A Scoresheet for Explainable AI

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

John Thangarajah
RMIT University
Melbourne, Australia
john.thangarajah@rmit.edu.au

Sebastian Rodriguez
RMIT University
Melbourne, Australia
sebastian.rodriguez@rmit.edu.au

ABSTRACT

Explainability is important for the transparency of autonomous and intelligent systems and for helping to support the development of appropriate levels of trust. There has been considerable work on developing approaches for explaining systems and there are standards that specify requirements for transparency. However, there is a gap: the standards are too high-level and do not adequately specify requirements for *explainability*. This paper develops a scoresheet that can be used to specify explainability requirements or to assess the explainability aspects provided for particular applications. The scoresheet is developed by considering the requirements of a range of stakeholders and is applicable to Multiagent Systems as well as other AI technologies. We also provide guidance for how to use the scoresheet and illustrate its generality and usefulness by applying it to a range of applications.

KEYWORDS

Explainable AI; scoresheet; Specifying Explainability; Assessing Explainability; Explainable Agency; Goal-Driven XAI

ACM Reference Format:
Michael Winikoff, John Thangarajah, and Sebastian Rodriguez. 2025. A Scoresheet for Explainable AI. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 10 pages.

1 INTRODUCTION

It is important for autonomous and intelligent systems[1] to be explainable for a range of reasons. Providing explanations can be required by legislation either directly (e.g. GDPR[2]) or indirectly as a consequence of legislation [42]. Providing explanations can also play a crucial role in helping to make autonomous and intelligent systems socially acceptable [11], transparent [1, 39], understandable [38], accountable [10], and to help establish an appropriate level of trust [11, 22, 32, 34, 35, 40].

The importance of explainability has also been recognised by various standards. For instance, the Ethics Guidelines for Trustworthy AI[3] and subsequent Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment[4] consider explainability as one of a range of factors (e.g. human agency and oversight, accountability, societal & environmental well-being). IEEE P7001 [20] also considers explainability as part of transparency, and defines a number of requirements relating to explainability (e.g. that information is provided on how a system works in general, or that the system provides the ability to answer "why?" questions).

However, this work does not provide adequate guidance for the development and evaluation of the explainability of systems. The Ethics Guidelines for Trustworthy AI only poses questions that ask whether the decisions and outcomes can be understood and whether an explanation is provided, and the Assessment List has just two questions: "Did you explain the decision(s) of the AI system to the users?" and "Do you continuously survey the users if they understand the decision(s) of the AI system?". Similarly, IEEE P7001 only provides a few explainability requirements ("why?" and "what if?" questions, as well as global[5] explanation - see §2), and Hoffman *et al.* [16] assign each system only a single number (1-7).

Following IEEE P7001, we propose to provide this guidance in the form of a scoresheet[6]. The P7001 scoresheet focuses on transparency, and is complementary to our scoresheet: our scoresheet is specifically for explainability, and provides details, whereas P7001 has considerably less detail on explainability (see §3.1).

The scoresheet can be used in various ways with the most obvious being to evaluate the explainability of candidate systems. Used this way, the responses affect which system is chosen because the scoresheet captures that a crucial explainability aspect is lacking, or that another system provides it better.

Explanations are used by different people for different purposes [4, 7, 13, 21, 29], and therefore we develop our scoresheet by considering the explainability needs of different stakeholders (§2).

This paper makes a number of contributions. Firstly, we develop (and justify) a scoresheet[7] for explainability (§3). Secondly, we provide additional detailed guidance on *how* to complete the scoresheet (§4), including an additional checklist for global explanations. Thirdly, we demonstrate that the scoresheet is applicable to a range of systems (§5), showing that the scoresheet is *usable* and *generic*, as well as that it is *useful* (i.e. that it provides a useful summary).

[1] Terminology: Since we consider both autonomous systems and other systems that use a range of Artificial Intelligence techniques, we use "autonomous and intelligent systems", sometimes compressed to just "intelligent systems". We also avoid the term "model" (unless we are specifically talking about machine learning) in favour of "module". Finally, we use "behaviour" as shorthand for "behaviour or outcome" which encompasses the system taking action or providing some output (e.g. a classification or recommendation).
[2] https://data.consilium.europa.eu/doc/document/ST-5419-2016-INIT/en/pdf

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

[3] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[4] https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment
[5] A common distinction [18, 23] is between *local* explanations that relate to a specific execution (e.g. "why did you do this?"), and *global* explanations that are not about a specific execution, and hence more general, but necessarily less detailed.
[6] Our scoresheet does not use numbers, but it contains more than just checkboxes, so we use the term "scoresheet" for consistency with P7001.
[7] The scoresheet was developed iteratively (define, apply, revise); the version presented in §3 is the final one.

2 STAKEHOLDER EXPLAINABILITY NEEDS

IEEE P7001 [20, 39] defines five stakeholder groups: end users, wider public & bystanders, safety certifiers, incident/accident investigators, and lawyers & expert witnesses. They consider the range of forms of transparency that each requires. For instance, that end users might want to be able to get natural language answers to "why did you do that?" and "what would you do if . . . ?" questions. Or that safety certifiers need information on what steps were taken to verify and validate a system. They go on to propose a simple transparency scale for each of the five stakeholder groups. For example, for an end user, the levels can be summarised as: 0: "no transparency"; 1: information provided on how the system works in general (including, if relevant, on data used); 2: same as 1, but interactive; 3: ability to answer"why?" questions for specific cases; 4: ability to answer hypothetical "what if?" questions; and 5: provision of "continuous explanation . . . that adapts . . . based on the user's information needs and context".

Arya *et al.* [3] argue that different stakeholders require different sorts of explanations. They propose a taxonomy (and associated toolkit) that allows stakeholders to select an explanation method that suits their needs. Their context is narrower than ours (machine learning systems that learn from data). Their taxonomy considers factors such as the following. Are explanations (of data) given as particular features (e.g. income or level of debt), examples, or distributions? Do explanations explain individual cases or overall behaviour (local vs. global)? Is the explanation derived directly from the model used to make decisions, or from another (surrogate) model? Elements of their taxonomy are relevant to our scoresheet, and are incorporated in Section 3. These are: explanation of data (where relevant, using examples, distributions, features) vs. explanation of the model/module; the distinction between global and local explanations (which is also raised by other literature); and the distinction between an explanation being derived from the module itself, or from a surrogate[8].

Liao *et al.* [23] interviewed 20 UX and design practitioners from IBM to "identify gaps between the current XAI [eXplainable AI] algorithmic work and practices to create explainable AI products". Their focus is narrower than ours (explanations of machine learning for end users). One useful contribution of their work is their interview framework: they developed a bank of questions to ensure that the interviews covered a range of important aspects. In order to develop this, they identified a range of question types that can be addressed by current XAI methods, including both widely used questions (How, Why, Why not, What if) and less widely-used questions (how to be that, how to still be this; explained in Section 3). Their XAI question bank covered six topics: input (i.e. data used), outputs produced, performance (e.g. accuracy, precision, limitations), how (global), why & why-not (one topic), and a topic covering hypothetical questions (what if, how to be that, how to still be this).

