Ethical Guidelines for Trustworthy AI Systems

Zahoor ul Islam [1] and **Andreas Theodorou** [2] and **Juan Carlos Nieves** [3] and **Virginia Dignum** [4]

Abstract.

1 INTRODUCTION

In order to engineer human-centric Artificial Intelligence (AI) and autonomous systems, comply with societal norms, ethical guidelines, and established standards, a well-established set of the development life cycle is needed. This development life cycle requires a continuous evaluation process for continuously evolving AI systems [9]. Furthermore, a strategy is required to initiate human responsibility in the development of AI systems from the start of the life cycle, address the gaps that emerge from the increased automation of decision, and provide tools to fill the gaps. [25]. A possible way to develop ethical guidelines for trustworthy AI systems from a development perspective is to look into software engineering (SE) domain. Software Development life cycle (SDLC) is defined by IEEE as "the process by which user needs are translated into a software product. The process involves translating user needs into software requirements, transforming the software requirements into design, implementing the design in code, testing the code, and sometimes, installing and checking out the software for operational use." [13]. Some of the most commonly used SDLC models is the *waterfall*, *incremental*, *prototype-driven*, *evolutionary*, *spiral* and *agile software development methods* [23]. Most SDLCs processes are available in all methodologies but applied and practised differently based on projects, problems and personal needs.

The rapid development in machine learning and neural networks has enabled machines and algorithms to effectively manage tasks such as natural language processing, translations, stock market predictions, and route planning and optimisation. AI systems can learn quickly from patterns and propose decisions and in some instances, autonomously take decisions but without paying in particular attention to the implications of those decisions. Therefore, applying SE methodologies is a fundamental prerequisite for delivering high quality, responsible, transparent, trustworthy, accountable, and robust smart software applications.

A well-established set of design methodologies has the potential to address many of the challenges of designing high-performing trustworthy intelligent systems. It can provide an explicit process for values elicitation and stakeholder involvement in the development of AI systems [9]. It can provide support for effective communication, re-usability, process enhancement, and process management. This methodology can help to maintain explicit formal links between values, norms, and systems functionalities that enable adaptation of the system to evolve perception and justification of implementation decisions in term of their underlying values [9]. It can provide support to choose system components based on their underlying societal and ethical conceptions [8]. These development methods are required for the explainability of black-boxes, identify and eradicate biases in training data, solve adversarial issues (slight changes in training data that could have serious implications), testing and formal verification of AI systems in terms of transparency, fairness, and accountability. These black boxes could have severe implications on society if not appropriately handled by Artificial Intelligent design methods [24]. Methods to introduce ethical values and question system reasoning must be included in the early design phase of AI systems [10].

Many high-level ethical guidelines have been proposed in the last few years, but often these offer no concrete governance mechanism or solutions to integrate ethical values in the architecture or the design and development process of AI systems [25]. The European Commission (EC) in June 2018 formed a High-level Expert Group on Artificial Intelligence (HLEG AI) to specify guidelines for trustworthy AI. These guidelines set out a framework for achieving trustworthy AI in Europe by setting out its vision for AI which support ethical, secure and cutting-edge AI made in Europe. The HLEG AI has published two documents: (1) AI ethics Guidelines; and (2) policy and investment recommendations[5].

In order to implement the vision, it is vital to understand the existing AI systems engineering practices by AI4EU partners. Thus, surveying existing development methods to explore how different processes are carried out in terms of designing, development, frameworks, standards and certifications, AI-specific architectures, testing, verification and validation, documentation, auditing, governance, ethical, and, legal compliance is important. It is also necessary to get a feedback for a proposed methodology regarding different areas of AI that would benefit from standard definitions and taxonomies, contents ethical guidelines on a scale of a low, medium, high, must include and not important. We design an online survey to understand what AI4EU partners already do and on what they think that they should be doing. The objective of the survey was to get a consensus on existing development practices for AI systems and expected requirements for the proposed methodology.

This paper is organised as follows. The next section will examine the related work. The research methodology section describes the design method for this survey. The result section outline results based on the classification of development methods and expected requirements for the proposed methodology for ethically aligned AI systems in industrial and academic context. Discussion section elaborate need for a methodology and how the posed challenges must be addressed with some future directions and conclusion section conclude the work by discussing some key findings regarding existing methods and future methodology.

