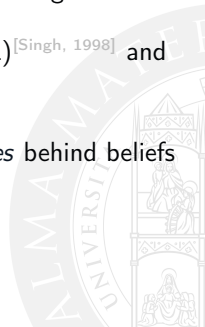
Agent Sharing in MAS

- *effective interaction amongst intelligent agents* in open and heterogeneous MAS mandates for **knowledge sharing among agents**
 - in the same way as humans share knowledge and cognition in cooperative contexts
 - so, beyond mere *inter-agent communication*
- it is not enough for agent A to know that B believes ϕ
 - in order to accept ϕ , A needs to understand (at least) *why* B believes ϕ
- also dealing with – and possibly focusing on – *knowledge acquired through sub-symbolic techniques*
 - which needs a *symbolic representation* and possibly a *transformation* to be shared rationally among heterogeneous agents



Agent Explaining to Agents

- effective interaction among agents in non-trivial MAS cannot be just based on mere agent communication
 - *complex agent cognitive processes* need to be *shared* not just in terms of their final results
- ! **explanation** as the main tool for **sharing knowledge among agents**
 - having the potential to work as an actual enabler of full-fledged *agent-to-agent communication*
 - so that tools like *agent communication languages* (ACL)^[Singh, 1998] and *ontologies*^[Huhns and Singh, 1997] are just technical premises
 - the syntactic and semantic premises, respectively
 - whereas explanation aims at sharing the *cognitive causes* behind beliefs
- ! so, first and foremost, **agents explaining to agents**



Self-explaining Agents

- agent-based intelligent systems made of **self-explaining agents**^[Omicini, 2020]
 - equipped with their own specific *rational explanation* capability, whatever their nature
 - able of providing a rational, *sharable representation* of their own specific cognitive process and results
 - capable of *manipulating* such a representation in order to build one or more explanations
- an **architectural requirement** for intelligent agents of any sort is that they need to be equipped with their own *explanation module*
 - matching the specific agent cognitive process so as to produce a symbolic representation of the process itself as well as of its results
- also, agents should be provided with enough **knowledge** of the **domain of discourse** to make them capable of *symbolic manipulation* of the *representation* in the form of a *semiotic transformation*
 - to make them able to provide *rational explanations* in symbolic form

Case Study: Decision Support System for Legal Process

- MAS with heterogeneous legal agents
 - some legal agents could be *deep-learning agents* trained over diverse sets of existing legal databases
 - others might be *logic-based agents*, rationally elaborating over some symbolic representation of some legal corpus
- each one not just providing its own argument and supporting its suggestions
 - not just directly providing humans with explanations
- instead, cooperatively building a *shared proposal* to be presented to human decision-makers in a potentially-understandable way
- ! this requires first of all that agents – symbolic vs. sub symbolic vs. hybrid – are capable of **explaining themselves** through a *rational form* that could be effectively **shared** not just with humans, but also **with other agents**
 - possibly exploiting *argumentation* techniques ^[Billi et al., 2021]

Next in Line...

- 1 Agents & Intelligent Systems
- 2 Intelligence, Rationality, Sharing
- 3 Explanators & Explainees
- 4 Back to Agents
- 5 Conclusion**



Overall

- ① **explanation** should be deemed as an *essential tool* for any *intelligent component*
 - in particular, **agents** in multi-agent systems
- ② intelligent agents should be able to **explicitly represent** their cognitive processes and their results
 - independently of their nature—symbolic / sub-symbolic / hybrid
- ③ and **manipulate** those representations
 - so that **rational** explanation would properly complement their ability to *reason* and *communicate*
- ④ intelligent agents should be able to *explain* themselves first of all **to other agents**
 - not just be part of a system explaining itself to humans
- ⑤ **symbolic** techniques are to be used for explanations—representing and manipulating cognitive processes and their results
 - so, as *first-class citizens* in both *agent modelling* and *intelligent systems engineering*
 - e.g., **argumentation** [Calegari et al., 2022]

Thanks. . .