The most directly relevant work to establishing stakeholder needs for explainability is the recent paper by Hoffman *et al.* [18] which seeks to establish what various stakeholders need by interviewing a range of stakeholders. One key point that they identify in their interviews is that the assumption that there are distinct, clearly distinguishable, stakeholders does not necessarily hold. Rather, they found that people had different roles, but that they adopted the viewpoints of different roles at different times, including roles other than their own. They highlighted the need for both global explanations (that are not too high-level, including holistic performance aspects such as biases, assumptions, bounding conditions and limitations) and local ones, and noted that it can be desirable to link them by having global explanations that refer to particular cases. They flagged the particular importance of edge cases in understanding how the system operates, and what are its limitations. More broadly, they identified the benefit of having access to the system development team and to (trusted) domain practitioners, and of having information about the system's context (e.g. what does it integrate with, how does it support users' goals) and the role of the company making the software, and trust in it, in a broader accountability and responsibility context.

3 AN XAI SCORESHEET

In this section we present the XAI scoresheet, focusing on *what* is included, and *why* it is included. Section 4 provides guidance on *how* to use the scoresheet.

The XAI scoresheet (Figure 1) has a number of sections that each collect different information. An initial section collects some **basic information**. Then there is a section that focuses on **veracity**, then **global explanations**, and finally a section focusing on a range of information relating to **local explanations**: features of explanations, the concepts used, the explanation types supported, and the level of automation.

Basic information: There are two pieces of basic information that the scoresheet collects. Firstly, whether the system's source code and (if relevant) training data is available. This is useful to know because access to code (and data) can help in understanding explanations, and in assessing the system's veracity (see below). However, this is of more use if there is access to the developers of the system, who can help to navigate the code (and data), and to (trusted) domain experts who can help to explain the context of use. Hoffman *et al.* [18] found that access to the system's developers and to trusted domain experts can be important to help understand the system's operation. In the case where the organisation assessing or using the system is also the one that is developing the system, then both these criteria would normally be met.

Veracity: An important basic requirement of explanations is that they actually correspond to the system's reasoning. An explanation system that invents explanations that do not reflect the actual reasons is clearly not useful, and could in fact mislead, and therefore be worse than not having an explanation at all. We therefore include in the scoresheet a high-level question to indicate the reliability[9] of explanations (Low/High[10]).

One approach to providing explanations with high reliability is to generate explanations directly from either the actual module used to

[8]Their taxonomy has two versions of this: for local explanations they distinguish between a self-explaining model and post-hoc explanations, whereas for global explanations they distinguish between directly interpretable models and post-hoc explanations such as a surrogate model, or a visualisation.

[9]We use "reliability" in the Cambridge dictionary sense of "the quality of being able to be trusted or believed because of working or behaving well".

[10]If the system does not provide local explanations then veracity is not applicable.

make decisions, or from a log that records what the system actually did and the factors considered (termed a "blackbox"[11] [39, 41]). This direct approach provides a high level of confidence that the explanation reflects the actual reasons.

An alternative approach is to construct explanations using an alternative proxy model. In this case it is possible for explanations to not correspond to the actual reasons, and so steps need to be taken to attempt to ensure alignment between the behaviour-generating model and the explanation model, and to assess the effectiveness of the alignment. For example, alignment can be attempted to be ensured by deriving the proxy model from the actual model by a systematic process or algorithm, and the alignment can be assessed by having a process of testing that evaluates for a range of system behaviours and explanations whether the generated explanation matches up with the real reasons for the system's behaviour. These real reasons can be identified by running the system on hypothetical scenarios to confirm that varying the reasons results in a change in behaviour. In some cases they may also be able to be identified by adding debugging probes to the system.

Regardless of the approach taken, it is important in order to be able to trust the explanations to know not just that explanations are reliable, but also *why* they are reliable, and so the scoresheet captures this information.

Global Explanations capture what sort of information is available about the system's *overall* functioning. One useful type of information is *how* the system works. Another is *how well* it works [18, 23]. This encompasses information on various limitations of the system (things it cannot do, including contexts in which it should not be used). It could also include information on the performance of the system (e.g. how accurate is it, how reliable, and in what scope/context can this level of performance be expected). These two questions can be addressed by providing a static document, or an interactive manual that allows the stakeholder to gain understanding of respectively how and how well the system functions [20]. Additionally, for systems where data plays an important role in decision-making, part of the answers to "how?" and "how well?" is information about data used (e.g. training data). This might usefully include the training data source, what steps were taken to ensure and/or assess its quality, information on distributions within the data (e.g. breakdown by demographic factors), how it was processed, and what limitations or assumptions exist. For example, a data set of facial photos from a particular country reflects that country's demographics, and may not be appropriate to use in a country with significantly different demographics.

Local Explanations are, unsurprisingly, a key part of the XAI scoresheet that capture a range of information. We begin with general information about the features of explanations that are generated. Firstly, since different people need different explanations, it can be useful to be able to generate different explanations for different people ("individually customised"). In order to do this it can be useful to be able 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?*" [41]. Secondly, since explanations can be quite complex and large, it can be useful to make them *interactive* [17]. For instance, provide a partial high-level answer to a question and allow the user to interactively get more information where needed. Finally, when an explanation is given, it can be useful for it to include indications of *confidence* [41] (e.g. that the system's decision was based on a particular belief that was held with a certain level of confidence), and of the *scope* [29] (the extent to which the explanation generalises, e.g. in a loan decision application that the key factor for a certain decision was the applicant's salary, but that this holds only as long as certain other factors are held within a certain range).

Next the XAI scoresheet records information on what *concepts* are used in explanations. This is useful to capture because it indicates at a high level what explanations look like. Furthermore, it has been argued [44] that since humans explain their behaviour in terms of particular concepts such as beliefs, goals, and valuings [24], using these same concept to explain autonomous systems

XAI scoresheet for _____

□ System source code is available
Is training data used available? Yes / No / Not Applicable
□ There is access to the system's developers
□ There is access to trusted domain experts

Veracity:
How reliable are explanations? Not Applicable / Low / High
What steps are taken to ensure explanation reliability?

* * *

Global Explanations: Has information been provided on:
□ *How* does the system work?
□ *How well* does it work?
(See checklist - Figure 2)

* * *

Local Explanations: Explanations ...
□ ...can be **individually customised**
□ ...are **interactive**
□ ...include an indication of **confidence**
□ ...include an indication of **scope of generalisation**

What **Concepts** are used in explanations?
□ Examples □ Features □ Beliefs □ Events/Percepts
□ Goals □ Actions □ Preferences □ Values
□ Other: _____

What forms of **Explanation Types** are provided?
Factual/Past: □ Did? □ Why? □ Why not? □ Contrastive
Future-looking: □ Will? □ Why? □ Why not? □ Contrastive
Hypothetical: □ What-if? □ How to be? □ How to still be?
Other: _____

Is **explanation generation** from questions?
□ Fully automated □ Partially automated □ Manual

Figure 1: XAI scoresheet. Notation: alternatives ("pick one") are separated by "/" whereas multiple options ("select all applicable") are indicated with "□".

