Towards Self-Explaining Agents

Johannes Fähndrich, Sebastian Ahrndt, and Sahin Albayrak

Abstract We advocate Self-Explanation as the foundation for the Self-* properties. Arguing that for system component to have such properties the underlining foundation is a awareness of them selfs and their environment. In the research area of adaptive software, self-* properties have shifted into focus pushing ever more design decisions to a applications runtime. Thus fostering new paradigms for system development like intelligent agents. This work surveys the state of the art methods of self-explanation in software systems and distills a definition of self-explanation.

1 Introduction

The development of distributed systems in heterogeneous environments is a challenging task for humans [12]. As a matter of fact, the management of such systems where different parties at different times make use of different technologies to reach their goals becomes ever more difficult. Additionally, systems can dynamically change due to the presence or absence of agents, services and/or devices leading to configuration problems as well as to the occurrence of emergent behavior meaning behavior which are not pre-programmed into the systems. To address the arising issues, developers attempt to shift evermore details to the application's runtime enabling the system to adjust their internal states as a result to exogenous and/or endogenous influences [14, 13]. In this process, the exogenous influences can be identified as the context the system is embedded in, whereas the endogenous influences stem from the system itself. Here, the identification and reaction as response to an influence depends on several self-* properties [23], where the initial set is known as self-CHOP (configuring, healing, optimizing, protecting) [11]. Admittedly these properties are rather high-level and can be distinguished into sev-

Johannes Fähndrich · Sebastian Ahrndt · Sahin Albayrak
DAI-Labor, Technische Universität Berlin, Ernst-Reuter-Platz 7, 10587 Berlin, Germany
e-mail: johannes.faehndrich@dai-labor.de (Corresponding author)

eral basic properties, where one of this is Self-Explanation. Self-Explanation is inspired not only from biological systems but also by the field of social science. In this context, self-explanation is defined as an ability "of explaining to oneself in an attempt to make sense of new information, either presented in a text or in some other medium" [4]. Commonly, explaining events, intentions and ideas is a well-known way of communicating information in everyday life. On the one hand, the explaining entity is able to impart knowledge to some audience. On the other hand, the audience is able to understand and comprehend the explainer's intentions and they may even understand the explainer's course of actions. The goal of this work is to foster the understanding about the self-explanation property, specialized on multi agent systems where the description an agent can provide about itself, is interpreted as a explanation. Therefore we will provide an overview about the research field and the requirements we identified. In addition, we will introduce a formal definition of self-explanation and a metric enabling to decide which description is more self-explanatory.

2 Self-Explanation in a Nutshell

In the Cambridge Dictionary[1] the term to explain is defined as "to make something clear or easy to understand by describing or giving information about it". By examining this definition we notice that explaining is the act of giving information about an subject of interest to an audience with the intend to foster both the knowing and the understanding of the subject of interest. Going back to the initial set of self-* properties one can imagine that self-explanation injects momentum not only to the self-configuration but also to the other properties. Indeed, these properties can not be considered independently. Consequently, the term self-explanation has different meanings, too. Taking into account the different parties involved – agents (the system itself), developers and (end)users – we can distinguish between two sides of self-explanation. To start with, the *system side* aiming to integrate new agents autonomously into the existing infrastructure [22, 15]. Following the idea of self-explanation this means that new agents as well as existing ones are able to learn the capabilities of each other and to comprehend in which way they are able to interact (e.g. which data format and expressions match). One can imagine this process in the way a new human introduces itself into a prior unknown group of other humans by explaining its name and capabilities. Further we refer to the *human side* aiming to integrate the user into the system. As those systems are typically goal-driven, humans should be enabled to set the pursuit goals, to restrict the systems using constraints and to observe the results of the self-organization process [21, 22, 3].
However, several definitions of explanations have been proposed. Each one specialized for the needs of some domain. We will look at some of them to see how they can help defining the term. To start with, in statistics we can identify evidence weights

