

# Not just for humans: Explanation for agent-to-agent communication

Andrea Omicini

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

## Abstract

Once precisely defined so as to include just the explanation's act, the notion of *explanation* should be regarded as a central notion in the engineering of intelligent system—not just as an add-on to make them understandable to humans. Based on *symbolic* AI techniques to match intuitive and rational cognition, explanation should be exploited as a fundamental tool for *inter-agent communication* among heterogeneous agents in open multi-agent systems. More generally, *explanation-ready agents* should work as the basic components in the engineering of intelligent systems integrating both symbolic and sub-/non-symbolic AI techniques.

## Keywords

explanation, rational vs. intuitive cognition, multi-agent systems, agent communication, noetics vs. semiotics, transformation of semiotic register

## 1. On the Meaning of Terms

Whereas computer science (CS) has gone so far and so deep that can nowadays be claimed to belong within the social sciences [1], its basic foundations still lay in the ground of hard sciences, possibly on the edge of classic disciplines such as logics and mathematics. Along with the contemporary claim that “All science is computer science” [2], this makes the use of basic terms and definitions a problematic matter. In fact, the need for precise and non-ambiguous definitions of concepts and notions somehow clashes with the number of terms and ideas that, coming from human sciences, have (also) a more or less widespread and consolidated “commonsense” meaning—which hardly copes well with the requirements of scientific definition.

In the so-called “AI Renaissance”, with the pervasive emergence of artificial intelligence (AI) techniques everywhere in human activity and social organisations, this has also become an issue for notions and concepts for intelligent systems—also because popularisation of AI achievements and results have further mixed up scientific and everyday words. Thus, terms such as ‘learning’, ‘understanding’, ‘explaining’, just to mention a few, have become at the same time more and more central in AI literature, and less and less well-defined and understood—as they are widely used in both the scientific context and popular science. As one may expect, a flow of specific literature is nowadays dealing with that issue: this is for instance the case of the field of *explainable artificial intelligence* – XAI [3] –, where the term and the very notion of *explanation* represent one of the main subject of scientific debate—see e.g., [4].

However, it also worth to point out that the same issue is particularly relevant in the AI field, where the very notion of what is *intelligence* – and so, what is *artificial intelligence* – have always been a

---


AIxIA 2020 Discussion Papers Workshop

✉ andrea.omicini@unibo.it (A. Omicini)

🌐 <http://andreaomicini.apice.unibo.it> (A. Omicini)

🆔 0000-0002-6655-3869 (A. Omicini)

© 2020 Copyright for this paper by its authors.  
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

powerful undercurrent of inner conflict in the field itself. In straight comparison, the same does not hold for the CS field, where the issues ‘what is computing’ and ‘what is computer science’ are the subject of a more or less conventional epistemic debate—see e.g. [5].

So, why is the issue of basic terms, concepts, and definitions so much more critical to AI than CS, where it is a key issue anyways? The problem is likely to be related to the main objects of study: whereas *computation* is mostly perceived as an artificial product of human invention, *intelligence* is a sort of natural phenomenon that we (i) observe around us, but mostly in ourselves, and (ii) use to understand the world and *make science*—in the broadest acceptance of the term. The fact is, understanding intelligence is not (just) an issue for AI: instead, it is first of all an issue for humans in general—and, all things considered, it always has been.

## 2. Rationality vs. Intuition

### 2.1. Esprit de finesse vs. esprit de géométrie

*Le cœur a ses raisons que la raison ne connaît point* [6]

The efforts to understand the ways in which human cognition works date back to the earlier human thinkers; however, it is in the XVII century that the distinction between *rational* and *non-rational* thought becomes radical. That in fact is the age when the two most influential rationalistic systems (by Descartes and Spinoza, respectively) were born—where rationalism, in Western philosophy, is “the view that regards reason as the chief source and test of knowledge”<sup>1</sup> In response to that (“Descartes useless and uncertain” [6]), philosophers such as Blaise Pascal fiercely opposed to the view that the rational process was to be considered as the only possible way for human cognition. Apart from the anatomical identification of the heart (*cœur*) as the main *locus* of non-rational cognition (“Nous connaissons la vérité non seulement par la raison mais encore par le cœur”), the distinction is clear between the things that we can *understand by means of reason* (such as mathematical proofs) and the things *reason can not comprehend*. This is well expressed in Pascal by the duality between the *esprit de géométrie* and the *esprit de finesse*: the former representing the rational process, seen as *vues lentes, dures et inflexibles* (slow, difficult, inflexible views); the latter representing the intuitive mind, bringing about immediate comprehension of the things, characterised by the “souplesse de la pensée” (flexibility of thought).