[11] As in an aircraft blackbox

can make explanations more accessible and understandable. The XAI scoresheet lists examples and features (since a range of explanation mechanisms use these), as well as a range of concepts for autonomous systems (see [31, Chapter 2]) and values [25, 26].

Next, the scoresheet captures what sort of *explanation types* the explanation generation system is able to generate. This is captured in terms of the sorts of questions that the system can answer. The most basic form of question is factual: e.g. did something happen? We also consider the possibility of future-looking factual expectations: is something expected to happen? For example, a system that does some form of planning (reactive or first principles planning) may be able to provide information about what it did (in the past), and what it intends to do or expects to be the case (in the future).

Perhaps the most common type of explanation considered in the literature is answering "Why?" questions. As for factual questions, explanations can potentially refer to both the past (e.g. a certain course of action was selected because of past information or beliefs) and the future (e.g. a certain action was performed in order to achieve a certain situation in the future). In addition to being able to pose "Why?" questions, it can also be useful to be able to ask "Why not?", and it has been argued [27] that as humans we naturally tend to ask *contrastive* questions ("Why did you do X rather than Y?", although sometimes the "rather than" part is implicit).

Finally, the literature identified a range of forms of *hypothetical* question types that can be useful. For instance, Hoffman *et al.* [17] note that contrastive and counterfactual explanations play a role in supporting a range of user goals. Liao *et al.* [23] identify a number of such question types: "what-if?" (what would happen in a different situation?), "how to be?" (how to change inputs to achieve a certain outcome), and "how to still be?" (what changes to inputs would leave the outcome unchanged).

The last part of the XAI scoresheet concerns automation. Ideally, when the user asks a question, the system generates the explanation. However, it is also possible to have the system support explanation construction by the user, or even provide enough information so the user can construct an explanation manually (e.g. see §5.3). However, a manual explanation construction process is clearly less desirable than having the system generate the explanation.

3.1 Comparing with IEEE P7001

Having explained what we have included in our XAI scoresheet and why, we now briefly compare it to the IEEE P7001 transparency scoresheet [20]. Like us, IEEE P7001 proposed a scoresheet in order to help bridge the gap between high-level statements about desirable properties of systems and actionable metrics. However, there are a number of significant differences. The most significant difference is that P7001 is broader in scope, focusing on transparency, whereas we focus specifically on explainability. For example, P7001 includes requirements about warning bystanders that sensors are collecting information, and providing certification agencies with information about verification and validation activities that were done. Focusing on explainability aspects, P7001 is fairly limited, making our scoresheet useful and complementary. For instance, we also include information on veracity, on how *well* the system works, and consider factors such as the level of confidence, scope of generalisation, concepts used, level of automation, and additional question types ("why not?", contrastive questions, "how to be?", "how to still be?").

To illustrate, consider the example (Appendix B.2 of IEEE P7001) of a medical diagnosis AI. With respect to explainability (as opposed to transparency more broadly), the assessment in the appendix of IEEE P7001 specifies only that end users (i.e. clinicians) need to be provided with (i) information on how the system functions (i.e. global explanation) specifically in an interactive form, and (ii) with the ability to pose "why?" and "what if?" questions to the system (levels 1-4). Some things that are is missing from this assessment but captured by our XAI scoresheet are:

- Veracity: how are explanations derived, and how can we know that an explanation corresponds to the actual reason?
- Are explanations interactive? Do they they include indications of the system's confidence? Do they include an indication of the scope within which they are valid?
- Does the system support contrastive questions? Does it support other forms of hypothetical questions? (e.g. "what would I need to change to get this (different) recommendation?")

There are also some other differences including that IEEE P7001 gives a set of orthogonal transparency requirements, classified by stakeholder, whereas we do not classify by stakeholder, since there is not a clear distinction between stakeholders [18]; and that for incident investigators P7001 requires a blackbox ("Event Data Recorder"), whereas we do not require a blackbox, since it is not essential to providing explanations.

4 OPERATIONALISING THE SCORESHEET

In this section we consider the question of *how* to use the scoresheet, in other words, when filling it out, how does one work out what the answers should be? We also note what other information is useful to capture (apart from what is in the scoresheet).

However, before starting to fill out the scoresheet for a given system that is being considered, we first need to identify who the relevant stakeholders are, and then what are their goals. We also need to identify for the application domain what are the risks that exist, and what level of risk is considered acceptable. This is required because to assess, for instance, whether there is adequate explanation of (globally) how the system operates, we are really answering the question of whether the provided information allows the stakeholders to gain an understanding of the system's functioning that is adequate for their goals. In other words, we need to know the stakeholders and their goals to assess this. For example, an elderly person using a domestic robot to support their independent living would need less information on how the robot functions and its limitations (e.g. tasks it cannot do well) than an agency responsible for certifying these robots for domestic use. Similarly, in order to assess the system's reliability, we need to know what the needs are: what can go wrong, and what are the potential consequences?

Basic information: this covers a few questions, that can be answered by asking the developer. However, although these questions appear to be answered by a simple "yes" or "no", they are actually an example of where there is additional information that is not in the scoresheet itself that is useful to capture. For example, when indicating that there is access to the developers of the system, there

Global Explanation checklist

There is an adequate description of:
- ☐ . . . *how* the system operates, including
  - ☐ . . . its (static) *structure*
  - ☐ . . . its (dynamic) *process*
- ☐ . . . *how well* the system functions, including information on
  - ☐ . . . the system's *performance*
  - ☐ . . . risks (including ethical issues)
  - ☐ . . . the system's *limitations* (e.g. situations in which it should (not) be used)

If the system uses training data:
- ☐ Information about the training data is available (e.g. its source, size)
- ☐ . . . including information on the process (e.g. data selection, cleaning, etc.)

Figure 2: Global Explanation Checklist

is a range of other information that is important to consider and record. For instance: How accessible are the developers? How reliably and quickly are they likely to be able to respond to queries or to meeting requests? To what extent might developers be reluctant to be transparent, especially when doing so might reveal an area of weakness in the system's performance? Similar considerations apply to access to trusted domain experts. This additional information is not included in the scoresheet itself in order to keep the scoresheet a brief and useful summary.

Veracity: If the system does not provide local explanations, then the answer to this is a simple "not applicable". Otherwise, to answer this question we need to consider the process by which (local) explanations are generated. The key question is: "if I get an explanation, how confident can I be that this actually reflects the real reasons for the behaviour I am seeing?". There are two approaches that can be used in order to complete this part of the scoresheet. Firstly, one can simply ask the system's developers to explain how explanations are generated (a meta-explanation), with particular emphasis on the links to the decision-making module and what steps were taken to ensure high reliability. Alternatively, it may be possible to evaluate veracity experimentally by setting up scenarios to see whether it is possible for explanations to deviate from the real reasons. For example, having (potentially adversarially-generated) scenarios A and B that give different behaviours but where the explanation provided for a question, such as "Why did you do A?" provides an explanation that only refers to features that are the same as in B.

