Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Exploring Ethical Requirements Elicitation for Applications in the Context of AI

José Antonio Siqueira de Cerqueira

Dissertação apresentada como requisito parcial para conclusão do Mestrado em Informática

Orientadora
Prof.a Dr.a Edna Dias Canedo

Brasília
2021

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Exploring Ethical Requirements Elicitation for Applications in the Context of AI

José Antonio Siqueira de Cerqueira

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Prof.a Dr.a Edna Dias Canedo (Orientadora)
CIC/UnB

Prof.a Dr.a Carla Taciana Lima Lourenço
Centro de Informática/UFPE

Prof. Dr. Vander Ramos Alves
CIC/UnB

Prof.a Dr.a Genaína Nunes Rodrigues
Coordenadora do Programa de Pós-graduação em Informática

Brasília, 22 de Junho de 2021

Dedication

I dedicate this work to my parents Teresa Cristina and Aluizio, who passed away during the master's course. Thank you for being a source of strength and inspiration throughout my life.

Acknowledgments

Thanks for my advisor, Dr Edna Dias Canedo, to whom I had the privilege of being advised. Thanks to Anayran Pinheiro de Azevedo and Heloise Acco Tives, for their significant contributions to the development of this work. In addition, I would like to thank Fernando de Albuquerque Guimarães for having made possible the evaluation through survey. I would also like to thank the professors of UniCeub who voluntarily allowed my presentation in their classes, and the survey participants for their time. I would like to express my gratitude to my dear colleagues who participated in the focus group, to Flávio Altoé for the logo, to Caio and Lucas Althoff for all the insights and motivation. To all my colleagues, friends and family, who supported me until the end of this stage, despite the immense difficulties and moments of mourning. Thanks to professors Carla Silva and Vander Alves for accepting to participate in the evaluation commission and contributing with their experiences and knowledge.

Resumo

Contexto: O uso crescente de sistemas baseados em Inteligência Artificial (IA) em diversas áreas do nosso cotidiano provoca o aumento da consciência sobre as implicações éticas do seu uso. A maioria dos estudos encontrados na literatura focam em diretrizes e princípios éticos para IA, e estes não atendem às demandas dos projetos de desenvolvimento de software do mundo real, uma vez que estes princípios são de alto nível e abstratos, implicando em pouco esforço em como operacionalizar a ética na IA. Assim, existe uma lacuna na operacionalização destes princípios na prática. **Objetivo**: Desta forma, o objetivo deste trabalho é propor um guia para apoiar os *Product Owners* e desenvolvedores de sistemas baseados em IA na elicitação de requisitos éticos. **Metodologia**: Utilizamos a metodologia Design Science Research e na etapa de compreensão do problema realizamos uma revisão sistemática de literatura. Além disso, desenvolvemos um guia online e realizamos a sua validação através de um *survey* e um grupo focal com profissionais da área. **Resultados**: Identificamos 33 estudos primários relacionados às diretrizes e aspectos práticos de ética em IA, dentre eles, destacamos 11 princípios éticos e o método ECCOLA. Desenvolvemos o Guia para Elicitação de Requisitos Éticos em IA (RE4AI Ethical Guide) como um sistema baseado na web, funcionando como um Planning Poker, composto de 26 cartas novas, abarcando os 11 princípios encontrados na primeira etapa. A validação do Guia foi realizada através de um *survey* com 40 estudantes de graduação e pós-graduação, e um grupo focal com 5 profissionais na área. Nossos achados sugerem que o RE4AI Ethical Guide, de acordo com a percepção dos profissionais da área, é interessante e tem o potencial de facilitar a elicitação dos requisitos éticos. **Conclusão**: O Guia proposto possui utilidade e praticidade e pode ajudar na elicitação de requisitos éticos no contexto de desenvolvimento ágil. Assim, nossos resultados preliminares revelam que o Guia contribui para preencher a lacuna entre princípios de alto nível e abstratos e a prática, auxiliando os desenvolvedores e *Product Owners*, principalmente em projetos de desenvolvimento ágil, a elicitar requisitos éticos e operacionalizar a ética em IA.

Palavras-chave: Ética, Ética em Inteligência Artificial, Engenharia de Requisitos, Requisitos Éticos, Aprendizagem de Máquina, Desenvolvimento de Software.

Resumo Expandido

Título: Explorando a Elicitação de Requisitos Éticos para Aplicações no Contexto da IA

O interesse em sistemas baseados em Inteligência Artificial (IA) vem ganhando força em um ritmo acelerado, tanto para times de desenvolvimento de software quanto para a sociedade como um todo. Este interesse crescente levou ao emprego de técnicas de IA, tais como Aprendizagem de Máquina e *Deep learning* para diversos fins, como medicina e sistemas de vigilância, e tais usos aumentaram a consciência sobre as implicações éticas do uso de sistemas de IA. Entretanto, as várias diretrizes e princípios éticos para Inteligência Artificial expostos na literatura não atendem às demandas do desenvolvimento do mundo real de sistemas baseados em IA, uma vez que estes princípios são de nível muito alto e abstratos. Além disso, a maioria dos estudos encontrados na literatura se concentrava exclusivamente em princípios e diretrizes, implicando em um pequeno esforço em como implementar a ética na IA.