In spite of the *naïveté* that contemporary scientist may observe in that distinction, modern epistemology has inherited the same dichotomy, when it opposes in the same way the “comprehension” against the objective and analytic “explication”.<sup>2</sup> Also, the same distinction seems to apply not just to the cognition process, but also to the knowledge itself.

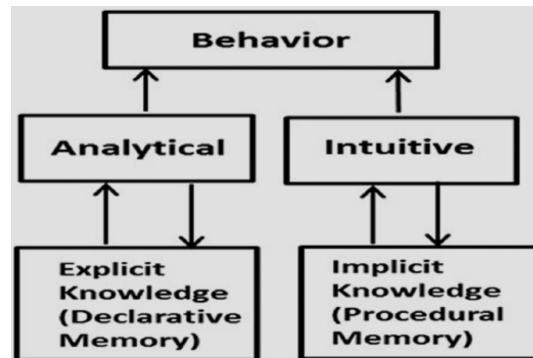
### 2.2. Analytical vs. intuitive

Psychology is one of the is foundational AI sciences: so, it cannot come as a surprise that the contrast between the two historical schools in psychology – that is, *cognitivism* against *behaviourism* [7] – was about the theory of the mental processes—cognitivism built around it, behaviourism refusing it. Nowadays, one of the most relevant distinctions in psychology is the one between *intuitive* vs. *analytical* cognition, as the sorts of cognition exploited in reasoning and decision making—see Figure 1. “Analytical cognition involves conscious deliberation that draws on limited working memory resources.

---

<sup>1</sup><https://www.britannica.com/topic/rationalism>

<sup>2</sup><https://la-philosophie.com/esprit-de-finesse-esprit-de-geometrie>



**Figure 1:** Two sorts of cognitive processes for human reasoning and decision making [8]

*Analytical cognition is voluntary, effortful, limited in capacity, and slow. Intuitive cognition involves unconscious situational pattern synthesis and recognition unconstrained by working memory limitations. Intuitive cognition is independent of conscious ‘executive’ control, large in capacity, and fast. These two types of reasoning and decision making can be dissociated experimentally and neurologically.” [8].*

Intuitive cognition is the default whenever analytical cognition is faced with tasks it cannot accomplish, and is critical to survival. As such, intuitive cognition – rather than analytical cognition – is seen as representing the core of human cognition, being instrumental in handling the situated complexities of everyday living for humans.

### 2.3. Symbolic vs. sub-/non-symbolic

Even though the existence and the importance of non-rational human cognition is so well-known and understood, the evolution of AI over the years have shown a somehow wavering awareness of this.

On the one hand, the dichotomy “intelligent process vs. intelligent behaviour” has informed the AI field since its very beginning, basically classifying the different AI approaches as, respectively, capturing and exploiting the basic mechanisms of human intelligence vs. behaving intelligently no matter how. Yet, most of the “intelligent process” approaches adopt a rational interpretation of human cognition, focussing on the mechanisms of rational human reasoning rather than on the whole spectrum of human cognitive processes. In fact, rational interpretation is where most of the *symbolic* AI techniques are grounded – e.g., automatic reasoning, planning, argumentation, logic programming –, and, by the way, also where most of the time is spent when teaching AI still today. For instance, the most widely used textbook for AI at some point defines AI as “the study of rational action” [9, page 366], there implicitly promoting a mostly-rational interpretation of the overall field.

However, non-rational, non-symbolic techniques basically terminated the first AI winter—e.g., proving that intelligence could exist without neither (rational, symbolic) reasoning [10] nor (symbolic) knowledge representation [11]. Also, the recent surge of the the so-called “AI Renaissance” has been mostly sparked by biologically-inspired yet non-rational, sub-symbolic techniques—such as deep learning ones.