[1] Cambridge Dictionary Online, visit `http://dictionary.cambridge.org/`

in a Bayesian believe network as explanations [10]. These weights represent the logarithmic likelihood ratio of the influence of an observation on a specific variable. Therefore they can and indeed are used to explain in which way the occurrence of an event influences the current systems state [6]. To ease the access of humans to these statistical explanations different classes of techniques can be applied (e.g. verbal explanations [8] and graphical explanations [5]). In addition, *Druzdzel* [6] identified two categories in which such explanations can be separated: *Explanation of Assumptions* focusing on the communication of the domain model of the system and *Explanation of Reasoning* focusing on how conclusions are made from those assumptions. It might be worthwhile to transfer these categories to self-explanation since the meaning of concepts used might differ depending the exogenous or endogenous origin of the fact explained. Therefore the reasoner has to distinguish between the explanation of the system itself and how it can be interpreted related to the current context. This work focuses on the explanation of assumptions, since the audience of such an description is seen as an external system component. As those approaches are quite fundamental and thus general we further want to list more practical approaches in the agent community:

- *Braubach et al.* [2] uses the beliefs, desires and intents to formulate goals, knowledge and capabilities for a multi-agent system
- *Grüninger et al.* [9] uses First-order Logic Ontology for Web Services (FLOWS) to describe the functionalities of a service
- *Sycara et al.* [25] formulates agent and service capabilities utilizing the Input, Output, Precondition and Effect (IOPE) approach
- *Martin et al.* [19] uses the Ontology Web Language to structure the description of services

Those approaches all explain something about the subject of interest in specific domains but all lack the ability to measure the amount of information transfered by such an explanation, making it impossible to distinguish the quality of such explanations. In this work, we want to subsume those approaches in an theoretical framework building the foundation for a measure of the amount of self-explanation. For the reminder of this work we will utilize the following definition for the term self-explanation:

Definition 1. *Self-explanation identifies the capability of systems and system components to describe themselves and their functionalities to other systems, components or human beings.*

3 Towards Self-Explanatory Descriptions

In order to enable a system to be self-explaining the system has to provide information about its capabilities, interaction ways and current state. Nowadays this information are provided by e.g. service descriptions. The problem at hand here is the

semi-optimal performance of AI algorithms using currently available descriptions like service matcher and planner [16]. As there are multiple improvement points (for example the reasoner, the knowledge-base, the used languages and the formalisms), self-explanatory descriptions try to improve the description side. To extend the current available description to self-explanatory description, we distinguish between three different types of information: *Syntax*, concerning the interpretation of signals, *Semantics* concerning the meaning and relationship between entities and *Pragmatics* concerning the interpretation of statements [20]. *Sooriamurthi* and *Leake* [24] follow this fragmentation and present in an early work their view point on explanations in the Artificial Intelligence (AI) research domain. The authors emphasize that the context should be incorporated in the interpretation and creation of explanations to enable systems to adapt to dynamic situations and therefore introduce the use of pragmatics as context-dependent interpretation of meanings. This is important since the explaining system might have to cope with partial observable situations while creating an explanation. In such situations the proposed approach suggest to take former explanations to guide the search for information to create a new explanation. *Leake* [17] underpin this finding while arguing that with changing system goals the interpretation of an explanation should change to. The author also emphasizes that this requirement holds in different research fields like Psychology, Philosophy and AI. At the same time, *Leake* [18] uses the factors plausibility, relevance and usefulness for explanations concerning anomalies in regard to a given goal. Coming to the conclusion that "(m)any explanations can be generated for any event, and only some of them are plausible" [18]. The requirement we identify here is that a self-explanatory description must include not only regular information but also semantic information (about the meaning of the regular information) and context information for the context dependent meaning. This correlates with the overall goal of self-explanation proposed by *Müeller-Schloer et al.* [21] to enable systems to explain its current state, which seams to be impossible without providing contextual information.

4 Formal foundation

Explanation of assumptions might informally be defined as a description to reveal the identity of some subject of interest. This might for example include information about its functionality. Imagine that we want do identify different boat types for tax reasons. We might not use the appearance to identify the difference of a rowing-, sailing- and a motor boat, because there might be different appearances in each class of boats. Instead, to identify the different boat classes, we need to describe some other details like the propulsion method and the tonnage of the boat. In contrast, if somebody wants to describe the different boat types to a child the functionality might be the detail separating the identities. In AI this fact is well known, since we seek different metrics to decrease intra class scatter and increase inter class scatter [7] (p. 121). Further, the explanation we must provide depends not only on

the context but also on the reasoner how infers about it. With this in mind, an explanation should help the audience, to identify the classes a Subject of Interest SOI might be part of and with that better describe its identity to foster understanding of the explanation whereas the understanding determines the goodness of an explanation [17].
We now formalize these definitions, easing the creation of measurable properties of explanations to determine the quality of an explanations.
We define the amount of information transfered to the audience as a measure of quality of the explanation. First we want to define a domain as a set of information concerning this domain:

Definition 2. The information available in one domain $\mathbb{D}$ with $\mathbb{D} \subset \mathbb{I}$ and $\mathbb{I}$ bing all information available.

