Trustworthy “blackbox” Self-Adaptive Systems

Beatriz Cabrero-Daniel[1], Yasamin Fazelidehkordi[1] and Olga Ratushniak[1]

[1] *University of Gothenburg, Hörselgången 5, 417 56, Göteborg, Sweden*

Abstract

For humans to trust Self-Adaptive Systems in critical situations, they must be robust, ethical, and lawful, but human intelligence is still needed to make ethical decisions. This paper presents a framework to discuss human values in the RE process for Self-Adaptive Systems and RE-specific challenges arising due to the AI paradigm shift towards foundation models: self-supervised blackboxes. Semi-autonomous heavy mining vehicles are a running example to present the requirements.

Keywords

Trustworthy AI, Human Oversight, Autonomous Vehicles.

1. Introduction

There is much public discussion on how Artificial Intelligence (AI) differs from human intelligence. We trust the latter, we are wary of the former. Industry practitioners share these concerns and put effort into measuring safety, privacy, etc. Their goal is ensuring AI-based Self-Adaptive Systems (SAS) can at least reach human performance in the tasks they were designed for [1]. However, these efforts are often insufficient for humans to trust SASs, especially with the introduction of foundation models, such as OpenAI’s ChatGPT, rapidly permeating society.

Foundation models are based on large-scale self-supervised deep learning algorithms [2], whose inner workings are not transparent, making them difficult to explain to and interpret by users. Moreover, foundation models often use large amounts of unlabelled data, often gathered disregarding ethical concerns, e.g., diversity. The more complex and accurate the models become, the more data is needed to train them, and the harder it is to explain their decision making process. Thus, the conflict between these powerful AI “blackboxes” and user trust [3].

Requirements Engineering (RE) guidelines for ethical AI were reviewed with the aim of building a framework for Trustworthy SASs (T-SASs). The outlined T-SAS framework is motivated by the emergence of semi-autonomous heavy vehicles for mining, as running example, which raise concerns addressed here. Nevertheless, the T-SAS framework could address human values in other fields. The focus will be on *human oversight*, still needed to promote trust in SASs [4, 5, 6]. The insights on human-on-the-loop (HOTL) expectations for T-SAS monitoring

In: A. Ferrari, B. Penzenstadler, I. Hadar, S. Oyedeji, S. Abualhaija, A. Vogelsang, G. Deshpande, A. Rachmann, J. Gulden, A. Wohlgemuth, A. Hess, S. Fricker, R. Guizzardi, J. Horkoff, A. Perini, A. Susi, O. Karras, A. Moreira, F. Dalpiaz, P. Spoletini, D. Amyot. Joint Proceedings of REFSQ-2023 Workshops, Doctoral Symposium, Posters /& Tools Track, and Journal Early Feedback Track. Co-located with REFSQ 2023. Barcelona, Catalunya, Spain, April 17, 2023.

beatriz.cabrero-daniel@gu.se (B. Cabrero-Daniel)
0000-0001-5275-8372 (B. Cabrero-Daniel)

and human intervention aim to foster discussions among the RE practitioners about creating T-SASs that adhere to ethical principles and laws [4, 7].

2. Background and Mining Context

Aristotle defined credibility in terms of *wisdom*, *virtue*, and *goodwill*. Centuries later, EU guidelines state that AI should be trustworthy, that is *robust*, *lawful*, and *ethical* [4]. Fig.1 shows requirements related to human autonomy and shared responsibility in EU guidelines. Evaluating whether adaptive systems meet stakeholders' needs often focuses on robustness verification, but this may not capture ethical values [1, 8, 9]. Nevertheless, embedding ethical values in SASs is challenging, partly due to the recent AI developments such as foundation models, e.g., text-to-image generators for non-expert users [10, 2].

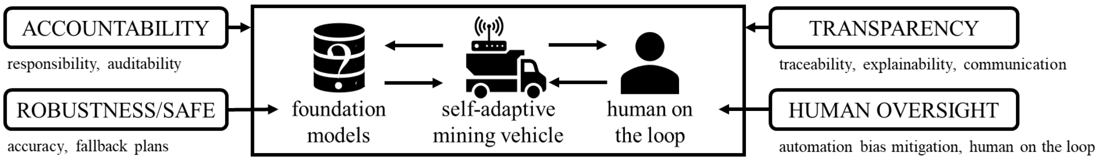
ACCOUNTABILITY
responsibility, auditability
ROBUSTNESS/SAFE
accuracy, fallback plans
foundation models
self-adaptive mining vehicle
human on the loop
TRANSPARENCY
traceability, explainability, communication
HUMAN OVERSIGHT
automation bias mitigation, human on the loop

Figure 1: Framework for trustworthy SAS using opaque self-supervised AI models.