As a result, nowadays most techniques in the AI field can be classified as either *symbolic* or *sub-/non-symbolic* ones. With some obvious degrees of approximation, this classification basically matches the aforementioned distinction between analytic vs. intuitive cognition in psychology: symbolic approaches are rational, have huge requirements in terms of resources, are slow, and fit for a limited range of problems; whereas sub-symbolic approaches are mostly non-rational, have limited requirements in terms of resources, and work fast in a huge number of scenarios—and we have no clear

understanding yet of why they do actually succeed.

So, in spite of the overwhelming successes of non- and sub-symbolic techniques, why does AI so much still revolves around symbolic techniques? Is there any motivation left for symbolic techniques *per se*, beyond the current wave of XAI calling for *integration* with sub-symbolic techniques [12]?

Overall, which are the reasons today for AI to be (also) rational, nourish its *esprit de géométrie*, cultivate analytic cognition?

### 3. Sharing

#### 3.1. Sharing is human

Our common understanding of science nowadays is grounded upon notions such as *reproducibility* and *refutability* [13], which basically reflect the fact that science for humans is a *social construct*. However, the need for sharing scientific results pre-dates our recent understanding of the conceptual foundation of science: even in the early years of Mediterranean science, some thousands years ago, scholars used to share their results by exchanging letters via sea mail.

This is because *sharing* is not just a means towards science: it is the foundation of human culture. Actually, when comparing humans against other living species (primates, in particular) it is quite apparent that *cumulative culture* is unique to people: “The remarkable ecological and demographic success of humanity is largely attributed to our capacity for cumulative culture, with knowledge and technology accumulating over time” [14]. Unlike most primates, humans have the innate disposition to *share knowledge*: as a result, hiding knowledge is mostly considered as an antisocial behaviour. By the way, this might also be one of the motivations behind open source technologies, the open access movement in science, the “recognise sharing as the default” principle for pervasive computing [15], as well as the hidden psychological cause of the common despise for “scholars hiding in their ivory towers”—knowledge bearers with no will to share.

So, the contemporary contempt for fact checking on the social networks (after the issues taken against the use of fake news in political competitions) is in some sense the social counterpart of the scientific concern for a shared process for stating scientific evidence and results— particularly emphasised these days by the worldwide run for COVID-19 therapies and vaccines. That is: at every level, human knowledge is shared, and the same holds for the processes to assess knowledge itself.

#### 3.2. Sharing is rational

Humans are *hyper-social* animals: “We never think alone” [16]. We may quite safely assume that our rational capability as humans evolved predominantly *after* our ability to interact socially—or, at most, they *co-evolved*.

Yet, in order to support sharing of knowledge, humans have invented and developed cognitive and physical *tools* – language, writing, books, the Web – which allow for rational *representation* of knowledge. This makes sharing effective, and extend its span over time and space, so that human can share knowledge with others even though they are not in the same place at the same time. Rational tools for knowledge sharing make it possible for human culture to be *systematic* and cumulative. Overall, sharing knowledge among humans definitely comes along with symbolic representation and reasoning.

So, though there is intelligence without representation [11] and reason [10], there is no (systematic) sharing without reason and representation. Sharing requires reason because it requires a form of

*conversation* between humans, and conversation requires *mind reading* – as the ability to understand motivations, beliefs, goals of others –, or, more generally, a *theory of mind* [17].

So, maybe the heart knows things that the reason does not, and maybe it knows more: however, what if *that sort* of knowledge cannot be (easily) shared? In other terms, how can we share the results of non-rational, intuitive cognitive processes? How can we share what we “humanly know”, without being hampered by those “*vues lentes, dures et inflexibles*” that sharing asks for in order to work with some predictability?

Humans actually have the answer: they do (try to) *explain*. When they need to share the results of their intuition, they try to (make up and) share a *rational explanation* with other humans around them: so, explanation is how humans try to make intuitive and analytical cognition *match*.