Here, the basic assumption we follow is, that in computer science where information is digitalized, information is a discrete entity. For example the chess move "Qxd4" (e.g. as move in the center game of a Danish Gambit) in the domain of playing chess is one piece of information $i \in \mathbb{I}$ in the domain of chess $\mathbb{D}_{chess}$. Where $\mathbb{I}$ is the amount of information available and $\mathbb{D}$ is the formal description of a domain as a proper subset of the information space $\mathbb{I}$. Consequently, a domain $\mathbb{D}$ contains those information necessary to create fully observable planing for the given domain. The following definition express what a reasoner is:
As illustrated in the boat example, the quality of explanations depends on the reasoner how infers about this explanation. Therefore we first need to define what a reasoner is.

Definition 3. A reasoner r is defined as an entity which includes new information $i \in \mathbb{I}$ into its knowledge-base $I_r \in \mathscr{D}$ where $\mathscr{D}$ is a σ-Algebra over $\mathbb{D}$.

This does not mean that all elements of $\mathbb{D}$ are available to each reasoner r. This offers the advantage that reasoners are able to infer in both fully and partial observable problems. To elude the problem of domain overarching knowledge, we define a domain as a σ-Algebra introducing the characteristic that all unions of information of one domain with information of the same domain is always part of the domain again. Later on we will utilize this and other characteristics of $\sigma - algebras$ to define a measure for explanations. Now, let $\mathscr{R}$ be a $\sigma - algebra$ of sets over all reasoners of concern where $r \in \mathscr{R}$. Further let $e \in \mathbb{E}$ be an explanation in some domain $\mathbb{D}$. Then we can define how an explanation maps to information by defining how the information in an explanation is transfered to the audience as follows:

Definition 4. $e \underset{r}{\rightarrow} i$ the explanation $e \in \mathbb{E}$ holds information $i \in \mathbb{I} \Leftrightarrow \exists r \in \mathbb{R}$ which is able to integrate i into its knowledge-base $I_r \in \mathscr{D}$ with the observation of e.

With this definition an explanation holds and transmits information to an audience if a reasoner of the audience can integrate new information into its knowledge-base. To avoid a philosophical discussion, we define that an explanation has to be understood by someone. Now that we have some definitions about a explanation, we will look at self-explanation to determine more specifically what exactly a explanation is.

The dictionary defines *self-explanatory* as: “easily understood from the information already given and not needing further explanation” [1]. This definition leads to the conclusion that the information given by self-explaining descriptions is sufficient for some reasoner in the audience to understand the SOI and that the explanation is given by the SOI. Taking this definition into account, we define a *degree of explanation* as follows:

Definition 5. Let $\mu : \mathscr{E} \to \bar{\mathbb{R}}$ be a measure with the σ-algebra $\mathscr{E}$ over $\mathbb{E}$ as some explanations to a affine extension of the real numbers $\bar{\mathbb{R}} := \mathbb{R} \cup \{+\infty, -\infty\}$ with $\mu(E) := \forall i \subseteq I \mid sup_{r \subseteq \mathbb{R}}(\sum_{\forall e \subseteq E : e \xrightarrow[r]{} i} (\delta_{i,i_r}))$. With $\delta_{i,i_r} = \begin{cases} 1, \text{ if } e \xrightarrow[r]{} i \text{ and } i_r \cup i \neq i_r \\ 0, \text{ if } e \xrightarrow[r]{} i \text{ and } i_r \cup i = i_r \\ -1, \text{ else} \end{cases}$

Where $E \in \mathscr{E}$ is the set of explanation, $R \in \mathscr{R}$ is the audience observing the explanation and $I \in \mathscr{D}$ is the set of pieces of information which should be explained in this domain. We acknowledge, that this is a practical measure, since the degree of explanations drops when a explanation is repeated in front of the same audience several times. Further we chose the supremum instead of an average since for a scientific ”proof of concept” we need one reasoner able to reason upon the explanation. With this definition of a measure for the degree of explanation, we can conclude that a theoretical complete self-explaining explanation with $\mu(E) =\mid \mathbb{D} \mid$ for some explanation $E \in \mathscr{E}$ could exist, so that no other explanation $e_i \in \mathscr{E}$ could explain the information i better to the audience. In practice such an explanation misses an example. But for a specific domain $\mathbb{D}$, an explanation e might be self-explanatory if the information space of $I_{r_1}, \ldots, I_{r_n} \in \mathscr{D}$ of the audience $r_1, \ldots, r_n, n \in \mathbb{N}$ of a domain d, is filled in that way, that the audience might reason to extract the entire information i hold by e by observing e.

