# Towards Engineering Explainable Autonomous Systems

Michael Winikoff[1 [0000−0002−5545−7003]]

Victoria University of Wellington, New Zealand
michael.winikoff@vuw.ac.nz

**Abstract.** Explanation is important to supporting appropriate levels of trust in autonomous systems. However, work in XAI (eXplainable AI) is focused on explanation of single system components, such as a machine learning algorithm or decision-making module. This paper: (1) argues that we need to develop ways to engineer explainable systems consisting of multiple components, and identifies this as a challenge for the community; (2) proposes an approach for explaining multi-component autonomous systems; (3) identifies integration issues that need to be addressed to make this vision a reality; and (4) poses a number of research challenges and questions that need to be addressed.

**Keywords:** Explanation · XAI · Engineering · Integration · Trust · Autonomous Systems

## 1 Introduction

*Scenario: The video footage was clear: the self-driving car had collided with the pedestrian. But why? The investigator queried the system: "why didn't you stop?". After a pause, the response came back: "I could not stop (or slow down significantly) because there was a car close behind me, and I could not perform an alternative manoeuvre". The investigator selected the second part of the response and queried it, eventually determining that the system had prioritised avoiding collisions with cars behind and beside it, due to a failure by the image processing module to classify the pedestrian as human, because they were obscured by packages they were carrying.*

Autonomous systems need to be explainable for a range of reasons [13, 38, 21, 35]: to be understandable [34], to help establish appropriate levels of trust [16, 29, 13], and to be accountable [7]. Explanations are used for a range of purposes [14, 3, 25], by a range of different stakeholders [20]. Stakeholders vary in their level of expertise and familiarity with the domain and with the system, and in their goals. For example, a software engineer trying to debug a system [37] has different needs than a lawyer seeking to construct a civil liability case in relation to a malfunctioning autonomous system [4].

This paper's contribution is to identify and articulate the challenge of engineering explainable multi-component systems, and to propose an approach to addressing this challenge, along with research challenges that we pose.

There is a whole body of work on explainable AI (XAI) [2]: techniques that allow explanations to be provided for the behaviour of AI modules. For example, why did a machine learning system recommend to decline a given loan application? However, XAI is typically focused on techniques for explaining individual (often machine learning) components, not whole systems [2, §5.3]. This paper argues that XAI is necessary, but not sufficient, and that we need to extend explanation from single components, to whole multi-component systems, and also consider engineering issues. This will allow realistic autonomous systems (with multiple components) to be explainable. Rodriguez *et al.* [31, 30] also call for an engineering focus on XAI, and they also identify, but do not address, the issue of explaining multi-component systems.

We next present and justify a proposed architecture for explainable autonomous systems (§2), including consideration of the sorts of questions that can be asked (§2.1), and the sorts of answers that can be given (§2.2). We then consider how the components in the proposed architecture integrate and interact (§3), and close (§4) with discussion of some broader issues and a summary of the research challenges that need to be addressed to enable the engineering of explainable autonomous systems.

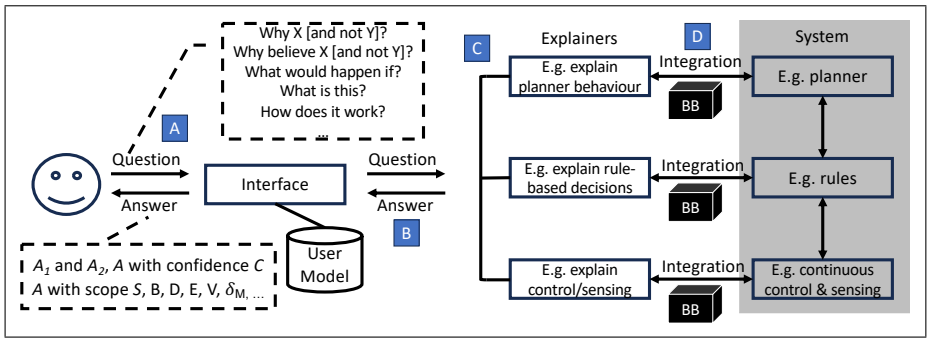


**Fig. 1.** Architecture (B=Belief, D=Desire, E=Emotion, V=Valuing, $\delta_M$=user model change, BB = Black Box)

## 2 Architecture

Figure 1 shows the proposed architecture for an explainable multi-component autonomous system. The explainable system can be naturally viewed as a multi-agent system. The original autonomous system (grey shaded on the right) consists of a number of components. Different systems might have a different number of components, but the architecture is generic: it can accommodate changes to the number of components in the autononous system by simply adding corresponding explainers. The rest of the architecture does not change.