Yet, what is an explanation? Or, at least, where can we find a suitable acceptation of the notion of explanation that could help us defining it as the rational act of an explainer – whichever his/her nature – trying to share his/her knowledge?

## 4. Explanation

### 4.1. Explanation everywhere

Since when *explainability* has become a hot topic in AI research, the very notion of *explanation* – and accessory notions such as *interpretation* and *understandability* alongside – has become the core of many research efforts, and undergone a constant flow of diverse and (sometimes) even extravagant definitions. This is likely to be due to the two problems mentioned at the beginning of this paper: the still weak settling of a commonly-acknowledged scientific method for CS and AI, and the pervasiveness of the common meaning of the terms.

For instance, when the GDPR [18] starts to recognise “the citizens’ right to explanation” [19], it encompasses in the same acceptation of the term ‘explanation’ both the *explainer* act aimed at making things understandable, and the *explainee* act aimed at understanding things. Whereas this might be a commonly-acknowledged acceptation of the term – and a very important stand for EU citizens, as well – at the same time it might be not the best choice for a well-grounded and technically-useful scientific definition of the term.

Also, as noticed in [4], many definitions in the literature do not really keep the two notions of explanation and interpretation well distinct and separated—as we probably do in our common use of the language. Whereas the validity of the related scientific results is most typically not affected by that, the scientific value of those definitions *per se* is somehow limited.

### 4.2. Noetics & semiotics

It should not come as a surprise that a useful contribution towards a well-founded definition of the notion of explanation comes from the theory of teaching—from the *teaching of mathematics*, specifically. There, given the fully-abstract nature of mathematical concepts, the understanding of mathematics can be labelled in terms of *noetics* – as the *conceptual acquisition* of an object – as opposed to *semiotics*—as the acquisition of a *representation built out of signs* [20]. Yet, noetics cannot happen without semiotics: no learning of mathematical concepts is actually possible if not via sign manipulation—since in math you do not learn to actually manipulate concepts, but the corresponding semiotic representations, instead.

Many different *semiotic representations* are possible for the same concept: a (straight) line could be represented in terms of common language, or, with a drawing, or, by means of the algebraic language—

by switching through different *semiotic registers*. Moving from a representation of a concept in a given register of semiotics to another in a different register (e.g., moving from a geometric representation to the analytic representation of the same curve) is a *transformation of conversion*. Instead, changing representation within the same register of semiotics (e.g., switching to a different analytic expression of a curve) is a *transformation of treatment*.

Changing semiotic register is what is used in math to *explain*. For instance, converting between polar and Cartesian coordinates in the representation of a curve (transformation of treatment), or, switching from the analytic to the geometric definition of a curve (transformation of conversion), are examples of manipulation of the semiotic representation that are usually adopted in order to give a grasp of abstract mathematical objects—to actually explain them. So, explanation could then be generally intended as an activity of *transformation of semiotic register*—of either conversion or treatment, or both.

Explanation is then an *operation by the explainer* that does neither require nor ensure *per se* any understanding by the explainee, yet can be intentionally geared in that direction once something of the explainee is known by the explainer (e.g., by mind reading [17])—as it happens in teaching, by the way. Limiting the definition of explanation to the act of an explainer only seems to better match acceptations of the term like those implicitly used in human textbooks: there in fact explanations are not explicitly directed towards any specific explainee, and just assume a general level of linguistic and cultural competence by the generic reader.

As an aside, the notion of *explanation as a transformation of semiotic register* also looks more easily in accordance with the original etymology of the word itself—from Latin *explanare*, which also means ‘spread’, ‘unfold’, ‘straighten out’. According to that, explanation can be seen as an *activity of symbolic representation and transformation by the explainer*, aimed at making the *subjective* activity of interpretation by the explainee easier—as in [4].

Once the notion of explanation has been framed, now: who is the explainer, who the explainee? Or, more specifically: can we proceed beyond the simplistic hypothesis that software systems / agents are the explainers, and explainees are just humans?

### 4.3. Explanation for humans

Almost all works on explainability have one aim in common: making *intelligent systems understandable to humans*. As one can observe from many surveys – e.g., [21, 22, 23] –, explainers are typically software components (such as agents in multi-agent systems), explainees are intended to be humans. Thus, explanations are mostly intended to work as *one-direction* tools: from agents to humans. First, agents do something bright and complex and unintelligible for humans; then, in order to make it acceptable for humans, agents are expected to explain themselves to humans in some way.