On the one hand, the degree of self-explanation can be interpreted as the additional information needed to create understanding. On the other hand, as a measure depending on the reasoning capability of the audience and how the explanation fits to those capabilities. If no further information/capability is needed for some reasoner to understand the SOI, then the degree of self-explanation rises. The more information is needed the less the degree of self-explanation becomes, where in the worst case no useful information about the SOI can be extracted from the explanation.

In a domain, the information about the domain might be limited, and with that, the possibility for a good explanation might be given. To come back to our chess example the move: “Qxd4” probably needs further explanation. First we could explain the steno-notation syntax: The first element represent a chess piece here Q for the queen. The second element represents an optional action, here x which stands for making a capture and the last element *d4* concerning the location on the chess board where the move ends. Further we could additionally explain the meaning of ”queen” or ”making a capture”. If e.g. the audience has watched the move of the chess game, the first explanation of the move given above has $\mu(E) = 0$. Under the assumption that the audience of this explanation does not know the steno-notation, the first explanation of the move given above could have $\mu(E) = -1$. Because there is one explanation and it is not understood. Now the second more detailed explanation can be of hight

or lower quality. Since we have added multiple sub-explanations (Q,x,d4, queen and capture), if the audience still does not understand the explanation the measure of explanation can become, $\mu(E) = -6$. In this case all explanations did fail to transport information to the audience.
Further this explanation does not contain information about where the move started from, thus not being completely self-explanatory, since this depends on the context of the chess game. As we argued above, such contextual information is needed in self-explanatory descriptions for the effect of this example move. This reaches in the explanation of reasoning (the description of effect) and thus is out of scope of this work.

5 Conclusion

We can conclude that an explanation e transports information i to an audience of reasoners. The quality of an explanation can be measured by how much information the audience can extract from the explanation. So far, we define that an explanation becomes of higher quality if the degree of explanation rises. Our future work will be concerned with properties of explanation, in the attempt to make the definition more tangible. Further we want to integrate the existing structures of explanations like BDI and IOPE into explanations, to become able to represent an measure of self-explanation for existing descriptions.

References

1. Cambidge dictionary online (2012). URL http://dictionary.cambridge.org/dictionary/british/self-explanatory?q=self-explanatory
2. Braubach, L., Pokahr, A., Moldt, D.: Goal representation for bdi agent systems. multiagent systems pp. 44–65 (2005). URL http://www.springerlink.com/index/NYNBB1658EAK50GB.pdf
3. Cheng, B.H., Lemos, R., Giese, H., Inverardi, P., Magee, J., Andersson, J., Becker, B., Bencomo, N., Brun, Y., Cukic, B., Marzo Serugendo, G., Dustdar, S., Finkelstein, A., Gacek, C., Geihs, K., Grassi, V., Karsai, G., Kienle, H.M., Kramer, J., Litoiu, M., Malek, S., Mirandola, R., Müller, H.A., Park, S., Shaw, M., Tichy, M., Tivoli, M., Weyns, D., Whittle, J.: Software engineering for self-adaptive systems: A research roadmap. In: B.H. Cheng, R. Lemos, H. Giese, P. Inverardi, J. Magee (eds.) Software Engineering for Self-Adaptive Systems, pp. 1–26. Springer-Verlag, Berlin, Heidelberg (2009). DOI 10.1007/978-3-642-02161-9_1. URL http://dx.doi.org/10.1007/978-3-642-02161-9_1
4. Chi, M.T.: Advances in Instructional Psychology, vol. 5, chap. Self-explaining expository texts: The dual processes of generating inferences and repairing mental models, pp. 161–238. Routledge (2000)
5. Cole, W.G.: Understanding bayesian reasoning via graphical displays. SIGCHI Bull. **20**(SI), 381–386 (1989). DOI 10.1145/67450.67522. URL http://doi.acm.org/10.1145/67450.67522
6. Druzdzel, M.J.: Qualitative verbal explanations in bayesian belief networks. Artificial Intelligence and Simulation of Behavior Quarterly **94**, 43–54 (1996)