In our example scenario we might have three components [12, §4.5]:

1. a component ("continuous control & sensing", bottom of Figure 1) that deals with continuous data from sensors and controlling actuators, and is able to recognise and classify obstacles and intersections, and control actuators to follow the road while managing speed to avoid collisions;
2. a component ("rules", middle of the Figure) that has rules for dealing with common situations, and is able to decide on the route and make decisions on what to do at an intersection, issue lane change commands in line with the navigation, and decide on a range of manoeuvres including pulling over or changing lane to avoid obstacles and moving out of the way of emergency vehicles; and
3. a component ("planner", top of Figure) that deals with unusual situations using planning based on the following principles: (i) avoiding damage to pedestrians (most important), (ii) avoiding damage to vehicles containing humans, and (iii) avoiding damage to unoccupied vehicles and property.

Each system component is associated with a corresponding explainer agent that applies appropriate XAI technique(s) to explain the behaviour of that component. For example, explaining a component that uses BDI (Belief-Desire-Intention) plans to generate behaviour could be done following the approach of Winikoff *et al.* [39]. This aspect of the architecture is required because each component might operate using quite different principles, and therefore require a different approach to generating explanations.

In order to generate explanations, a "black box" that captures relevant details from the system's execution is used. For example, to explain a component that uses rule-based reasoning, the black box would likely need to capture the facts that were believed to be true at a given point in time that were the basis for the choice of rule that was made.

The user ("☺" on the left) interacts with an interface. Having a single interface agent is required because we want to be able to hide the internal structure of the system from the user: as far as the user is concerned, there is a single autonomous system that is exhibiting behaviour that needs to be explained[1]. This interface maintains a model of the user (e.g. what does the user already know?) that is used to filter answers.

When the user asks a question, the interface agent passes on the question to one or more of the explainer agents. This process is iterative: as seen in the scenario, an answer may prompt the user to ask follow-up questions. In some situations (see §2.1) the system's reply might take the form of a question, with the user providing an answer. Argumentation [8, 1, 32, 22] can be a good approach for structuring the dialogue, and there has also been some work on dialogue for explaining BDI agent behaviour [10, 11].

The interactions within the system (e.g. between user and interface, and between interface and explainer agents) are done by asking questions, and receiving answers. We therefore need to define what sorts of questions can be asked, and

[1] Mualla *et al.* [27] and Calvaresi *et al.* [6] also propose a single interface agent, but differ in context, and lack the details of our architecture, as described in the remainder of this paper.

what constructs are used to provide answers. Although this may vary from system to system, for a given system we need to define this so the interface is able to indicate to the user what sorts of questions can be asked.

In defining the forms of questions and answers that are used we are guided by the extensive literature that aims to inform XAI researchers about relevant work in social sciences [24, 5]. Key findings include that the explanations that humans naturally use are *contrastive* (see §2.1), *selective* (i.e. incomplete, covering only (some) relevant factors), and *social* (presented relative to what the explainer believes the listener already knows).

There are a few design decisions that need to be made to realise an instance of our architecture. The first design decision is whether the system is designed so that the interface can determine from the question which explainer agent to send it to, or whether it simply sends it to all the explainer agents, and explainer agents can reply with a response of "don't ask me" (we return to this in §6). The second design decision is whether to have a single black box for the whole system, or to have each system component have its own black box. We propose to have a black box for each component, since each component's requirements for the black box might be quite different to other components.

### 2.1 Questions ...

In this section we consider what forms of questions the user should be able to ask. The most basic and obvious form is "Why?", for example "Why did you do this?" [18] or "Why did you *believe* this?" [37]. Additionally, when the system does something other than what the user was expecting, it can be useful to ask "Why did you *not* do *(something else)*?". However, in fact evidence from the social sciences shows that as humans we tend to use a more general form of *contrastive* questions [24]: "Why did you $X$ (fact) instead of $Y$ (foil)?" (although the foil is sometimes implied and omitted). Another (related) question form is the *counter-factual*: "What would happen if ...?" [28, Page 23][2]. Additionally, it may be useful to allow the system to answer more basic informational questions such as "what is this?", "how does this work?" and "how do I use this?", which Haynes *et al.* [15] define respectively as ontological, mechanistic, and operational, and provide patterns for how to engineer systems that can answer these sorts of questions.