. . . for your kind attention!

and, to Prof. Sartor for inviting me here



Self-Explaining Agents

Andrea Omicini
andrea.omicini@unibo.it

Dipartimento di Informatica – Scienza e Ingegneria (DISI)
ALMA MATER STUDIORUM – Università di Bologna

ERC CompuLaw Final Conference
Florence, Italy, 14 July 2025



References I

- [Anjomshoe et al., 2019] Anjomshoe, S., Najjar, A., Calvaresi, D., and Främling, K. (2019).
Explainable agents and robots: Results from a systematic literature review.
 In *18th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'19)*, pages
 1078–1088. IFAAMAS
 DOI:10.5555/3306127.3331806
- [Billi et al., 2021] Billi, M., Calegari, R., Contissa, G., Lagioia, F., Pisano, G., Sartor, G., and Sartor, G. (2021).
Argumentation and defeasible reasoning in the law.
J, 4(4):897–914.
 Special Issue The Impact of Artificial Intelligence on Law
 DOI:10.3390/j4040061
- [Brooks, 1991a] Brooks, R. A. (1991a).
Intelligence without reason.
 In Mylopoulos, J. and Reiter, R., editors, *12th International Joint Conference on Artificial Intelligence (IJCAI 1991)*, volume 1, pages 569–595, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
<http://dl.acm.org/citation.cfm?id=1631258>
- [Brooks, 1991b] Brooks, R. A. (1991b).
Intelligence without representation.
Artificial Intelligence, 47:139–159
 DOI:10.1016/0004-3702(91)90053-M
- [Calegari et al., 2020] Calegari, R., Ciatto, G., and Omicini, A. (2020).
On the integration of symbolic and sub-symbolic techniques for XAI: A survey.
Intelligenza Artificiale, 14(1):7–32
 DOI:10.3233/IA-190036



References II

- [Calegari et al., 2022] Calegari, R., Pisano, G., Omicini, A., and Sartor, G. (2022).
 Arg2P: An argumentation framework for explainable intelligent systems.
Journal of Logic and Computation, 32(2):369–401.
 Special Issue from the 35th Italian Conference on Computational Logic (CILC 2020)
 DOI:10.1093/logcom/exab089
- [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>
- [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
- [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
- [Guidotti et al., 2019] Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., and Giannotti, F. (2019).
 A survey of methods for explaining black box models.
ACM Computing Surveys, 51(5):1–42
 DOI:10.1145/3236009



References III

- [Gunning, 2016] Gunning, D. (2016).
Explainable artificial intelligence (XAI).
 Funding Program DARPA-BAA-16-53, Defense Advanced Research Projects Agency (DARPA)
<http://www.darpa.mil/program/explainable-artificial-intelligence>
- [Huhns and Singh, 1997] Huhns, M. N. and Singh, M. P. (1997).
Ontologies for agents.
IEEE Internet Computing, 1:81–83
 DOI:10.1109/4236.643942
- [Kahneman, 2011] Kahneman, D. (2011).
Thinking, Fast and Slow.
 Farrar, Straus and Giroux, USA
- [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
<https://philarchive.org/archive/KIRDCC>
- [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>



References IV

- [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
- [Rosenfeld and Richardson, 2019] Rosenfeld, A. and Richardson, A. (2019).
Explainability in human-agent systems.
Autonomous Agents and Multi-Agent Systems, 33(6):673–705
DOI:10.1007/s10458-019-09408-y
- [Singh, 1998] Singh, M. P. (1998).
Agent communication languages: Rethinking the principles.
Computer, 31(12):40–47
DOI:10.1109/2.735849
- [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



References V

[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



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
- “CompuLaw” project, funded by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (G.A. 833647)
- “ENGINES — ENGINEERING INtelligent Systems around intelligent agent technologies” project funded by the Italian MUR program “PRIN 2022” under grant number 20229ZXBZM