7. Duda, R.O., Stork, D.G., Hart, P.E.: Pattern classification and scene analysis. Part 1, Pattern classification, 2 edn. Wiley (2000)
8. Elsaesser, C.: Explanation of probabilistic inference. In: L.N. Kanal, T.S. Levitt, J.F. Lemmer (eds.) UAI, pp. 387–400. Elsevier (1987)
9. Grüninger, M., Hull, R., McIlraith, S.: A short overview of flows: A first-order logic ontology for web services. Data Engineering p. 3 (2008)
10. Heckerman, D.E., Horvitz, E.J., Nathwani, B.N.: Toward normative expert systems: Part i. the pathfinder project. In: Methods of information in medicine, vol. 31, pp. 90–105 (1992). URL http://www.ncbi.nlm.nih.gov/pubmed/1635470
11. Hinchey, M.G., Sterrit, R.: Self-managing software. IEEE Computer **39**(2), 107–109 (2006)
12. Jennings, N.R.: An agent-based approach for building complex software systems. Communications of the ACM, Forthcoming **44**(4), 35–41 (2001)
13. Kaddoum, E., Raibulet, C., George, J.P., Picard, G., Gleizes, M.P.: Criteria for the evaluation of self-* systems. In: Proceedings of the 2010 ICSE Workshop on Software Engineering for Adaptive and Self-Managing Systems, SEAMS '10, pp. 29–38. ACM, New York, NY, USA (2010). DOI 10.1145/1808984.1808988
14. Kephart, J.O.: Autonomic computing: The first decade. In: Proceedings of the 8th ACM international conference on Autonomic Computing, ICAC '11, pp. 1–2. ACM, New York, NY, USA (2011). DOI 10.1145/1998582.1998584. URL http://doi.acm.org/10.1145/1998582.1998584
15. Kephart, J.O., Chess, D.M.: The vision of autonomic computing. Computer **36**(1), 41–50 (2003). DOI http://dx.doi.org/10.1109/MC.2003.1160055
16. Klusch, M., Küster, U., Leger, A., Martin, D., Paolucci, M.: 4th international semantic service selection contest - performance evaluation of semantic service matchmakers (2010). URL http://www-ags.dfki.uni-sb.de/~klusch/s3/s3c-2010-summary-report-v2.pdf. Last visited: 2013-01-11
17. Leake, D.B.: Goal-based explanation evaluation. Cognitive Science **15**(4), 509–545 (1991)
18. Leake, D.B.: Evaluating Explanations A Content Theory. Psychology Press (1992)
19. Martin, D., Paolucci, M., McIlraith, S., Burstein, M.: Bringing Semantics to Web Services: The OWL-S Approach. In: First International Workshop on Semantic Web Services and Web Process Composition (SWSWPC 2004)., pp. 26–42. Springer-Verlag Berlin Heidelberg 2005, Berlin, Heidelberg (2004). URL http://www.springerlink.com/index/rl5r1c8v64xvf0r8.pdf
20. Morris, C.: Foundations of the Theory of Signs, vol. 1. University of Chicago Press (1938)
21. Müller-Schloer, C.: Organic computing – on the feasibility of controlled emergence. In: A. Orailoglu, P.H. Chou (eds.) Proceedings of the 2nd IEEE/ACM/IFIP International Conference on Hardware/Software CoDesign and System Synthesis, CODES+ISSS04, pp. 2–5. ACM, New York, NY, USA (2004)
22. Müller-Schloer, C., Schmeck, H.: Organic computing: A grand challenge for mastering complex systems. it – Information Technology **52**(3), 135–141 (2010). DOI 10.1524/itit.2010.0582. URL http://www.oldenbourg-link.com/doi/abs/10.1524/itit.2010.0582
23. Salehie, M., Tahvildari, L.: Self-adaptive software: Landscape and research challenges. ACM Transactions on Autonomous and Adaptive Systems **4**(2), 1–42 (2009). DOI 10.1145/1516533.1516538. URL http://doi.acm.org/10.1145/1516533.1516538
24. Sooriamurthi, R., Leake, D.: Towards situated explanation. In: In Proceedings of the Twelfth National Conference on Artifical Intelligence, p. 1492 (1994)
25. Sycara, K., Klusch, M., Widoff, S., Lu, J.: Dynamic service matchmaking among agents in open information environments. SIGMOD Record **28**, 47–53 (1999)