Finally, in addition to posing a question, it may also be useful to provide some information on what is desired in a good answer. For instance, how complete does the answer need to be? What is the aim of the person asking the question - are they a novice trying to clarify why something slightly unexpected occurred, i.e. to learn, or are they an expert seeking to dig deep to ascribe blame for something that should not have occurred?

[2] This is different from a contrastive question in that the question includes a *difference* and the answer is the (alternative) outcome (the foil), whereas a contrastive question provides the actual outcome and (optionally) the alternative outcome, and gives the difference as the answer.

### 2.2 . . . and Answers

We now turn to consider how answers are formed: what concepts can be used in answers, and what other features are important to have to support our approach to explaining multi-component autonomous systems.

We begin with generic features. Firstly, answers need to be *decomposable*: when the system says "I could not stop because . . . and I could not perform an alternative manoeuvre", the user needs to be able to decompose the answer to select the second part and ask "But why couldn't you perform an alternative manoeuvre?". This can be realised by defining a number of general-purpose combining forms for answers (e.g. "and", "but"). Additionally, answers need to be able to include references to indicate where the explanation for something involving a module depends on another module. For example, a rule-based module chose to perform a certain action because of information that it had earlier received from a video-processing module. These links allow the interface agent to direct follow-up questions. Secondly, it can be desirable to be able to include levels of confidence in answers. For example, the system might indicate that it believed a particular key fact held with roughly 75% probability, or it might indicate that it was "very" confident of a particular classification of an image. Additionally, when explaining machine learning it has been argued that we need to have a way to indicate the "scope" in which the explanation is relevant and reliable [25].

Turning to consider the concepts that can be used to form answers, these obviously depend on the component: a system component that uses image processing will have different explanatory concepts than, say, one that uses BDI plans to realise goals. There are a range of approaches for explaining various machine learning approaches by providing examples. For instance [26] annotating an image to highlight the parts that were most influential in a particular decision, or indicating that had certain features been different, an alternative behaviour would have occurred (e.g. a higher salary would have led to the loan application being approved). Turning to cognitive agents, it has been argued [39] that the concepts of "belief" and "desire", as used in BDI architectures, match directly with the same concepts that humans naturally use to explain their behaviour in similar contexts, and therefore that these concepts, along with the additional concept of "valuing" [23] form a natural basis for explaining autonomous systems. Furthermore, Kaptein *et al.* [17] argue that in addition to beliefs and desires, explanations sometimes are in terms of emotions, for instance: "I called the hospital because I was **scared** (emotion) that I might have a hypo (too low blood sugar level)" [17, p.304, emphasis in original]. Another form of explanation is a correction to the human's mental model, for example, noting that a particular action has a pre-condition that the user appears to be unaware of, and which justifies the need for an action to establish the pre-condition [33].

Figure 1 summarises the range of constructs that might be used as explanations, including generic ones ("and", confidence, and scope), and component-specific building blocks (beliefs, desires, valuings, emotions, and human model changes).

## 3 Integration

We now turn to a number of issues that relate to the integration of the system.

The first issue corresponds to the arrow labelled C in Figure 1: how does the interface know where to send a given question? One option, which may be the simplest in some cases, is to simply send it to all explainers, if it is easy to have each explainer identify whether it is able to answer a given question. Another approach is to capture information that allows the interface to determine which explainer agent can answer a given question. This might be a static index. For example, if each system component has a distinct set of actions it can perform, then from the action one can determine which explainer agent to ask about it. Another option, if this is not the case, is to tag each action when it is performed with an indication of which system component generated that action. So, for example, when a rule is applied to decide to change lanes to avoid an obstacle, then the action would be tagged with "rules". However, there can be more complex cases (we return to these in §4). For instance, a decision to not stop might involve decision-making in all three system components: the continuous sensing component might detect an obstacle but, due to a car close behind, decide against stopping and invoke the rules component to consider whether to change lanes. In turn, the rules component might decide against a lane change and invoke the planner component to consider alternative approaches to avoid the obstacle. Finally, the planner component might decide that reducing speed and hitting the obstacle is the least bad option available.