Global Explanations: There are just two yes/no questions to be answered, but in order to answer them there is additional information that needs to be considered, and in fact we create an auxiliary checklist[12] (Figure 2) to ensure that it is considered. In addition to the checklist, it can also be useful to capture in what form the global explanation is provided. For example, is it a static document or in an interactive form [20]? The explanation could also use a range of (possibly derived) models such as decision tree, rules, or weighted features [23].

The essential question here is whether information is provided (on "how?" and "how well?") in a form and at the level of detail that is appropriate for the relevant stakeholder(s), and whether the information provided is adequate for their needs. For example, the level of understanding of how a system operates may be lower for a user of a system and higher for someone certifying the system for use in a given context.

The two unindented checkboxes in Figure 2 correspond to the two questions under Global Explanations in Figure 1. We would normally expect that in order to get an overall tick, the indented questions would also need to be ticked. It would clearly be unusual to indicate, for example, that there is an adequate description of how the system functions, without there being both descriptions of the system's (static) structure and its (dynamic) process of operation. Similarly, it would be unusual to consider information on how well the system functions to be adequate if it did not address the system's performance, the risks associated with its use, and its limitations.

With regard to the questions under "how well", the first ("performance") indicates whether information has been provided on how well the system operates within the intended domain of application. In other words, when the system is being used as intended, in a domain that it is designed for, how well does it perform, for instance, how accurate is it? The second ("risk") indicates whether information has been provided on what risks exist in relation to the use of the system (including any ethical issues). The third ("limitations") indicates whether information has been provided on the boundaries of intended use: in what situations is the system's effectiveness reduced, or, indeed, the system should not be used? For example, an application for assessing loan applications may only be appropriate to use when the applicants are salaried employees. Edge cases can play a role in documenting these boundaries.

If the system uses training data, then it can be important to also have information on the training data (such as where/how it was obtained, its size, and other characteristics such as demographic distributions), and on the process that was used to prepare the data (e.g. selection, cleaning, quality assessment) and to use it (e.g. training methodology, hyper-parameters). There can also be relevant data-related information included in the discussion of limitations. For example, that a given data set only covers certain demographic groups adequately, so should not be used for other demographic groups. There is a range of work on how to provide information about data in this context that can be leveraged (e.g. [2, 9, 12, 19, 28]).

Finally, moving on to **Local Explanations**, recall that information in the scoresheet covers a range of things: features of explanations, the concepts used, the explanation types supported, and the level of automation.

The explanatory features provided (e.g. individual customisation, interactivity) should be able to be determined by asking or just by using the system. For the first one (individual customisation), the question is whether it is possible for different people/roles asking the same question to be able to get different (relevant to each) answers. If the answer is yes, then it can be useful to also capture (not on the scoresheet) the *extent* and *forms* of individual customisation. For example, can a question include an explicit indication of what level of detail is sought, or what concepts should be used?

[12]We use the term "checklist" here since, unlike the scoresheet in Figure 1, all the responses here are ticks in boxes.

> "*A bicycle was not available, money was available, the made choice (catch bus) has the shortest duration to get home (in comparison with walking) . . . I needed to buy a bus ticket in order to allow you to go by bus, and I have the goal to allow you to catch the bus.*"

Figure 3: Example Explanation from [43, §2]

Identifying the concepts used in explanations requires looking at a range of explanations (and documentation). It may not always be immediately clear which parts of the explanation correspond to which concepts. For example, the explanation in Figure 3 has a number of elements, and it may not be immediately clear which are beliefs, goals, or preferences. Identifying instances of concepts can be done by applying the definitions of the concepts (see [45, Appx A]). For example, "money was available" is a factual statement about the environment, i.e. a belief. On the other hand "to allow you to catch the bus" is a single desired state, i.e. a goal, whereas a statement that compares more than one alternative indicates a preference (e.g. "made choice . . . has the shortest duration . . . in comparison with . . . "). Finally, if an explanation (or part of it) does not appear to map to any of the concepts, then it is an "Other" (e.g. "I needed to buy a bus ticket in order to allow you to go by bus" is an example of doing one thing in order to enable a later action).

Similarly, identifying the forms of explanation types provided requires looking at a range of questions (and documentation), and may require some interpretation. For example a question of the form "What situation would give an outcome of X?" does not immediately correspond to the question types in the scoresheet. However, considering what is provided to the system (the desired behaviour) and what it provides to the human (the situation, i.e. conditions under which the desired outcome occurs) can allow us to see that it corresponds to a form of "How to be?" - what situation will lead to desired behaviour (see also §5.5).

Finally, identifying the level of automation should be straightforward.

5 APPLYING TO DIFFERENT USE CASES

In this section we demonstrate the utility and versatility of the scoresheet by applying it to a range of systems. This shows that it can be applied to a diverse range of systems, and also demonstrates that the scoresheet for a system summarises information about the explainability of the system in a useful form.

We have selected the following six systems, which represent a broad range of types of intelligent or autonomous systems: (1) ChatGPT being used to recommend travel activities; (2) Generative AI being used to generate medical images; (3) A planner being used in a robotic application; (4) A search and rescue application implemented using BDI (Belief-Desire-Intention) concepts (goals, plans); (5) A multi-agent reinforcement learning system applied in a number of domains including a multi-robot search and rescue; and (6) A taxi scheduling domain where the system combines learning and planning. See Figure 4 for the corresponding scoresheets.

We note that the scoresheets are based on the specific systems mentioned on an as-is basis, rather than what could be done to the systems to make them more explainable, as there are certainly ways to do so.

5.1 ChatGPT for activity recommendation

We selected ChatGPT as an example of a general-purpose LLM, and applied it to the domain of generating recommendations for activities when visiting a city. The transcripts from our interaction with ChatGPT are available [45, Appx B]. In addition to asking ChatGPT for recommendations, we also asked for a range of explanations. We were expecting ChatGPT to do relatively poorly, but in fact it did quite well in providing explanations (as indicated in the bottom of the scoresheet, see Figure 4).

However, it is important to note that there is no information on what measures (if any) have been taken to attempt to ensure that answers, including explanations, reflect the actual reasons. Since ChatGPT is known to bullshit [15] (sometimes euphemistically termed "hallucinate"), this is an issue, since it means that the explanations cannot be relied upon. This is highlighted in the scoresheet.

5.2 PET Image Generation

This system uses generative AI to generate PET (Positron Emission Tomography) images [30]. It takes PET images from one radiotracer and generates pseudo-PET images of another radiotracer. The training and test data were obtained from a hospital with appropriate privacy and ethics approvals. The scoresheet clearly captures that while there is information provided on both *how* and *how well* the system works, the system does not have the ability to explain specific images generated, other than providing a confidence level (e.g. 0.85).

5.3 Planning for mobile service robot

This system [37] uses a hybrid planning system (CHIMP), that combines HTN-style task decomposition and meta-CSP search, resulting in an HTN planner able to handle very rich domain knowledge. This is applied to an application of a mobile service robot that performs tasks such as serving hot coffee with sugar. For such a task, it must reason not just about the consequences of each action but also the duration of the action, whilst considering alternative possibilities for accomplishing the same task.

The system keeps a log of what was done and why. This makes it possible to obtain information to answer a broad range of questions. However, as highlighted in the scoresheet, this needs to be done manually by the developers. On the other hand, because this information is generated directly from the planner, the explanations can be relied upon.