Before we proceed further, it is worth pointing out how this assumption is already quite unsatisfactory *as it is*. In fact, the ability of software systems to effectively interact with humans is such a problematic issue in general that we are likely to devote some of the forthcoming decades to the full development of *conversational informatics* [24] as a multi-disciplinary field aimed at capturing all the diverse dimensions and nuances of human conversational skills in terms of *conversational agents*. Since our goal there is basically to *level agents up* to the human conversational abilities, it looks somehow pretentious today to assume that an explanation from an agent to a human is something that we can easily handle already in general.

Also, as much as this may look at a first glance as obvious and natural, it is not the only possibility: an agent providing a human with an explanation is not the only way in which explanation could be useful and work. This is apparent in teaching theory, of course: both explainers and explainees there

are humans, even when they are supported by artificial systems of any sorts—as in the case of tools for distance learning.

Furthermore, by exploring in principle every possible direction that explanation could follow in *socio-technical systems* (STS) – that is, artificial systems where both *humans* and *artificial components* play the role of system components [25] –, humans explaining to agents would require agents to exhibit just the same level of conversational abilities that the agent-to-human explanation mandates for—not an easy task still nowadays.

What about software systems, instead? Is there any reasonable way in which software components could use some notion of explanation to interact with each other, and not just with some human? Or, by changing the viewpoint over software systems, is there any meaningful way to make explanation a general tool for intelligent systems engineers, used to make interaction and communication between intelligent components more expressive and effective?

#### 4.4. Agents communicating in MAS

Agents and multi-agent systems (MAS) are the richest providers of abstractions, technologies, and methodologies for complex intelligent systems [26]. Agent interaction within MAS at its most fundamental level typically exploits *agent communication languages* (ACL) [27], which are one of the oldest standard in the MAS field [28], and are shaped around Searle’s theory of human communication based on speech acts [29]. ACL generally set the stage for agent interoperability in MAS, by providing communicating agents with shared syntax and ontology.

This is particularly relevant in the case of *open* and *heterogeneous* MAS, where ACL standards allow in principle diverse agents to effectively interact and communicate with each other overcoming their difference in terms of model, architecture, and technology. For instance, within an intelligent system used in the legal domain, 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. In spite of their fundamental heterogeneity, ACL would make communication possible between the two very different sorts of agents.

More generally, *agreement technologies* [30] – e.g., semantic web, coordination models and languages, norms, e-Institutions, dialogue, negotiation, argumentation, ... – go beyond the mere level of agent communication. For instance, agreement technologies are essential in the case of case of *computable law*, described in [31] as a “multi-agent system based on argumentation, dialogue, and conversation”. There, agents do not just communicate with each other: when they make statements, agents should provide arguments to support them, to be understood and possibly accepted.

Yet, agreement technologies do not directly addresses in general issues such as explainability, interpretability, transparency, and understandability. For instance, in the field of computable law, some highly-heterogenous agents in a decision-support system within a legal process might be asked to go beyond just arguing and supporting their suggestions and decision proposals—so, beyond explaining themselves to humans. Instead, they could be required to interact with each other – and, more specifically, *to understand each other* – in order to (first of all) cooperatively build a common, well-motivated proposal – or, a well-reasoned list of possible alternatives –, to be (then) presented to human decision makers in a potentially-understandable way.

This, of course, would require that agents, whatever their nature – symbolic vs. sub symbolic vs. hybrid – are capable not only of communicating via ACL and interacting with each other via agreement technologies, but also of *representing* their cognitive process and achievements in a *rational form* that could be effectively *shared* not just with humans, but also, first and foremost, *with other agents*.

#### 4.5. Agent sharing in MAS

Broadly speaking, analogy is not always the best way to proceed in multi- and inter-disciplinary contexts, and should be feared in the scientific practice for how easily it may mislead researchers. Yet, when dealing with intelligent agents and MAS – possibly for the way in which agent reasoning agent action are modelled after their human equivalent, and MAS are modelled after human societies –, this has worked quite well and quite often in the last few decades: *intentional stance* in agent reasoning [32], *speech acts* in agent communication [29], *activity theory* for agent coordination [33] are just some examples.