Designing comprehensive evaluation strategies for these complex and industrial systems is difficult due to the lack of auditability and sustainability analysis, and the emergence of unforeseen skills during training [2]. Moreover, the lack of open APIs and benchmarks hinders research on foundation models' transparency, robustness, fairness, etc. Moreover, the resources needed to train and test such systems hinder academics' access to evaluating their benefits and harms [2]. Nevertheless, high-risk SASs like Autonomous Vehicles (AV), potentially using foundation models, must nevertheless show transparency to allow for human oversight and intervention [3, 4]. SASs must inform diverse end-users, e.g., end-users or third-party audits, about their capacities and limitations and trace them back to input data to enable responsibility reasoning [11, 12]. Responsibility sharing and mitigation of foreseeable misuse are challenging and raise ethical questions that need to be answered during the RE process [3, 13].

Mining AVs in safety-critical situations are high-risk AI products, therefore a HOTL to monitor the AVs and intervene when prompted is needed [3, 4]. Human drivers and AVs primarily rely on vision, or Computer Vision (CV), to avoid danger and their responsibilities must be balanced [14, 15]. AI algorithms can help mining vehicles remote operators in critical situations: by measuring user attention, either driver or remote operator, to reduce reaction times or by facilitating fallback to human control in case of low AI confidence [16, 7]. Even HOTL AVs can be involved in incidents, potentially fatal with heavy mining machinery, so risks arising from faulty interactions must be mitigated. Human-AI interaction is receiving increasing academic attention together with limitations of AV, including benefits, harms, and development practices [11, 17, 18]. The AI paradigm is shifting to blackbox models, hindering HOTL-SAS interaction and raising the question of how to split the responsibility of decision-making.

Deep Learning algorithms are increasingly popular to detect edge cases where human intervention might be needed, but they rely on large amounts of annotated data, which is difficult or impossible to gather, expensive and time-consuming to curate [2]. However, sensor difficulties, e.g., extreme weather affecting visibility, or cognitive limits, e.g., insufficient training data, might cause malfunctions [19]. The RE process therefore needs to set standards for data quality, security, and privacy [10]. Based on the data, robustness needs to be periodically evaluated by stakeholders, using performance metrics and criteria that reflect their values and goals, e.g., ore throughput rate [1, 10]. Limitations of mining AVs should be clearly explained to the HOTL at all times, e.g., to prevent incidents, improve throughput rates, or audit accidents [12, 7]. Transparency, though, is not always possible when using these algorithms, especially in opaque blackbox algorithms or foundation models.

3. Framework for Trustworthy Self-Adaptive Systems

This section outlines a framework to guide the RE process for T-SAS focusing on requirements for HOTL-mechanisms (see Figure 1) in light of the trend to incorporate foundation models such as GPT-3, DALL-E, or BERT [20]. The relationship between the concepts is also discussed:

Robustness. Classic AIs use annotated data, whilst foundation models use large volumes of unlabeled data, removing the difficult and time-consuming task of curating data sets. This paradigm can particularly benefit AVs for mining, which inherently need to deal with previously unseen scenarios. Nevertheless foundation models, especially learning online, can be affected by incorrect, redundant, or unstable data, which could lead to safety-critical situations. Therefore, the T-SAS framework promotes the usage of high-quality, diverse, self-updating, and self-augmenting data sets [21, 4]. Appropriate requirements for data availability, usability, consistency, and integrity, must be discussed [2, 1].

Human oversight. Whilst foundation models can accomplish complex tasks, e.g., image synthesis, they still show limitations, e.g., generalizing to new scenes, mainly due to self-supervised training [2]. Even if totally reliable, SASs incorporating such models would still need to be transparent to facilitate human oversight, foster human autonomy, and, ultimately, be trustworthy. HOTL-SAS interaction is an open and important problem for humans, who should be able to supervise and override SAS decisions at all times. Therefore, T-SASs must integrate HOTL strategies and monitoring interfaces adequate to the end-users, designed to address the transparency and accountability needs of T-SASs [17, 7, 10].