[1] Umeå University, email: zahoor.ul.islam@umu.se
[2] Umeå University, email: andreas.theodorou@umu.se
[3] Umeå University, email: juan.carlos.nieves@umu.se
[4] Umeå University, email: virginia.dignum@umu.se

[5] See `https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence`.

2 RELATED WORK

AI systems are software artefacts that can get benefit from structured methods for design and development like any other traditional software using well-structured SE design methods [9]. SE is "a systematic, disciplined, quantifiable approach to the analysis, design, assessment, implementation, test, operation, maintenance and development of software, that is, the application of engineering to software. In the SE approach, several design methods and models for the software life cycle are defined, and many methodologies for the definition and assessment of the different phases of a life-cycle model are applied " [13, 20]. One essential category of SE is software development process by which requirements are elicited from stakeholders, transformed into design requirements, implement the design in code, and testing the code, and the process is called software development life cycle. Software processes are specified to facilitate human understanding, communication, and coordination; to aid management of software projects; to measure and improve the quality of software products in an efficient manner; to support process improvement; to provide a basis for automated support of process execution [2].

However, existing traditional methodologies do not fulfil the requirements for the complex structure of AI systems where the environment is multi-dimensional and heterogeneous [1]. Software designing is a challenging phase of creating intelligent agents that is why optimizing agent architecture as a design tool is of utmost importance [3]. Software designing can be related to abstraction, patterns, separation of data, modulating, functional independence, and refactoring. Several methodologies and processes are designed, developed and proposed for developing AI systems based on different approaches, mainly extending existing object-oriented and knowledge engineering methodologies. Due to a considerable number of challenges faced by AI design technologies, these methods have been re-designed, re-engineered, altered, and updated significantly over the years.

Several methodologies have been developed specifically for intelligent multi-agent systems, including the Multi-Agent System Engineering (MASE) [6], The Gaia methodology for agent-oriented analysis and design [27], a methodology and modelling technique for systems of belief-desire and intention (BDI) agents [18], and OperA, a methodology to design multi-agent organisations [7] Apart from developing methodologies from scratch, researchers have been extending existing methodologies to fit agent designing mainly in two areas object-oriented methodologies and knowledge engineering methodologies [14]. Agent-oriented analysis and design (AOAD) [4], agent modelling technique for systems of BDI agents [17] and multi-agent scenario-based methods(MASB) [22, 21] and agent-oriented methodology for enterprise modelling [16] combining integration definition for function modelling (IDEF) [11] and computer integrated manufacturing open system architecture (CIMOSA) [19] are some of the earlier extensions of object-oriented methodologies. At the same time, CoMoMAS [12], MAS-CommonKADS [15] are an extension of knowledge engineering methodologies. Some other approaches, including the methodology of Cassiopeia [5], and the cooperative information agents design proposed by Verharen [26], have been tested.

Most works focus on developing systems that can incorporate societal and organizational concepts at large, mostly combining essential characteristics of intelligent agents, multi-agent systems and their relation to SE. Methodologies like Gaia and OperA introduced conceptual frameworks that allow system analyst to systematically design agent model, services model, and acquaintance model from a detailed set of requirements and analysis stage where roles model and interaction models are developed [27, 7]. However, recent developments in AI, grounded on machine learning methods, that result on stochastic, non-monotonic models, are not sufficiently covered by these methodologies. At the same time, as we will describe in the remainder of this paper, AI developers do not sufficiently use formal design methodologies on their work.

3 METHODOLOGY

Understanding and exploring various types of development environment for AI both in academic and industrial domains is essential to design a methodology that could integrate and foster ethical and legal values. There is a broad spectrum in terms of the types and maturity of AI system development conducted in research and industrial environment. A survey is conducted to characterise, analyse, and identify the different types of system development practices conducted in academic research and industrial environment. The survey will help to identify needs for different AI systems design and development approaches from a software development life cycle perspective.

To achieve this objective, we systematically designed an online survey with forty-six questions covering the complete life cycle: requirements, design, implementation, quality assurance (testing), auditing and compliance, deployment, and maintenance. Results are constructed based on two main categories academic and industrial context. Eighteen participants participated in the survey from academic and industrial domains. Nine are affiliated with universities and academic research institute working as an academic researcher. At the same time, nine are from industry(Large-sized companies, small-sized companies including startups and Non-Governmental Organisations (NGO's) working mostly as AI developers, engineers, architectures, and project, product managers on AI-related products and services.