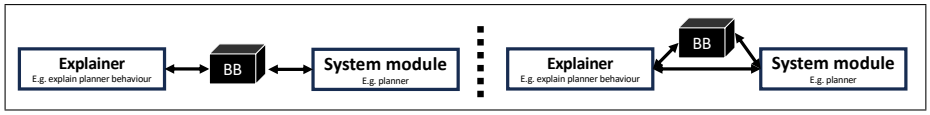


**Fig. 2.** Integration designs: indirect (via black box only) on left, direct on right

The second issue corresponds to the arrow labelled D in Figure 1: how does each explainer interact with its corresponding system component to generate explanations? One option, which may be the simplest in some cases, is to have the interaction be *solely* using the black box (left side of Figure 2): the system generates a trace in the black box as it runs, and the explainer constructs explanations using (only) this trace, without interacting with the system components. The advantage of this approach is that the system is cleanly separated from the explanation. However, it might require storing large amounts of data in the black box that could be reduced by allowing an explainer agent to interact with its corresponding system component. The alternative (right side of the Figure) is to also have direct interaction between the explainer and the module it is explaining. For example, to explain why a given image was not classified as having a pedestrian, the system might present the image (and perhaps some variants of it) to the component. Furthermore, in situations where the user has an im-

plicit foil, the system component may need to be used to generate the plausible alternatives for the situation.

## 4 Research Challenges

In order to be able to engineer explainable multi-component autonomous systems a range of research challenges need to be addressed. We group the challenges into three groups: those relating to the broader context, those relating to the research process itself, and specific research questions.

Our first group of challenges concerns the broader context of use. To use explanation in a particular context, for example a civil liability lawsuit, there may be a range of information that is relevant and useful other than explanations of behaviour. Buiten *et al.* [4] flag a range of such information. These include the development process, what steps were taken to mitigate risk, "second order explanations" (i.e. what explanations were previously provided to the user), the situations in which the system tends to fail, and how common are system failures. The research challenge is how to effectively collect and meaningfully present this information. Some of the information is about human social processes (e.g. development process), but some pose specific engineering research questions. For example: how can we identify and communicate the situations in which a system tends to fail, or the probability of system failures? Furthermore, this needs to be done in a way that cannot be manipulated.

The second group of challenges concern the endeavour of research: how to help the research community develop and evaluate solutions? More specifically: what scenarios would be useful? What test beds and benchmarks would help the research community to meaningfully and usefully evaluate ongoing work to assess and guide progress? Is there a role for standardisation? For competitions? These sorts of issues are ones where the XAI community could benefit from the experience and expertise of the EMAS community.

The third group of challenges is more specific research questions that relate to specific aspects of the architecture proposed in this paper. We believe these are perhaps the most useful to the EMAS community, and so close our paper with these questions, in the hope that they will spur further work to address them.

1. Regarding the first issue in §3: how can we manage the tagging of actions with the system component that is responsible?
   In complex cases, the decision to perform a certain action may have involved multiple system components. This raises a number of questions for how the architecture functions. Can the decision-making process be extended to keep a record that allows one module to be identified as the one making the final decision? Or instead can we just send questions to all explainer agents, and if so, how does an explainer agent determine whether the question is one that it can answer? And what should the interface do if it does not get exactly one meaningful answer? More broadly, what is there isn't a single responsible component (e.g. multiple agents interacting following a protocol)?

One additional challenge here is dealing with contrastive questions: if the question is "why did you do $X$?" then it can be possible to track which modules were involved in selecting action $X$. However, for the question "why did you *not* do $Y$?" we need to be able to identify how $Y$ might have been selected. This implies either that the black boxes capture information on possible alternatives (see questions 2 and 3 below), or that we need to have a way for explainer agents to explore what might have been with their corresponding module (see second part of question 2 below)