So, in the same way as humans share knowledge and cognition in cooperative contexts, agent cooperation in open and heterogeneous MAS seemingly mandates for *sharing among agents*, since (i) agents are generally-opaque entities, (ii) trivial agent semantics à la ARCOL (where agents are assumed to be sincere—that is, they say only what they do believe [27]) are not practical in real-world open MAS, (iii) notions such as trust and reputation are hardly effective in open and dynamics systems, and do not scale up easily with system size. Sharing not just knowledge, but also their own cognitive processes, would in principle allow agents to understand each other at a deepest level possible in open and heterogeneous MAS, by making agent intelligence as *transparent* as possible to other agents.

As a means to make sharing effective, explanation – as a mere act of the explainer, and in the general acceptance of transformation of semiotic register – has the potential to work as a powerful enhancer of agent-to-agent communication, as well as an expressive extension to agreement technologies.

#### 4.6. Explanation for agents

Along this line, one may picture agent-based intelligent systems as made of *explanation-ready* agents, as agents of any nature and sort equipped with their own specific *rational explanation* capability. So, in principle, any intelligent agent would be built with the ability of providing a *rational*, sharable *representation* of their own specific cognitive process and results, as well as of *manipulating* such a representation in order to build one or more *explanations*.

If an explanation is a transformation of the semiotic register, the expressive power of an explanation-ready agent could be basically measured by its ability to deal with *different semiotic registers* in its domain of discourse, and to switch between *diverse representations* within each of them—thus producing explanations in terms of both transformations of conversions and transformations of treatment, respectively.

Architecturally, this would imply that intelligent agents of any sort should 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: and this should hold for both *intuitive* agents – as those adopting sub-/non-symbolic techniques – and *rational* ones—as those built around symbolic techniques. Also, intelligent 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 transformation of the semiotic register—so to make them able to provide one or more *explanations* in symbolic form.



## 5. Conclusions

After pointing out that the issue of understanding and modelling (human) intelligence dates far before the birth of AI, we start from the observation that the classic philosophical distinction between rationally and intuition in human cognition roughly matches the well-known dichotomy of contemporary AI techniques—that is, symbolic vs. sub/non-symbolic, respectively. We then note that the overwhelming success of sub-symbolic techniques (such as deep learning ones) has not moved AI really far from its symbolic core. Yet, symbolic techniques nowadays are mostly seen as supporting *actually-working* AI technologies such as deep learning in real-world intelligent systems. This is particularly evident in the XAI field, where symbolic techniques are exploited to complement sub-symbolic components and provide them with human-understandable explanations.

Thus, in this paper we first advocate that the very notion of *explanation* needs to be defined more precisely than it currently is in the literature, specifically as an *explanator's act*—so, as a premise for the possible explainee's understanding, not including it. Then, we suggest that explanation should not be seen as a mere add-on for intelligent systems, and should instead work as an *essential tool* for any *intelligent component*—in particular, *agents* in multi-agent systems. Also, we argue that intelligent agents should be able to *explicitly represent* their cognitive process and its results, and *manipulate* those representations, so that *rational* explanation would properly complement their ability to *reason* and *communicate*. As a result, intelligent agents should be able to *explain* themselves first of all *to other agents*, not just towards humans. In this context, *symbolic* techniques are to be used for explanations, to enable agents to represent and manipulate their own cognitive processes and their results, and to understand explanations from other agents.

Overall, we advocate that both explanation and symbolic techniques should play the role of *first-class citizens* in both *agent modelling* and *intelligent systems engineering*.

## Acknowledgments

This work has been partially supported by the H2020 Project “AI4EU” (G.A. 825619).

I would like to thank my collaborators Roberta Calegari and Giovanni Ciatto for the many discussions we had in the last few years around the main topics of this paper. Also, I am grateful to Giuseppe Pisano and Giuseppe Vizzari for basically forcing me to write this paper – for totally different reasons –, and to the anonymous reviewers, too, who forgivingly accepted quite a preliminary version of this paper.