4 RESULTS

The survey results reveal an important and compelling finding of the AI system development environment in an academic and industrial research and development environment. Among the findings are the following:

1. Academia is playing an equally vital role in the design and development of AI systems with totally different requirements from industry in many domains, including the production, deployment, and development.
2. In terms of different areas of AI that would benefit from standard definitions and taxonomies academia (100%) stated "Yes."
3. Scrum is a highly preferred method proposed by academia (44%) and industry (67%).
4. In terms of ethical guidelines to be included in the methodology: the top requirements from industry are documentation (89%) and standardised reporting (78%). At the same time, academia stated Challenger benchmarks (Production model benchmarking against other similar systems or purpose-made challenges, e.g. ImageNet) (89%) and Adversarial example testing (Systematically testing the model's behaviour on developer generated potential extreme / rare instances) (89%).
5. In terms of contents of the documentation, academia stated High-level description of the decision-making system used, e.g. methods (100%), Identified risks (100%). In contrast, industry stated

Requirements (89%), High-level description of the decision-making system used, e.g. methods (89%), and System boundaries (78%).

6. From the survey, it appeared that no AI-specific architecture is being used in the industry. The methodology to make an explicit recommendation on what architecture system should have, both academia and industry results stated "No" with (56%) average. In terms of methodology to integrate a specific architecture into the methodology, academia (78%) and industry (67%) stated "No".
7. Both academia and industry stated that methodology should include recommendations and instructions on why, how and when to use version control with (66%) and (78%) acceptance, respectively.
8. Respondents from academia (100%) agreed on limiting system performance to ensure legal and ethical compliance while industry agreed on legal compliance (100%) and ethical compliance (89%).
9. In terms of auditing, academia and industry are auditing mostly for legal compliance with (67%) and (66%) acceptance, respectively. In terms of auditing projects compliance with organisational policy academia stated (56%) acceptance ratio on requirements gathering while industry stated during development with (56%) acceptance ratio.
10. Using tools for verification and validation are non-existent in both academia (78%) and industrial (89%) practices.
11. In terms of keeping internal documentation for data governance in academia (56%) stated "Yes" while in the industry (56%) stated "No".
12. In terms of methodology to require specific standards to follow in academia (78%) stated "No" while in the industry (56%) stated "Yes". In terms of methodology to include procedures of software updates and provide a specification for testing both academia (89%) and industry (100%) stated "Yes" respectively.
13. In terms of producing reports of the effects of products, services, research output both in academic and research environment is mostly done when relevant, in academia (78%) stated for environmental. In contrast, the industry (78%) opted for Risk of misuse.
14. In terms of disclosure of system architecture technical details including a list of agent pre-defined goals academia stated it should be for all users (44%), for experts (44%) while industry stated it should be only for expert users with (56%) average. Disclosure of Indication of sensors location & capabilities both academia and industry with (56%) average agreed it should be for expert users only. Disclosure of agent using machine learning (including prior training) both academia and industry agreed with (56%) average it should be for expert users only. Disclosure if the agent has online learning academia stated (44%) for all and expert users while industry with (67%) opted for expert users only. Disclosure of any data filtering and manipulation used both academia and industry with (67%) average agreed it should be for expert users only. Disclosure of training Data both academia and industry with (56%) average agreed it should be for expert users only. Disclosure trained/production model's characteristics, including accuracy and tuning academia stated with (56%) average for all users while an industry with (56%) average stated it should be for expert users only and (44%) stated for none. Disclosure of security testing in academia (78%) stated it should be for expert users only while in industry average was (44%). Disclosure of identify risks of using or deployment academia stated it should be for all users with (67%) average while in the industry (56%) thinks the same while (44%) stated for none. Disclosure of testing conducted at high-level description academia stated with (44%) each for all users and expert users while industry stated with (67%) ratio for expert users only. Disclosure of testing and stress testing results, e.g. accuracy, tested uptime, etc. academia stated with (67%) average for expert users only while industry stated with (44%) average for expert users and none (44%) respectively.
15. In terms of methodology to contain sample tools/libraries, both academia and industry stated "Yes" with an average of (89%).