2. Regarding the second issue in §3: how do the explainer agents interact with the blackboxes? More specifically, what protocol is followed, and what interface (API) do the blackboxes need to provide?
   Furthermore, if explainer agents interact directly with the module that they are explaining, what protocol would be followed, and what interface (API) needs to be provided by the module?
3. What information is captured in the black box (e.g. [36])? How do we determine (in advance and/or at runtime) what needs to be captured, and how do we extend the system to do so?
   One issue is the management of storage: a complete trace of everything could be very large. There are approaches (e.g. [19]) for capturing system execution that manage to capture only key information and significantly reduce storage need, while permitting system execution to be "rewound" to any given point in the execution.
   Another issue is ensuring that the information captured allows explanations to be generated that include references where needed, i.e. where another module played a role in the observed behaviour, and follow-up questions relating to that factor (e.g. a belief held) should be directed to that module.
4. How can we ensure that provided explanations can be verified to be authentic and honest [9]? A key factor is how to ensure that the black box is tamper proof. A number of approaches are possible, depending on the system's operating environment and the amount of information captured. Approaches can include having a separate hardware component that provides the black box, and using encryption.
5. Are there situations where the interface agent may need to share information from the user model with explainer agents? For example, instead of generating explanations and then having the user agent filter them using the user model, it may be more efficient to share relevant parts of the user model with the explainer module, so it can not use them to guide to explanation generation process to avoid generating things that would subsequently be filtered out.

In addition to the above research questions, that are more focussed on EMAS-related topics, there are also XAI-oriented questions such as how to specify the parameters of a desired answer (e.g. level of detail, preferred explanatory factor types), how should confidence be expressed, how should the scope of validity of an explanation be indicated, and whether other question types or answer types are needed.

## References


1. Amgoud, L., Prade, H.: Using arguments for making and explaining decisions. Artif. Intell. **173**(3-4), 413–436 (2009). https://doi.org/10.1016/J.ARTINT.2008.11.006
2. Anjomshoae, S., Najjar, A., Calvaresi, D., Främling, K.: Explainable agents and robots: Results from a systematic literature review. In: Elkind, E., Veloso, M., Agmon, N., Taylor, M.E. (eds.) Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019. pp. 1078–1088 (2019), http://dl.acm.org/citation.cfm?id=3331806
3. Biran, O., McKeown, K.: Justification narratives for individual classifications. In: ICML 2014 AutoML Workshop. p. 7 (2014)
4. Buiten, M.C., Dennis, L.A., Schwammberger, M.: A vision on what explanations of autonomous systems are of interest to lawyers. In: Schneider, K., Dalpiaz, F., Horkoff, J. (eds.) 31st IEEE International Requirements Engineering Conference, RE 2023 - Workshops, Hannover, Germany. pp. 332–336. IEEE (2023). https://doi.org/10.1109/REW57809.2023.00062
5. Byrne, R.M.J.: Good explanations in explainable artificial intelligence (XAI): evidence from human explanatory reasoning. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China. pp. 6536–6544. ijcai.org (2023). https://doi.org/10.24963/ijcai.2023/733
6. Calvaresi, D., Ciatto, G., Najjar, A., Aydogan, R., van der Torre, L., Omicini, A., Schumacher, M.: Expectation: Personalized explainable artificial intelligence for decentralized agents with heterogeneous knowledge. In: Calvaresi, D., Najjar, A., Winikoff, M., Främling, K. (eds.) Third International Workshop on Explainable and Transparent AI and Multi-Agent Systems (EXTRAAMAS), Revised Selected Papers. LNCS, vol. 12688, pp. 331–343. Springer (2021). https://doi.org/10.1007/978-3-030-82017-6_20
7. Cranefield, S., Oren, N., Vasconcelos, W.W.: Accountability for practical reasoning agents. In: Lujak, M. (ed.) Agreement Technologies - 6th International Conference, AT 2018, Bergen, Norway, December 6-7, 2018, Revised Selected Papers. LNCS, vol. 11327, pp. 33–48. Springer (2018). https://doi.org/10.1007/978-3-030-17294-7_3
8. Cyras, K., Rago, A., Albini, E., Baroni, P., Toni, F.: Argumentative XAI: A survey. In: Zhou, Z. (ed.) Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021. pp. 4392–4399. ijcai.org (2021). https://doi.org/10.24963/IJCAI.2021/600
9. Dazeley, R., Vamplew, P., Foale, C., Young, C., Aryal, S., Cruz, F.: Levels of explainable artificial intelligence for human-aligned conversational explanations. Artif. Intell. **299**, 103525 (2021). https://doi.org/10.1016/j.artint.2021.103525
10. Dennis, L.A., Oren, N.: Explaining BDI agent behaviour through dialogue. In: Dignum, F., Lomuscio, A., Endriss, U., Nowé, A. (eds.) AAMAS '21: 20th International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, United Kingdom, May 3-7, 2021. pp. 429–437. ACM (2021). https://doi.org/10.5555/3463952.3464007
11. Dennis, L.A., Oren, N.: Explaining BDI agent behaviour through dialogue. Auton. Agents Multi Agent Syst. **36**(1), 29 (2022). https://doi.org/10.1007/S10458-022-09556-8