A necessidade de ética em IA surge do fato que sistemas baseados em IA com um projeto mal-apresentado pode ter um impacto muito negativo na sociedade, podendo haver um mau uso ou comportar-se de forma imprevisível e potencialmente prejudicial, agravados pela crescente presença desses sistemas no nosso dia a dia. Atualmente, a atividade de pesquisa no campo de ética em IA vem sendo focada na criação e elaboração de princípios e diretrizes, contendo guias teóricos do quê é a ética em IA. Já foi indicado que, em parte, é devido à falta de métodos e ferramentas formais que a ética em IA não está sendo amplamente implementada. Assim, é um desafio em progresso diminuir a lacuna entre a pesquisa e prática nesta área. Dessa forma, dado este cenário, há uma demanda de métodos e ferramentas práticas para a implementação de ética em IA. Portanto, existe uma necessidade de se incluir práticas tradicionais da Engenharia de Software, voltadas ao contexto de IA, que tenham como foco o processo de Engenharia de Requisitos, mais especificamente na fase elicitação de requisitos, pois, é nesse processo que ocorre uma maior interação entre diferentes partes interessadas, sendo a fase ideal para abordar as diversas questões éticas provenientes das diretrizes de ética em IA.

O objetivo geral deste trabalho é investigar como pode ser realizada a implementação dos princípios éticos em sistemas baseados em IA durante o processo de desenvolvimento de software, propor um guia para apoiar a implementação desses princípios, além de avaliar e propor melhorias para o guia. Desta forma, este trabalho tem como objetivo a elaboração de um guia prático para que desenvolvedores de sistemas baseados em IA e *Product Owners* possam abordar requisitos éticos nas primeiras fases do Ciclo de desenvolvimento de software, de forma iterativa em contextos ágeis, com auxílio visual, apresentando também documentação suficiente e exemplo de uso. Além disso, este trabalho visa realizar avaliações do guia proposto através de *survey* com estudantes de graduação e pós-graduação, e grupo focal com especialistas na área.

Neste trabalho foi desenvolvido o Guia para Elicitação de Requisitos Éticos para Inteligência Artificial (RE4AI Ethical Guide), para auxiliar a implementação de ética em IA por times de desenvolvimento. A metodologia adotada para este fim foi o Design Science Research em cinco passos: 1) Compreensão do problema, 2) sugestão de um projeto piloto, 3) desenvolvimento de um protótipo, 4) sua avaliação e 5) conclusão. A adoção deste método de pesquisa justifica-se pela sua capacidade de auxiliar os pesquisadores a construir um artefato e melhorá-lo através de um processo contínuo de refinamento e avaliação.

1. **Compreensão do problema:** Esta fase envolve a busca de informações sobre o problema a ser investigado, porém, sem tentar solucioná-lo ainda. Esta fase tem como objetivo o entendimento e a descrição do problema. Nesta fase são identificados os principais atores envolvidos e afetados pelo problema, além dos objetivos a serem atingidos, as causas/razões que originam o problema, os efeitos e as contribuições quando os objetivos forem alcançados. Assim, nesta etapa foi conduzida uma revisão sistemática de literatura, em um processo empírico e exploratório de busca, análise e descrição do conhecimento, envolvendo as abordagens para implementação de ética em IA. Esta fase tem como saída uma proposta para um novo esforço de pesquisa.

2. **Sugestão:** Estreitamente relacionando com a fase anterior, a sugestão é primordialmente uma etapa criativa na qual novas configurações são concebidas e assentadas em uma nova configuração de novos elementos ou de elementos previamente existentes. Para esta fase, foi concebido um artefato, considerando os dados obtidos na Compreensão do problema. O artefato proposto foi a implementação de um método para apoiar a operacionalização de ética em IA em equipes de desenvolvimento. Esta fase tem como saída uma proposta conceitual para o guia proposto (projeto piloto).

3. **Desenvolvimento:** Nesta fase, o projeto piloto é aprimorado e definido em um protótipo. O guia criado estabeleceu uma coleção de princípios, métodos e ferra-

mentas relevantes para se realizar a operacionalização de ética em IA no processo de desenvolvimento de software.

4. **Avaliação:** Nesta etapa, foi verificado o artefato proposto por meio de um *survey* com 40 estudantes de graduação e pós-graduação, além de um grupo focal com 5 especialistas da área de IA e elicitação de requisitos, que receberam uma breve descrição de como utilizar o guia.

5. **Conclusão:** Na última etapa, foram realizadas as análises do trabalho, resultados da fase anterior, apontando as contribuições encontradas e possíveis trabalhos futuros.