5 Discussion

Our survey exposed various shortcomings in existing development methods and how lack of structured software engineering life-cycles approaches can affect the development of high-quality, transparent, and trustworthy AI systems. On the other hand, the survey also identifies some key finding for future methodology and how various operational elements and metrics from a well-established software development process can lead different stakeholders of the system and their development activities towards trustworthiness. Based on recent advances in AI development methods and techniques, it appears that much work has focused on developing working prototypes and applications without focusing on adapting effective process and project management methods based on SDM. Furthermore, a well- structured evaluation policy is missing that can guide capturing norms, assign and implement governance mechanism at different levels. Effective verification and validation methods and tools are also required in order to align the ethical guidelines with the methodology. The survey highlighted some of the major processes that must be integrated into development methodology including effective management, communication, use of appropriate standards, sufficient documentation, a mechanism for version control, ethical and legal compliance mechanisms, the introduction of use cases, and finally the support of tools and libraries for policy evaluation and compliance mechanisms. Transparency seems to be a significant concern for both the industry and academia. Our survey identifies that different approaches will be required by different stakeholders when it comes to system disclosure in industry and academia. Mechanisms for evaluating policy compliance and system's compliance with organisational policy has appeared in our survey, but much work is needed in proposing ethical guidelines that evaluate these compliance mechanisms with organisational policy through a methodology.

From this survey, we envision several possible future research directions. Firstly industry and academia require different approaches, frameworks, and tools to integrate ethical guidelines and to comply with organisational policies.

Secondly, the methodology should introduce methods to have continuous evaluation and justification of the whole development process, not just a following sequence of steps in different stages of SDLC.

Thirdly, with AI being dynamic and adaptable, it is crucial to have the proper calibration of trust between human users and AI systems by introducing transparency mechanisms that disclose the decision making of a system [24].

Lastly, incorporating ethical guidelines for trustworthy AI systems with the development of methodology related to existing software engineering approaches can reduce risks of evaluation and justification issues in different stages of development life cycle by ensuring that societal and ethical values are central to the development process.

In order to adapt a methodology that could help ethical decision making by humans in developing and deploying trustworthy AI systems, more work on collecting data about various design and development processes both in an industrial and academic domain is re-

quired.

6 CONCLUSION

The results of our survey will support the development of a methodology for AI4EU platform. Our results will also help developing AI systems aligned with ethical and societal requirements, as well as to support the evaluation of the resources shared through the AI4EU platform in terms of ELS (ethical, legal, societal) alignment. It will also benefit researchers in observing the patterns for adapting an ethical methodology for different domains. Different areas of AI require standard definitions and taxonomies including fundamental technical terms(machine learning, cognitive architecture), low-level technical terms(reasoning, learning, functions, goals), guideline terms(transparency, explainability, accountability) Job roles and expertise (AI ethics, AI auditor, AI architecture). The methodology must be aligned with different existing development frameworks. Ethical guidelines must include early warning systems, challenger benchmarks, documentation, and adversarial example testing. Documentation must include requirements, high-level description of the decision-making systems, identified risks and system boundaries. The methodology should include instructions on why how and when to use version control. Based on survey results it is concluded that the survey raises several hypotheses that merit further research about the views of AI practitioners about ethical guidelines and methodology and outreach efforts to address concerns about integrating into design and development process of AI systems.

ACKNOWLEDGEMENTS

This work was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, and the H2020 project AI4EU.

REFERENCES

[1] Huib Aldewereld, Virginia Dignum, and Yao-hua Tan, 'Design for values in software development', *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, 831–845, (2015).

[2] *SWEBOK: Guide to the Software Engineering Body of Knowledge*, eds., Pierre Bourque and Richard E. Fairley, IEEE Computer Society, Los Alamitos, CA, version 3.0 edn., 2014.

[3] Joanna J Bryson and Lynn Andrea Stein, 'Modularity and design in reactive intelligence', in *International Joint Conference on Artificial Intelligence*, volume 17, pp. 1115–1120, (2001).

[4] Birgit Burmeister, 'Models and methodology for agent-oriented analysis and design', *Working Notes of the KI*, **96**(96-06), 52, (1996).