12. Fisher, M., Mascardi, V., Rozier, K.Y., Schlingloff, B., Winikoff, M., Yorke-Smith, N.: Towards a framework for certification of reliable autonomous systems. Auton. Agents Multi Agent Syst. **35**(1), 8 (2021). https://doi.org/10.1007/s10458-020-09487-2
13. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., Vayena, E.: AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines (Nov 2018). https://doi.org/10.1007/s11023-018-9482-5
14. Gregor, S., Benbasat, I.: Explanations from intelligent systems: Theoretical foundations and implications for practice. MIS Q. **23**(4), 497–530 (1999), http://misq.org/explanations-from-intelligent-systems-theoretical-foundations-and-implications-for-practice.html
15. Haynes, S.R., Cohen, M.A., Ritter, F.E.: Designs for explaining intelligent agents. Int. J. Hum. Comput. Stud. **67**(1), 90–110 (2009). https://doi.org/10.1016/j.ijhcs.2008.09.008
16. High-Level Expert Group on Artificial Intelligence: The assessment list for trustworthy artificial intelligence. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment (2020)
17. Kaptein, F., Broekens, J., Hindriks, K.V., Neerincx, M.A.: Evaluating cognitive and affective intelligent agent explanations in a long-term health-support application for children with type 1 diabetes. In: 8th International Conference on Affective Computing and Intelligent Interaction, ACII 2019, Cambridge, United Kingdom, September 3-6, 2019. pp. 1–7. IEEE (2019). https://doi.org/10.1109/ACII.2019.8925526
18. Koeman, V.J., Dennis, L.A., Webster, M., Fisher, M., Hindriks, K.V.: The "why did you do that?" button: Answering why-questions for end users of robotic systems. In: Dennis, L.A., Bordini, R.H., Lespérance, Y. (eds.) Engineering Multi-Agent Systems - 7th International Workshop, EMAS 2019, Montreal, QC, Canada, May 13-14, 2019, Revised Selected Papers. LNCS, vol. 12058, pp. 152–172. Springer (2019). https://doi.org/10.1007/978-3-030-51417-4_8
19. Koeman, V.J., Hindriks, K.V., Jonker, C.M.: Omniscient debugging for cognitive agent programs. In: Sierra, C. (ed.) Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017. pp. 265–272. ijcai.org (2017). https://doi.org/10.24963/IJCAI.2017/38
20. Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., Baum, K.: What do we want from explainable artificial intelligence (xai)? - A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artif. Intell. **296**, 103473 (2021). https://doi.org/10.1016/j.artint.2021.103473
21. Langley, P., Meadows, B., Sridharan, M., Choi, D.: Explainable agency for intelligent autonomous systems. In: Singh, S., Markovitch, S. (eds.) Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA. pp. 4762–4764. AAAI Press (2017), http://aaai.org/ocs/index.php/IAAI/IAAI17/paper/view/15046
22. Madumal, P., Miller, T., Vetere, F., Sonenberg, L.: Towards a grounded dialog model for explainable artificial intelligence. CoRR **abs/1806.08055** (2018), http://arxiv.org/abs/1806.08055