## References

- [1] R. Connolly, Why computing belongs within the social sciences, *Communications of the ACM* 63 (2020) 54–59. doi:10.1145/3383444.
- [2] G. Johnson, The world: In silica fertilization; all science is computer science, *The New York Times* (2001). URL: <https://www.nytimes.com/2001/03/25/weekinreview/the-world-in-silica-fertilization-all-science-is-computer-science.html>.
- [3] D. Gunning, Explainable artificial intelligence (XAI), Funding Program DARPA-BAA-16-53, Defense Advanced Research Projects Agency (DARPA), 2016. URL: <http://www.darpa.mil/program/explainable-artificial-intelligence>.
- [4] G. Ciatto, M. I. Schumacher, A. Omicini, D. Calvaresi, Agent-based explanations in AI: Towards an abstract framework, in: D. Calvaresi, A. Najjar, M. Winikoff, K. Främling (Eds.), *Explainable,*

- Transparent Autonomous Agents and Multi-Agent Systems, volume 12175 of *Lecture Notes in Computer Science*, Springer, Cham, 2020, pp. 3–20. doi:10.1007/978-3-030-51924-7\_1.
- [5] P. J. Denning, Ubiquity symposium ‘What is computation?’: Opening statement, *Ubiquity* 2010 (2010) 1:1–1:11. doi:10.1145/1880066.1880067.
- [6] B. Pascal, *Pensées*, Guillaume Desprez, Paris, France, 1669.
- [7] B. F. Skinner, Cognitive science and behaviourism, *British Journal of Psychology* 76 (1985) 291–301. doi:10.1111/j.2044-8295.1985.tb01953.x.
- [8] R. E. Patterson, R. G. Eggleston, Intuitive cognition, *Journal of Cognitive Engineering and Decision Making* 11 (2017) 5–22. doi:10.1177/1555343416686476.
- [9] S. J. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed., Pearson Education Limited, Harlow, Essex, UK, 2010.
- [10] R. A. Brooks, Intelligence without reason, in: J. Mylopoulos, R. Reiter (Eds.), *12th International Joint Conference on Artificial Intelligence (IJCAI 1991)*, volume 1, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1991, pp. 569–595. URL: <http://dl.acm.org/citation.cfm?id=1631258>.
- [11] R. A. Brooks, Intelligence without representation, *Artificial Intelligence* 47 (1991) 139–159. doi:10.1016/0004-3702(91)90053-M.
- [12] R. Calegari, G. Ciatto, A. Omicini, On the integration of symbolic and sub-symbolic techniques for XAI: A survey, *Intelligenza Artificiale* 14 (2020) 7–32. doi:10.3233/IA-190036.
- [13] K. R. Popper, *The Logic of Scientific Discovery*, Routledge, 2002. URL: <https://www.routledge.com/The-Logic-of-Scientific-Discovery/Popper/p/book/9780415278447>, 1st English Edition:1959.
- [14] L. G. Dean, R. L. Kendal, S. J. Schapiro, B. Thierry, K. N. Laland, Identification of the social and cognitive processes underlying human cumulative culture, *Science* 335 (2012) 1114–1118. doi:10.1126/science.1213969.
- [15] R. Grimm, J. Davis, E. Lemar, A. Macbeth, S. Swanson, T. Anderson, B. Bershad, G. Borriello, S. Gribble, D. Wetherall, System support for pervasive applications, *ACM Transactions on Computer Systems* 22 (2004) 421–486. URL: <http://portal.acm.org/citation.cfm?id=1035582.1035584>. doi:10.1145/1035582.1035584.
- [16] S. Sloman, P. Fernbach, *The Knowledge Illusion: Why We Never Think Alone*, Riverhead Books, New York, NY, USA, 2017. URL: <https://www.penguinrandomhouse.com/books/533524/the-knowledge-illusion-by-steven-sloman-and-philip-fernbach/>.
- [17] H. Wimmer, J. Perner, Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children’s understanding of deception, *Cognition* 13 (1983) 103–128. doi:10.1016/0010-0277(83)90004-5.
- [18] P. Voigt, A. von dem Bussche, *The EU General Data Protection Regulation (GDPR). A Practical Guide*, Springer, 2017. doi:10.1007/978-3-319-57959-7.
- [19] B. Goodman, S. Flaxman, European Union regulations on algorithmic decision-making and a “right to explanation”, *AI Magazine* 38 (2017) 50–57. doi:10.1609/aimag.v38i3.2741.
- [20] B. D’Amore, Noetica e semiotica nell’apprendimento della matematica, in: A. R. Laura, F. Eleonora, M. Antonella, P. Rosa (Eds.), *Insegnare la matematica nella scuola di tutti e di ciascuno*, Ghisetti & Corvi Editore, Milano, Italy, 2005. URL: <http://www.dm.unibo.it/rsddm/it/articoli/damore/676noeticaesemioticaBari.pdf>.
- [21] A. Rosenfeld, A. Richardson, Explainability in human-agent systems, *Autonomous Agents and Multi-Agent Systems* 33 (2019) 673–705. doi:10.1007/s10458-019-09408-y.
- [22] S. Anjomshoae, A. Najjar, D. Calvaresi, K. Främbling, Explainable agents and robots: Results from a systematic literature review, in: *18<sup>th</sup> International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS’19)*, IFAAMAS, 2019, pp. 1078–1088. URL: <https://dl.acm.org/>