[5] Anne Collinot, Alexis Drogoul, and Philippe Benhamou, 'Agent oriented design of a soccer robot team', in *Proceedings of the Second International Conference on Multi-Agent Systems (ICMAS-96)*, pp. 41–47, (1996).

[6] Scott A DeLoach, Mark F Wood, and Clint H Sparkman, 'Multiagent systems engineering', *International Journal of Software Engineering and Knowledge Engineering*, **11**(03), 231–258, (2001).

[7] MV Dignum, *A model for organizational interaction: based on agents, founded in logic*, SIKS, 2004.

[8] Virginia Dignum, 'Responsible artificial intelligence: Designing ai for human values', *world*, **13**, 23, (2018).

[9] Virginia Dignum, *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*, Springer International Publishing, 2019.

[10] Virginia Dignum, Frank Dignum, Javier Vázquez-Salceda, Aurélie Clodic, Manuel Gentile, Samuel Mascarenhas, and Agnese Augello, 'Design for values for social robot architectures.', in *Robophilosophy/TRANSOR*, pp. 43–52, (2018).

[11] PUB FIPS, '183 (1984). federal information processing standards, united states national institute of standards and technology (nist)', *Computer Systems laboratory, Gaithersburg*, (1993).

[12] Norbert Glaser, 'Contribution to knowledge modelling in a multi-agent framework (the comomas approach)', *PhDthesis, Lniverstit Henri Poincar, Nancy I, France*, (1996).

[13] Ieee, 'IEEE Standard Glossary of Software Engineering Terminology', *Office*, **121990**(1), 1, (1990).

[14] Carlos A Iglesias, Mercedes Garijo, and José C González, 'A survey of agent-oriented methodologies', in *International Workshop on Agent Theories, Architectures, and Languages*, pp. 317–330. Springer, (1998).

[15] Carlos A Iglesias, Mercedes Garijo, José C González, and Juan R Velasco, 'Analysis and design of multiagent systems using mas-commonkads', in *International Workshop on Agent Theories, Architectures, and Languages*, pp. 313–327. Springer, (1997).

[16] Elisabeth A Kendall, Margaret T Malkoun, and Chong H Jiang, 'A methodology for developing agent based systems for enterprise integration', in *Modelling and Methodologies for Enterprise Integration*, 333–344, Springer, (1996).

[17] David Kinny, M Georgeff, and A Rao-A Methodology, 'Modeling technique for systems of bdi agents-in', in *Proc. of the 7th European Workshop on MAAMAW*, pp. 56–71.

[18] David Kinny, Michael Georgeff, and Anand Rao, 'A methodology and modelling technique for systems of bdi agents', in *European Workshop on Modelling Autonomous Agents in a Multi-Agent World*, pp. 56–71. Springer, (1996).

[19] Kurt Kosanke, 'Cimosa—overview and status', *Computers in industry*, **27**(2), 101–109, (1995).

[20] Philip A Laplante, *What every engineer should know about software engineering*, CRC Press, 2007.

[21] Bernard Moulin and Mario Brassard, 'A scenario-based design method and an environment for the development of multiagent systems', in *Australian Workshop on Distributed Artificial Intelligence*, pp. 216–232. Springer, (1995).

[22] Bernard Moulin and Louis Cloutier, 'Collaborative work based on multiagent architectures: A methodological perspective', in *Soft computing*, pp. 261–296. Prentice-Hall, Inc., (1994).

[23] Ieee Computer Society, *Guide to the Software Engineering Body of Knowledge Version 3.0 (SWEBOK Guide V3.0)*.

[24] A Theodorou, *AI Governance Through A Transparency Lens*, Ph.D. dissertation, University of Bath, 2019.

[25] Andreas Theodorou and Virginia Dignum, 'Towards ethical and socio-legal governance in AI', *Nature Machine Intelligence*, **2**(1), 10–12, (1 2020).

[26] Egon M Verharen et al., 'A language-action perspective on the design of cooperative information agents', Technical report, Tilburg University, School of Economics and Management, (1997).

[27] Michael Wooldridge, Nicholas R Jennings, and David Kinny, 'The gaia methodology for agent-oriented analysis and design', *Autonomous Agents and multi-agent systems*, **3**(3), 285–312, (2000).