5.4 Search & Rescue using BDI

This system is a simulation that controls UAVs carrying out a search and rescue task [36]. It is implemented using BDI concepts (goals and plans) in SARL [33], and uses the TriQPAN pattern [34, 35] to extend SARL to be able to provide a range of (local) explanations. The scoresheet clearly indicates that the system is able to provide a range of explanations, and that this is fully automated. It also indicates that the explanations are directly derived from logs of the actual system, so the explanations can be relied upon.

(a) Chat GPT used for Itinerary Recommendation

XAI scoresheet for Chat-GPT

☐ System source code is available
Is training data used available? Yes / No / Not Applicable
☐ There is access to the system's developers
☐ There is access to trusted domain experts

Veracity:
How reliable are explanations? Not Applicable / Low / High
What steps are taken to ensure explanation reliability?
unsure
* * *

Global Explanations: Has information been provided on:
☑ *How does the system work?*
☑ *How well does it work?*
(See checklist - Figure 2)

Local Explanations: Explanations ...
☑ ...can be **individually customised**
☑ ...are **interactive**
☐ ...include an indication of **confidence**
☐ ...include an indication of **scope of generalisation**

What Concepts are used in explanations?
☑ Examples ☐ Features ☐ Beliefs ☐ Events/Percepts
☐ Goals ☐ Actions ☐ Preferences ☐ Values
☐ Other: _____

What forms of **Explanation Types** are provided?
Factual/Past: ☐ Did? ☐ Why? ☐ Why not? ☐ Contrastive
Future-looking: ☐ Will? ☐ Why? ☐ Why not? ☐ Contrastive
Hypothetical: ☐ What-if? ☐ How to be? ☐ How to still be?
Other: _____

Is **explanation generation** from questions?
☑ Fully automated ☐ Partially automated ☐ Manual

In general, yes, but not for this particular case
ChatGPT did surprisingly well!

(b)Multiagent RL – Search and Rescue Simulation

XAI scoresheet for MARL

☑ System source code is available
Is training data used available? Yes / No / Not Applicable
☑ There is access to the system's developers
☑ There is access to trusted domain experts

Veracity:
How reliable are explanations? Not Applicable / Low / High
What steps are taken to ensure explanation reliability?
Explanations derived from model
* * *

Global Explanations: Has information been provided on:
☐ *How does the system work?*
☑ *How well does it work?*
(See checklist - Figure 2)

Local Explanations: Explanations ... It is more like a static check of likely outcomes
☐ ...can be **individually customised**
☐ ...are **interactive**
☐ ...include an indication of **confidence**
☐ ...include an indication of **scope of generalisation**

What Concepts are used in explanations?
☑ Examples ☐ Features ☑ Beliefs ☐ Events/Percepts
☑ Goals ☑ Actions ☐ Preferences ☐ Values
☐ Other: _____

What forms of **Explanation Types** are provided?
Factual/Past: ☐ Did? ☐ Why? ☐ Why not? ☐ Contrastive
Future-looking: ☐ Will? ☐ Why? ☐ Why not? ☐ Contrastive
Hypothetical: ☑ What-if? ☑ How to be? ☐ How to still be?
Other: They use different terms, mapped to the above

Is **explanation generation** from questions?
☑ Fully automated ☐ Partially automated ☐ Manual

(c) Generative AI used in PET Imaging

XAI scoresheet for GenAI-PET

☑ System source code is available
Is training data used available? Yes / No / Not Applicable
☑ There is access to the system's developers
☑ There is access to trusted domain experts

Veracity:
How reliable are explanations? Not Applicable / Low / High
What steps are taken to ensure explanation reliability?
* * *

Global Explanations: Has information been provided on:
☑ *How does the system work?* This is based on confidence ratings
☐ *How well does it work?*
(See checklist - Figure 2)

Local Explanations: Explanations ... The system does not provide any explanations
☐ ...can be **individually customised**
☐ ...are **interactive**
☐ ...include an indication of **confidence**
☐ ...include an indication of **scope of generalisation**

What Concepts are used in explanations?
☐ Examples ☐ Features ☐ Beliefs ☐ Events/Percepts
☐ Goals ☐ Actions ☐ Preferences ☐ Values
☐ Other: _____

What forms of **Explanation Types** are provided?
Factual/Past: ☐ Did? ☐ Why? ☐ Why not? ☐ Contrastive
Future-looking: ☐ Will? ☐ Why? ☐ Why not? ☐ Contrastive
Hypothetical: ☐ What-if? ☐ How to be? ☐ How to still be?
Other: _____

Is **explanation generation** from questions?
☐ Fully automated ☐ Partially automated ☐ Manual

(d) SARL APL – Search and Rescue Simulation

XAI scoresheet for SARL

☑ System source code is available
Is training data used available? Yes / No / Not Applicable
☑ There is access to the system's developers
☑ There is access to trusted domain experts

Veracity:
How reliable are explanations? Not Applicable / Low / High
What steps are taken to ensure explanation reliability?
Derived from execution logs, traces, system model and XAg Engine
* * *

Global Explanations: Has information been provided on:
☑ *How does the system work?*
☑ *How well does it work?*
(See checklist - Figure 2)

Local Explanations: Explanations ...
☑ ...can be **individually customised**
☑ ...are **interactive**
☐ ...include an indication of **confidence**
☐ ...include an indication of **scope of generalisation**

What Concepts are used in explanations?
☑ Examples ☑ Features ☑ Beliefs ☑ Events/Percepts
☑ Goals ☑ Actions ☑ Preferences ☑ Values
☐ Other: _____

What forms of **Explanation Types** are provided?
Factual/Past: ☑ Did? ☑ Why? ☑ Why not? ☑ Contrastive
Future-looking: ☐ Will? ☐ Why? ☐ Why not? ☐ Contrastive
Hypothetical: ☐ What-if? ☐ How to be? ☐ How to still be?
Other: _____

Is **explanation generation** from questions?
☑ Fully automated ☐ Partially automated ☐ Manual

(e) CHIMP HTN planner – Farm Robot Simulation

XAI scoresheet for CHIMP-HTN

☑ System source code is available
Is training data used available? Yes / No / Not Applicable
☑ There is access to the system's developers
☑ There is access to trusted domain experts

Veracity:
How reliable are explanations? Not Applicable / Low / High
What steps are taken to ensure explanation reliability?
Derived from execution traces and simulations. Visualisation tools * * * also assist
* * *

Global Explanations: Has information been provided on:
☑ *How does the system work?*
☐ *How well does it work?*
(See checklist - Figure 2)

Local Explanations: Explanations ... Manual. tools available to log, trace and visualise
☑ ...can be **individually customised**
☑ ...are **interactive**
☐ ...include an indication of **confidence**
☐ ...include an indication of **scope of generalisation**

What Concepts are used in explanations?
☑ Examples ☑ Features ☑ Beliefs ☑ Events/Percepts
☑ Goals ☑ Actions ☐ Preferences ☐ Values
☐ Other: _____