Transparency. T-SASs should provide concise, complete, correct, and clear explanations that are relevant, accessible and comprehensible to users in a context (use or foreseeable misuse), to avoid risks to health, safety, or fundamental rights [4, 3]. These requirements intend to ensure *human autonomy* and *responsibility sharing* but integrating these needs into SAS is challenging. Previous work has focused on highly trained operators, e.g., aircraft pilots, but there is still the need to investigate how to design interactions with non-expert users [11]. Training end-users while using SASs could be considered. For that, appropriate metrics and criteria, adapted to the user and the operation context, would be needed to ensure clarity and avoid ambiguity about the state of the T-SAS.

Accountability. As discussed above, many SASs, including AV, cannot ensure safety on their own and need to be monitored by humans during operation. Even when SASs are not entirely robust, might be able to produce priors and convey information that greatly helps the HOTL in critical situations. This has long been a focus of Human-Computer Interaction research [3, 17, 7]. Moreover, T-SASs must also be accountable to justify their goals, motivations and rationale in *post hoc* analysis by third parties. This topic is strongly related to detecting, leveraging, and mitigating risks by public authorities. Therefore, the framework should explicitly connect these needs to open communication requirements, critical for T-SASs that closely interact with humans, e.g., AV drivers [4].

4. Conclusion

Humans often mistrust SASs or show automation bias [3, 11]. Both are concerning as SASs increasingly integrate foundation models, far from being transparent or auditable [20, 6, 22, 2]. Much effort has been devoted to support practitioners in addressing human values in the RE process but the absence of clear guidelines, benchmarks, metrics, and evaluation criteria, makes this task challenging. As a result, there is still a need for human oversight, e.g., fallback procedures [11, 17, 16]. Academics from different backgrounds should examine the models' biases and limitations, and inform society about their trustworthiness [2]. These recommendations are based on existing international laws, domestic legislation, and AI development frameworks and aim to increase awareness among RE practitioners and inspire the development of a generic framework for creating T-SASs.

Efforts to homogenise mining processes are already being made but further research is needed to adequately address human values in HOTL mining SASs. For instance, it is necessary to consider the implications that foundation models will entail with respect to other ethical considerations. Agreeing on appropriate recommendations with practitioners to address human values in the RE process for T-SAS would be a necessary next step. Frameworks from other disciplines and the ad-hoc practices of RE practitioners could be studied to propose adaptations to existing frameworks to better address human values in T-SAS development. Data governance should in turn be aligned with stakeholders' values, e.g., non-discrimination, and requirements such as privacy or fairness. These considerations are left out for future work.

This work is based on European Union guidelines but different values might prevail in non-EU countries. Even within the EU, revisions to the AI legislation, which is still in draft form, might have a significant impact on the SAS now in development. As such, it is important for the framework to adapt to new, unforeseeable trust elements introduced by public authorities that might, directly and indirectly, impact the expectations for T-SASs. As a final note, future research must also address the question of how to allow for diverse legislation and context-dependent interpretation of T-SAS requirements.

Acknowledgments

This work is thanks to the University of Gothenburg's Amanuens program. Thanks to Prof. Berger and Assoc. Prof. Horkoff for their valuable guidance. This work was supported by the Vinnova project ASPECT [2021-04347].

References

[1] D. M. Berry, Requirements engineering for artificial intelligence: What is a requirements specification for an artificial intelligence?, volume 13216 LNCS, Springer Science and Business Media Deutschland GmbH, 2022, pp. 19–25. doi:10.1007/978-3-030-98464-9_2.

[2] R. Bommasani, et al., On the opportunities and risks of foundation models, 2021. URL: https://arxiv.org/abs/2108.07258. doi:10.48550/ARXIV.2108.07258.

[3] EUR-Lex - 52021PC0206 - EN - EUR-Lex, 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

[4] European Commission and Directorate-General for Communications Networks, Content and Technology, The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment, Publications Office, 2020. doi:doi/10.2759/002360.

[5] R. Calinescu, D. Weyns, S. Gerasimou, M. Iftikhar, I. Habli, T. Kelly, Entrust: engineering trustworthy self-adaptive software with dynamic assurance cases, 2018, pp. 495–495. doi:10.1145/3180155.3182540.