23. Malle, B.F.: How the Mind Explains Behavior. MIT Press (2004), ISBN: 9780262134453
24. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. **267**, 1–38 (2019). https://doi.org/10.1016/j.artint.2018.07.007
25. Mittelstadt, B.D., Russell, C., Wachter, S.: Explaining explanations in AI. In: danah boyd, Morgenstern, J.H. (eds.) Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*), Atlanta, GA, USA, January 29-31, 2019. pp. 279–288. ACM (2019). https://doi.org/10.1145/3287560.3287574
26. Montavon, G., Samek, W., Müller, K.R.: Methods for interpreting and understanding deep neural networks. Digital Signal Processing **73**, 1–15 (2018). https://doi.org/https://doi.org/10.1016/j.dsp.2017.10.011
27. Mualla, Y., Tchappi, I., Kampik, T., Najjar, A., Calvaresi, D., Abbas-Turki, A., Galland, S., Nicolle, C.: The quest of parsimonious XAI: A human-agent architecture for explanation formulation. Artif. Intell. **302**, 103573 (2022). https://doi.org/10.1016/j.artint.2021.103573
28. Mueller, S.T., Hoffman, R.R., Clancey, W.J., Emrey, A., Klein, G.: Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. CoRR **abs/1902.01876** (2019), http://arxiv.org/abs/1902.01876
29. Robinette, P., Li, W., Allen, R., Howard, A.M., Wagner, A.R.: Overtrust of robots in emergency evacuation scenarios. In: Bartneck, C., Nagai, Y., Paiva, A., Sabanovic, S. (eds.) The Eleventh ACM/IEEE International Conference on Human Robot Interation, HRI 2016, Christchurch, New Zealand, March 7-10, 2016. pp. 101–108. IEEE/ACM (2016). https://doi.org/10.1109/HRI.2016.7451740
30. Rodriguez, S., Thangarajah, J.: Explainable agents (XAg) by design (blue sky ideas track). In: Alechina, N., Dignum, V., Dastani, M., Sichman, J. (eds.) Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS). ACM (2024)
31. Rodriguez, S., Thangarajah, J., Davey, A.: Design patterns for explainable agents (XAg). In: Alechina, N., Dignum, V., Dastani, M., Sichman, J. (eds.) Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS). ACM (2024)
32. Sklar, E.I., Azhar, M.Q.: Explanation through argumentation. In: Imai, M., Norman, T., Sklar, E., Komatsu, T. (eds.) Proceedings of the 6th International Conference on Human-Agent Interaction, HAI 2018, Southampton, United Kingdom, December 15-18, 2018. pp. 277–285. ACM (2018). https://doi.org/10.1145/3284432.3284470
33. Sreedharan, S., Srivastava, S., Kambhampati, S.: Using state abstractions to compute personalized contrastive explanations for AI agent behavior. Artif. Intell. **301**, 103570 (2021). https://doi.org/10.1016/j.artint.2021.103570
34. Verhagen, R.S., Neerincx, M.A., Tielman, M.L.: A two-dimensional explanation framework to classify AI as incomprehensible, interpretable, or understandable. In: Calvaresi, D., Najjar, A., Winikoff, M., Främling, K. (eds.) Third International Workshop on Explainable and Transparent AI and Multi-Agent Systems (EXTRAAMAS), Revised Selected Papers. LNCS, vol. 12688, pp. 119–138. Springer (2021). https://doi.org/10.1007/978-3-030-82017-6_8
35. Winfield, A.F.T., Booth, S., Dennis, L.A., Egawa, T., Hastie, H.F., Jacobs, N., Muttram, R.I., Olszewska, J.I., Rajabiyazdi, F., Theodorou, A., Underwood, M.A., Wortham, R.H., Watson, E.N.: IEEE P7001: A proposed standard on transparency. Frontiers Robotics AI **8**, 665729 (2021), https://doi.org/10.3389/frobt.2021.665729

36. Winfield, A.F.T., van Maris, A., Salvini, P., Jirotka, M.: An ethical black box for social robots: a draft open standard. CoRR (2022), https://doi.org/10.48550/arXiv.2205.06564
37. Winikoff, M.: Debugging agent programs with why?: Questions. In: Larson, K., Winikoff, M., Das, S., Durfee, E.H. (eds.) Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2017, São Paulo, Brazil, May 8-12, 2017. pp. 251–259. ACM (2017), http://dl.acm.org/citation.cfm?id=3091166
38. Winikoff, M.: Towards trusting autonomous systems. In: Seghrouchni, A.E.F., Ricci, A., Son, T.C. (eds.) Engineering Multi-Agent Systems - 5th International Workshop, EMAS 2017, Sao Paulo, Brazil, May 8-9, 2017, Revised Selected Papers. LNCS, vol. 10738, pp. 3–20. Springer (2017). https://doi.org/10.1007/978-3-319-91899-0_1
39. Winikoff, M., Sidorenko, G., Dignum, V., Dignum, F.: Why bad coffee? explaining BDI agent behaviour with valuings. Artif. Intell. **300**, 103554 (2021). https://doi.org/10.1016/J.ARTINT.2021.103554