What forms of **Explanation Types** are provided?
Factual/Past: ☑ Did? ☑ Why? ☑ Why not? ☑ Contrastive
Future-looking: ☑ Will? ☑ Why? ☑ Why not? ☑ Contrastive
Hypothetical: ☑ What-if? ☑ How to be? ☑ How to still be?
Other: _____

Is **explanation generation** from questions?
☐ Fully automated ☑ Partially automated ☑ Manual
and requires tools

We can manually inspect and explain

(f) Hybrid Deep RL and Symbolic planning – Taxi Simulation

XAI scoresheet for SAGE Hybrid DRL/Planning

☑ System source code is available
Is training data used available? Yes / No / Not Applicable
☑ There is access to the system's developers
☑ There is access to trusted domain experts

Veracity:
How reliable are explanations? Not Applicable / Low / High
What steps are taken to ensure explanation reliability?
Only applicable for planning component for DRL
* * *

Global Explanations: Has information been provided on:
☑ *How does the system work?*
☐ *How well does it work?*
(See checklist - Figure 2)

Local Explanations: Explanations ...
☑ ...can be **individually customised**
☑ ...are **interactive**
☐ ...include an indication of **confidence**
☐ ...include an indication of **scope of generalisation**

What Concepts are used in explanations?
☑ Examples ☑ Features ☑ Beliefs ☑ Events/Percepts
☑ Goals ☑ Actions ☐ Preferences ☐ Values
☐ Other: _____

What forms of **Explanation Types** are provided? manual tracing
Factual/Past: ☑ Did? ☑ Why? ☑ Why not? ☑ Contrastive
Future-looking: ☑ Will? ☑ Why? ☑ Why not? ☑ Contrastive
Hypothetical: ☑ What-if? ☑ How to be? ☑ How to still be?
Other: _____

Is **explanation generation** from questions?
☐ Fully automated ☑ Partially automated ☑ Manual
and requires tools

Global Explanation checklist

There is an adequate description of:
☑ ...*how* the system operates, including
☑ ...its (static) *structure*
☑ ...its (dynamic) *process*
☐ ...*how well* the system functions, including information on
☑ ...the system's *performance*
☑ ...risks (including ethical issues)
☐ ...the system's *limitations*
(e.g. situations in which it should (not) be used)

If the system uses training data:
☑ Information about the training data is available
(e.g. its source, size)
☐ ...including information on the process
(e.g. data selection, cleaning, etc.)

Global Explanation checklist

There is an adequate description of:
☑ ...*how* the system operates, including
☐ ...its (static) *structure*
☐ ...its (dynamic) *process*
☑ ...*how well* the system functions, including information on
☑ ...the system's *performance*
☑ ...risks (including ethical issues)
☑ ...the system's *limitations*
(e.g. situations in which it should (not) be used)

If the system uses training data:
☐ Information about the training data is available
(e.g. its source, size)
☐ ...including information on the process
(e.g. data selection, cleaning, etc.)

Figure 4: Completed Scoresheets for the Six Systems

5.5 Multi-Agent Reinforcement Learning

This work extends multi-agent reinforcement learning with explanation features [5], building on earlier work on single agent reinforcement learning explanation [14]. They apply their approach to three domains: a multi-robot search and rescue scenario, a multi-robot cooperative delivery task, and a grid-based game where agents cooperate and compete to collect food.

In essence, they provide two things: an algorithm to create a summary of a policy, and an algorithm to provide explanations for given queries (they extend this in a subsequent paper to temporal logic queries [6]).

The first contribution, a summary of a policy, is a *global* explanation ("*policy summarization provides a global view of the agent behavior under a MARL policy*" [5, §4]). However, while the query-based explanations provide what look like typical local explanations, in fact the explanations are in terms of *likely paths*, rather than in terms of a particular execution of the system.

Regardless of this though, it is interesting to observe that the three question types they support do not match in an obvious way to the question types that we have included in our scoresheet. Specifically, the first question type ("When do [agents] do [actions]?") is used "*for identifying conditions for action(s) of a single or multiple agent(s)*" [5, §4]. This can be seen, in intent, if not phrasing, as being related to "how to be?": it is identifying conditions that allow particular actions (i.e. behaviours) to occur. The second question type ("Why don't [agents] do [actions] in [states]?") is clearer, corresponding to our "Why not?". Finally, the third question type that they support ("What do [agents] do in [predicates]?") is used "*for revealing agent behavior under specific conditions*" (ibid) and can be seen as a form of "what if?": given particular conditions, what would happen?

The scoresheet clearly captures that this system provides local explanations of various types, and that the explanation generation is done directly from the behaviour-generating module, and hence the explanations can be relied upon.

5.6 Taxi planning using learning & planning

This work [8] proposes an architecture that combines planning and learning, and demonstrates it in a taxi planning domain. The architecture has three levels: a top-level that uses reinforcement learning to identify what are the best goals to select, a middle level that uses an off-the-shelf planner to develop plans to achieve these goals, and a low-level module that uses deep reinforcement learning to perform low-level actions within the plans.

In terms of using the scoresheet to assess the explainability aspects of this system a key challenge is that it has three modules, each of which has different explainability features. The planning module (similar to §5.3) captures information that can be used to (manually) generate (highly reliable) explanations. However, the deep reinforcement learning module does not provide any form of explainability.

There are two ways in which this can be captured using the scoresheet. The first (which is preferred) is to use a single scoresheet for the whole system, but annotate it to indicate when answers apply to only parts of the system. For example, for veracity we might indicate that it is "Not Applicable" for the RL part of the system and "High" for the planning component. The second way, which may be required if the first approach yields an overly cluttered and complex scoresheet, is to have a separate scoresheet for different modules in the system (perhaps with a system-wide scoresheet that refers to them).

6 DISCUSSION & CONCLUSION

We have presented a scoresheet for explainability, along with detailed guidance for how to use it. The scoresheet was then applied to a broad range of systems, demonstrating its usability and generality. Looking at the results of applying the scoresheet (Figure 4) we can see that important explainability features of the different systems are captured. For example, for ChatGPT it is clear that explanations may not be reliable, but that the system provides a range of explanation types. On the other hand, for PET image generation, the scoresheet captures clearly that only global explanations are available. For the mobile service robot the scoresheet clearly indicates that a range of (local) explanations are available, and that they can be relied upon (because they are generated directly from the planner), but that the construction of explanations from the information is a manual process. The search and rescue (using SARL) and Multi-agent reinforcement learning are similar in providing a range of (reliable) explanations, and do not require manual construction of these explanations. Finally, the taxi planning application scoresheet captures clearly that there are multiple modules in the system, and that these have different explainability characteristics.

6.1 Limitations & Future Work

One limitation is that the scoresheet has only been used by the authors. Therefore, future work includes further use and evaluation of the scoresheet. This could include having a range of people (e.g. various roles, covering the stakeholder types discussed in §2, as well as a range of experience levels and diverse demographics) use it to assess systems. It could also include assessing how well the scoresheet can be used for other use cases (e.g. specifying the explainability requirements of an application, rather than assessing a given system). This would be done by indicating what XAI features are required of a system that is to be used in a certain context, e.g. if a bank was looking to develop a system for making loan decisions it could use the scoresheet to specify what XAI features would be required for the system-to-be. Indeed, it might be possible to use a scoresheet to specify the explainability requirements for a whole sector or domain (e.g. transport, policing), or even to specify regulatory requirements relating to explainability.