[6] M. Rahimi, J. L. Guo, S. Kokaly, M. Chechik, Toward requirements specification for machine-learned components, IEEE, 2019, pp. 241–244. doi:10.1109/REW.2019.00049.

[7] J. Dimatteo, D. M. Berry, K. Czarnecki, Requirements for monitoring inattention of the responsible human in an autonomous vehicle: The recall and precision tradeoff (2020). URL: https://ceur-ws.org/Vol-2584/RE4AI-paper2.pdf.

[8] E. Halme, M. Agbese, J. Antikainen, H.-K. Alanen, M. Jantunen, A. A. Khan, K.-K. Kemell, V. Vakkuri, P. Abrahamsson, Ethical user stories: Industrial study (2022). URL: http://ceur-ws.org.

[9] F. B. Aydemir, F. Dalpiaz, A roadmap for ethics-aware software engineering (2018). URL: https://doi.org/10.1145/3194770.3194778. doi:10.1145/3194770.3194778.

[10] I. Ozkaya, Ethics is a software design concern, IEEE Software 36 (2019) 4–8. doi:10.1109/MS.2019.2902592.

[11] A.-Q. V. Dao, S. L. Brandt, V. Battiste, K.-P. L. Vu, T. Strybel, W. W. Johnson, The impact of automation assisted aircraft separation on situation awareness, in: G. Salvendy, M. J. Smith (Eds.), Human Interface and the Management of Information. Information and Interaction, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 738–747.

[12] N. E. Gold, Virginia dignum: Responsible artificial intelligence: How to develop and use ai in a responsible way, Genetic Programming and Evolvable Machines 2020 22:1 22 (2020) 137–139. URL: https://link.springer.com/article/10.1007/s10710-020-09394-1. doi:10.1007/S10710-020-09394-1.

[13] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, K. Scott, S. Schieber, J. Waldo, D. Weinberger, et al., Accountability of ai under the law: The role of explanation, arXiv preprint arXiv:1711.01134 (2017).

[14] Y. Chen, Y. Tian, M. He, Monocular human pose estimation: A survey of deep learning-based methods, Computer Vision and Image Understanding 192 (2020) 102897. doi:10.1016/J.CVIU.2019.102897.

[15] A. Bulat, J. Kossaifi, G. Tzimiropoulos, M. Pantic, Toward fast and accurate human pose estimation via soft-gated skip connections, Proceedings - 2020 15th IEEE International

Conference on Automatic Face and Gesture Recognition, FG 2020 (2020) 8–15. doi:10.1109/FG47880.2020.00014.

[16] I. Kotseruba, J. Tsotsos, Attention for vision-based assistive and automated driving: A review of algorithms and datasets, IEEE Transactions on Intelligent Transportation Systems (2022) 1–22. doi:10.1109/TITS.2022.3186613.

[17] C. Mutzenich, S. Durant, S. Helman, P. Dalton, Updating our understanding of situation awareness in relation to remote operators of autonomous vehicles, Cognitive Research: Principles and Implications 6 (2021) 1–17. doi:10.1186/S41235-021-00271-8/FIGURES/6.

[18] N. Hutchins, Z. Kirkendoll, L. Hook, Social impacts of ethical artifical intelligence and autonomous system design, 2017 IEEE International Symposium on Systems Engineering, ISSE 2017 - Proceedings (2017). doi:10.1109/SYSENG.2017.8088298.

[19] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, 2d human pose estimation: New benchmark and state of the art analysis, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2014) 3686–3693. doi:10.1109/CVPR.2014.471.

[20] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding (2018). URL: https://arxiv.org/abs/1810.04805. doi:10.48550/ARXIV.1810.04805.

[21] M. Borg, H.-M. Heyn, J. Horkoff, K. M. Habibullah, A. Knauss, E. Knauss, P. J. Li, Precog: Requirements Engineering toward Safe Machine Learning-Based Perception Systems for Autonomous Mobility | Vinnova, 2021. URL: https://www.vinnova.se/en/p/precog-requirements-engineering-toward-safe-machine-learning-based-perception-systems-for-autonomous-mobility/.

[22] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, Manipulating and measuring model interpretability, Conference on Human Factors in Computing Systems - Proceedings (2021). doi:10.1145/3411764.3445315.