- doi/10.5555/3306127.3331806.
- [23] R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, F. Giannotti, A survey of methods for explaining black box models, *ACM Computing Surveys* 51 (2019) 1–42. doi:10.1145/3236009.
  - [24] T. Nishida, A. Nakazawa, Y. Ohmoto, Y. Mohammad, *Conversational Informatics. A Data-Intensive Approach with Emphasis on Nonverbal Communication*, Springer Japan, Tokyo, 2014. URL: <http://link.springer.com/10.1007/978-4-431-55040-2>. doi:10.1007/978-4-431-55040-2.
  - [25] B. Whitworth, Socio-technical systems, in: C. Ghaou (Ed.), *Encyclopedia of Human Computer Interaction*, IGI Global, 2006, pp. 533–541. doi:10.4018/978-1-59140-562-7.ch079.
  - [26] F. Zambonelli, A. Omicini, Challenges and research directions in agent-oriented software engineering, *Autonomous Agents and Multi-Agent Systems* 9 (2004) 253–283. doi:10.1023/B:AGNT.0000038028.66672.1e, Special Issue: Challenges for Agent-Based Computing.
  - [27] M. P. Singh, Agent communication languages: Rethinking the principles, *Computer* 31 (1998) 40–47. doi:10.1109/2.735849.
  - [28] FIPA ACL, Agent Communication Language Specifications, Foundation for Intelligent Physical Agents (FIPA), 2002. URL: <http://www.fipa.org/repository/aclspecs.html>.
  - [29] J. Searle, *Speech Acts: An Essay in the Philosophy of Language*, Cambridge University Press, 1969.
  - [30] S. Ossowski, *Agreement Technologies*, volume 8 of *Law, Governance and Technology Series*, Springer Netherlands, 2012. doi:10.1007/978-94-007-5583-3.
  - [31] R. Calegari, A. Omicini, G. Sartor, Computable law as argumentation-based MAS, in: R. Calegari, G. Ciatto, E. Denti, A. Omicini, G. Sartor (Eds.), *WOA 2020 – 21st Workshop “From Objects to Agents”*, volume 2706 of *CEUR Workshop Proceedings*, Sun SITE Central Europe, RWTH Aachen University, Aachen, Ger, 2020, pp. 54–68. URL: <http://ceur-ws.org/Vol-2706/paper10.pdf>, Bologna, Italy, 14–16 September.
  - [32] D. Dennett, Intentional systems, *Journal of Philosophy* 68 (1971) 87–106.
  - [33] L. S. Vygotskiĭ, *Mind in Society: Development of Higher Psychological Processes*, Harvard University Press, Cambridge, MA, USA, 1978. URL: <http://www.hup.harvard.edu/catalog/VYGMIX.html>.