Finally, we highlight some broader research challenges for the XAI research community. There is a need to move beyond explaining particular decisions or actions (local explanations) to be able to provide useful information on *how* the system works, using local explanations to illustrate (i.e. "global-local" explanations), including highlighting edge cases [18]. There is also a need to be able to identify and include information in particular about behaviours that are *surprising* [17].

ACKNOWLEDGMENTS

This research is partially supported by the C2IMPRESS project funded by the EU.

REFERENCES

[1] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable Agents and Robots: Results from a Systematic Literature Review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*, Edith Elkind, Manuela Veloso, Noa Agmon, and Matthew E. Taylor (Eds.). 1078–1088. http://dl.acm.org/citation.cfm?id=3331806

[2] Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, Karthikeyan Natesan Ramamurth, Darrell Reimer, Alexandra Olteanu, David Piorkowski, Jason Tsay, and Kush R. Varshney. 2019. FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. *CoRR* abs/1808.07261 (2019). arXiv:1808.07261 http://arxiv.org/abs/1808.07261

[3] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilovic, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John T. Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2019. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. *CoRR* abs/1909.03012 (2019). arXiv:1909.03012 http://arxiv.org/abs/1909.03012

[4] Or Biran and Kathleen McKeown. 2014. Justification Narratives for Individual Classifications. In *ICML 2014 AutoML Workshop*. 7.

[5] Kayla Boggess, Sarit Kraus, and Lu Feng. 2022. Toward Policy Explanations for Multi-Agent Reinforcement Learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, Luc De Raedt (Ed.). ijcai.org, 109–115. https://doi.org/10.24963/IJCAI.2022/16

[6] Kayla Boggess, Sarit Kraus, and Lu Feng. 2023. Explainable Multi-Agent Reinforcement Learning for Temporal Queries. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*. ijcai.org, 55–63. https://doi.org/10.24963/IJCAI.2023/7

[7] Miriam C. Buiten, Louise A. Dennis, and Maike Schwammberger. 2023. A Vision on What Explanations of Autonomous Systems are of Interest to Lawyers. In *31st IEEE International Requirements Engineering Conference, RE 2023 - Workshops, Hannover, Germany*, Kurt Schneider, Fabiano Dalpiaz, and Jennifer Horkoff (Eds.). IEEE, 332–336. https://doi.org/10.1109/REW57809.2023.00062

[8] Andrew Chester, Michael Dann, Fabio Zambetta, and John Thangarajah. 2023. SAGE: Generating Symbolic Goals for Myopic Models in Deep Reinforcement Learning. In *AI 2023: Advances in Artificial Intelligence - 36th Australasian Joint Conference on Artificial Intelligence, AI 2023, Brisbane, QLD, Australia, November 28 - December 1, 2023, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 14472)*, Tongliang Liu, Geoffrey I. Webb, Lin Yue, and Dadong Wang (Eds.). Springer, 274–285. https://doi.org/10.1007/978-981-99-8391-9_22

[9] Kasia S. Chmielinski, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. 2022. The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence. *CoRR* abs/2201.03954 (2022). arXiv:2201.03954 https://arxiv.org/abs/2201.03954

[10] Stephen Cranefield, Nir Oren, and Wamberto Weber Vasconcelos. 2018. Accountability for Practical Reasoning Agents. In *Agreement Technologies - 6th International Conference, AT 2018, Bergen, Norway, December 6-7, 2018, Revised Selected Papers (LNCS, Vol. 11327)*, Marin Lujak (Ed.). Springer, 33–48. https://doi.org/10.1007/978-3-030-17294-7_3

[11] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. 2018. 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

[12] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for Datasets. *CoRR* abs/1803.09010 (2018). arXiv:1803.09010 http://arxiv.org/abs/1803.09010

[13] Shirley Gregor and Izak Benbasat. 1999. Explanations From Intelligent Systems: Theoretical Foundations and Implications for Practice. *MIS Q.* 23, 4 (1999), 497–530. http://misq.org/explanations-from-intelligent-systems-theoretical-foundations-and-implications-for-practice.html

[14] Bradley Hayes and Julie A. Shah. 2017. Improving Robot Controller Transparency Through Autonomous Policy Explanation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2017, Vienna, Austria, March 6-9, 2017*, Bilge Mutlu, Manfred Tscheligi, Astrid Weiss, and James E. Young (Eds.). ACM, 303–312. https://doi.org/10.1145/2909824.3020233

[15] Michael Townsen Hicks, James Humphries, and Joe Slater. 2024. ChatGPT is bullshit. *Ethics Inf. Technol.* 26, 2 (2024), 38. https://doi.org/10.1007/S10676-024-09775-5

[16] Robert R. Hoffman, Mohammadreza Jalaeian, Connor Tate, Gary Klein, and Shane T. Mueller. 2023. Evaluating machine-generated explanations: a "Scorecard" method for XAI measurement science. *Frontiers Comput. Sci.* 5 (2023). https://doi.org/10.3389/FCOMP.2023.1114806

[17] Robert R. Hoffman, Tim Miller, Gary Klein, Shane T. Mueller, and William J. Clancey. 2023. Increasing the Value of XAI for Users: A Psychological Perspective. *Künstliche Intell.* 37, 2 (2023), 237–247. https://doi.org/10.1007/S13218-023-00806-9

[18] Robert R. Hoffman, Shane T. Mueller, Gary Klein, Mohammadreza Jalaeian, and Connor Tate. 2023. Explainable AI: roles and stakeholders, desirements and challenges. *Frontiers Comput. Sci.* 5 (2023). https://doi.org/10.3389/FCOMP.2023.1117848

[19] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *CoRR* abs/1805.03677 (2018). arXiv:1805.03677 http://arxiv.org/abs/1805.03677

[20] IEEE. 2022. IEEE Standard for Transparency of Autonomous Systems. IEEE Std 7001-2021. , 54 pages. https://doi.org/10.1109/IEEESTD.2022.9726144

[21] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. 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 (2021), 103473. https://doi.org/10.1016/j.artint.2021.103473

[22] Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. 2017. Explainable Agency for Intelligent Autonomous Systems. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, Satinder Singh and Shaul Markovitch (Eds.). AAAI Press, 4762–4764. https://doi.org/10.1609/aaai.v31i2.19108

[23] Q. Vera Liao, Daniel M. Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik (Eds.). ACM, 1–15. https://doi.org/10.1145/3313831.3376590

[24] Bertram F. Malle. 2004. *How the Mind Explains Behavior*. MIT Press. ISBN: 9780262134453.

[25] Siddharth Mehrotra, Catholijn M. Jonker, and Myrthe L. Tielman. 2021. More Similar Values, More Trust? - the Effect of Value Similarity on Trust in Human-Agent Interaction. In *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, Marion Fourcade, Benjamin Kuipers, Seth Lazar, and Deirdre K. Mulligan (Eds.). ACM, 777–783. https://doi.org/10.1145/3461702.3462576

[26] Siddharth Mehrotra, Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. 2024. Integrity-based Explanations for Fostering Appropriate Trust in AI Agents. *ACM Trans. Interact. Intell. Syst.* 14, 1 (2024), 4:1–4:36. https://doi.org/10.1145/3610578

[27] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[28] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2018. Model Cards for Model Reporting. *CoRR* abs/1810.03993 (2018). arXiv:1810.03993 http://arxiv.org/abs/1810.03993

[29] Brent D. Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining Explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*), Atlanta, GA, USA, January 29-31, 2019*, danah boyd and Jamie H. Morgenstern (Eds.). ACM, 279–288. https://doi.org/10.1145/3287560.3287574

[30] Prabath Hetti Mudiyanselage, Ruwan B. Tennakoon, John Thangarajah, Robert Ware, Jason H. Callahan, and Lucy Vivash. 2023. Preliminary Study of Pseudo-PET Image Synthesis of Glucose Metabolism from Early-Phase PET Images of an Uncorrelated Radiotracer. In *International Conference on Digital Image Computing: Techniques and Applications, DICTA 2023, Port Macquarie, Australia, November 28 - Dec. 1, 2023*. IEEE, 237–244. https://doi.org/10.1109/DICTA60407.2023.00040

[31] Lin Padgham and Michael Winikoff. 2004. *Developing intelligent agent systems - a practical guide*. Wiley. https://doi.org/10.1002/0470861223

[32] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M. Howard, and Alan R. Wagner. 2016. Overtrust of Robots in Emergency Evacuation Scenarios. In *The Eleventh ACM/IEEE International Conference on Human Robot Interation, HRI 2016, Christchurch, New Zealand, March 7-10, 2016*, Christoph Bartneck, Yukie Nagai, Ana Paiva, and Selma Sabanovic (Eds.). IEEE/ACM, 101–108. https://doi.org/10.1109/HRI.2016.7451740

[33] Sebastian Rodriguez, Nicolas Gaud, and Stéphane Galland. 2014. SARL: A General-Purpose Agent-Oriented Programming Language. In *The 2014 IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, Vol. 3. IEEE Computer Society Press, Warsaw, Poland, 103–110. https://doi.org/10.1109/WI-IAT.2014.156

[34] Sebastian Rodriguez and John Thangarajah. 2024. Explainable Agents (XAg) by Design (Blue Sky Ideas Track). In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, N. Alechina, V. Dignum, M. Dastani, and J.S. Sichman (Eds.). ACM. https://doi.org/10.5555/3545946.3598908

[35] Sebastian Rodriguez, John Thangarajah, and Andrew Davey. 2024. Design Patterns for Explainable Agents (XAg). In *Proc. of the 23rd International Conference*

on Autonomous Agents and Multiagent Systems (AAMAS), N. Alechina, V. Dignum, M. Dastani, and J.S. Sichman (Eds.). ACM. https://doi.org/10.5555/3545946.3598908

[36] Sebastian Rodriguez, John Thangarajah, and Michael Winikoff. 2023. A Behaviour-Driven Approach for Testing Requirements via User and System Stories in Agent Systems. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems* (London, United Kingdom) *(AAMAS '23)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1182–1190.

[37] Sebastian Stock, Masoumeh Mansouri, Federico Pecora, and Joachim Hertzberg. 2015. Hierarchical Hybrid Planning in a Mobile Service Robot. In *KI 2015: Advances in Artificial Intelligence*, Steffen Hölldobler, Rafael Peñaloza, and Sebastian Rudolph (Eds.). Springer International Publishing, Cham, 309–315.

[38] Ruben S. Verhagen, Mark A. Neerincx, and Myrthe L. Tielman. 2021. A Two-Dimensional Explanation Framework to Classify AI as Incomprehensible, Interpretable, or Understandable. In *Third International Workshop on Explainable and Transparent AI and Multi-Agent Systems (EXTRAAMAS), Revised Selected Papers (LNCS, Vol. 12688)*, Davide Calvaresi, Amro Najjar, Michael Winikoff, and Kary Främling (Eds.). Springer, 119–138. https://doi.org/10.1007/978-3-030-82017-6_8

[39] Alan F. T. Winfield, Serena Booth, Louise A. Dennis, Takashi Egawa, Helen F. Hastie, Naomi Jacobs, Roderick I. Muttram, Joanna I. Olszewska, Fahimeh Rajabiyazdi, Andreas Theodorou, Mark A. Underwood, Robert H. Wortham, and Eleanor Nell Watson. 2021. IEEE P7001: A Proposed Standard on Transparency. *Frontiers Robotics AI* 8 (2021), 665729. https://doi.org/10.3389/frobt.2021.665729

[40] Michael Winikoff. 2017. Towards Trusting Autonomous Systems. In *Engineering Multi-Agent Systems - 5th International Workshop, EMAS 2017, Sao Paulo, Brazil, May 8-9, 2017, Revised Selected Papers (LNCS, Vol. 10738)*, Amal El Fallah Seghrouchni, Alessandro Ricci, and Tran Cao Son (Eds.). Springer, 3–20. https://doi.org/10.1007/978-3-319-91899-0_1

[41] Michael Winikoff. 2024. Towards Engineering Explainable Autonomous Systems. In *Engineering Multi-Agent Systems - 12th International Workshop, EMAS 2024, Auckland, New Zealand, May 6-7, 2024, Revised Selected Papers (Lecture Notes in Computer Science, Vol. 15152)*, Daniela Briola, Rafael C. Cardoso, and Brian Logan (Eds.). Springer, 144–155. https://doi.org/10.1007/978-3-031-71152-7_9

[42] Michael Winikoff and Julija Sardelić. 2021. Artificial Intelligence and the Right to Explanation as a Human Right. *IEEE Internet Comput.* 25, 2 (2021), 116–120. https://doi.org/10.1109/MIC.2020.3045821

[43] Michael Winikoff and Galina Sidorenko. 2023. Evaluating a Mechanism for Explaining BDI Agent Behaviour. In *Explainable and Transparent AI and Multi-Agent Systems - 5th International Workshop, EXTRAAMAS 2023, London, UK, May 29, 2023, Revised Selected Papers (Lecture Notes in Computer Science, Vol. 14127)*, Davide Calvaresi, Amro Najjar, Andrea Omicini, Reyhan Aydogan, Rachele Carli, Giovanni Ciatto, Yazan Mualla, and Kary Främling (Eds.). Springer, 18–37. https://doi.org/10.1007/978-3-031-40878-6_2

[44] Michael Winikoff, Galina Sidorenko, Virginia Dignum, and Frank Dignum. 2021. Why bad coffee? Explaining BDI agent behaviour with valuings. *Artif. Intell.* 300 (2021), 103554. https://doi.org/10.1016/J.ARTINT.2021.103554

[45] Michael Winikoff, John Thangarajah, and Sebastian Rodriguez. 2025. A Scoresheet for Explainable AI. arXiv:2502.09861 [cs.AI] https://arxiv.org/abs/2502.09861