Para a primeira etapa foi realizada uma revisão sistemática da literatura a fim de: 1) identificar as técnicas, metodologias, métodos, frameworks, processos e ferramentas existentes na literatura para apoiar a operacionalização de requisitos éticos em IA; 2) identificar os trabalhos que investiguem a ética na elicitação de requisitos para aplicações no contexto de Inteligência Artificial e Aprendizagem de Máquina; 3) identificar os princípios e diretrizes éticos existentes na literatura e na indústria no contexto de Inteligência Artificial.

Dos 33 estudos primários selecionados, poucos abordam explicitamente o uso ou apresenta novas propostas de meios práticos para se implementar ética em IA. As propostas encontradas apresentam um baixo nível de maturidade, poucos exemplos práticos e falta de documentação. Isso demonstra como a ética em IA prática ainda está em seus primeiros estágios, sobre tudo em relação às orientações práticas, requisitos éticos, e ferramentas.

Após a análise das técnicas, ferramentas, métodos, frameworks e processos encontrados na literatura, identificamos o método ECCOLA como o mais adequado ao nosso contexto, consistindo em um baralho de cartas, baseado no Planning Poker, para a elicitação de requisitos éticos em IA, disponibilizado de forma estática. Também encontramos a necessidade da inclusão de práticas tradicionais da Engenharia de Software, como a elicitação de requisitos, para a o contexto de Inteligência Artificial, além das características de um Guia para implementar ética em IA: 1) Amplo; 2) Operacionalizável; 3) Flexível; 4) Iterativo; 5) Guiado; 6) Participativo.

Existe uma sobreposição entre os princípios que são elencados pelas diversas diretrizes publicadas. Várias são as críticas em relação às diretrizes de ética em IA disponíveis: as diversas diretrizes de ética em IA publicadas pelo setor privado servem principalmente como estratégia de marketing, visto que não há consequências no não cumprimento destas diretrizes. Além disso, a falta de sentimento de responsabilização e a distribuição de responsabilidade, a falta de conhecimento prévio do impacto ético, e acima de tudo, os incentivos financeiros (e.g., empresas esperam produzir mais em menos tempo), prejudicam o comprometimento aos princípios éticos presentes nas diretrizes, durante o desenvolvimento e aplicação dos sistemas baseados em IA.

Dentre os princípios éticos encontrados na revisão sistemática da literatura, nós selecionamos os que foram utilizados na proposta do Guia: 1) Transparência; 2) Justiça e equidade (*fairness*); 3) Não-maleficência; 4) Responsabilidade; 5) Privacidade; 6) Beneficência; 7) Liberdade e autonomia; 8) Confiança; 9) Sustentabilidade; 10) Dignidade; 11) Solidariedade.

Para a segunda etapa, foi criado um projeto piloto, funcionando como uma implementação do método ECCOLA e uma prova de conceito para a próxima etapa, onde são apresentadas as 21 cartas, abarcando 8 princípios, originais dos autores. O projeto piloto está disponível em https://josesiqueira.github.io/eccola/index.html, e seu código fonte em https://github.com/josesiqueira/eccola.

Para a terceira etapa, foi desenvolvido um protótipo – RE4AI Ethical Guide – composto de 26 cartas novas, abarcando os 11 princípios encontrados na primeira etapa, elaborado em cima das funcionalidades do projeto piloto da etapa anterior. Para utilizar o Guia em reuniões de sprint backlog, os atores devem escolher as cartas que irão ser usadas naquele sprint, realizar a leitura em voz alta do conteúdo da carta, em seguida o time de desenvolvimento irá elicitar os requisitos éticos em forma de história de usuário, anotando também o raciocínio que os levaram àquelas histórias de usuário. A validação deve ser feita pelos times de desenvolvimento conjuntamente com os clientes e as múltiplas partes interessadas, que podem solicitar alterações, por sprint. O Guia foi desenvolvido como um sistema baseado na web (utilizando HTML, CSS e JS), permitindo interatividade na seleção das cartas através de filtros e comparações entre múltiplas cartas, além de amplo material de apoio (como utilizar, princípios, ferramentas, trade-offs) e a adição de sugestão de ferramentas no conteúdo das cartas. O código fonte do guia está disponível em https://github.com/josesiqueira/RE4AIEthicalGuide e o sistema em https://josesiqueira.github.io/RE4AIEthicalGuide/.

Finalmente, para a quarta etapa, avaliação, foi realizado um estudo misto através de um *survey* com 40 estudantes de graduação e pós-graduação que avaliaram o Guia através de um questionário on-line, além de um grupo focal com 5 especialistas em IA em distintas áreas de atuação. Apontamos que o RE4AI Ethical Guide é percebido como de grande interesse pelos participantes, recebendo uma avaliação geral positiva em ambos os estudos, e pode ajudar a mitigar o desafio de operacionalizar os princípios éticos na prática. Durante o sessão do grupo focal, os participantes foram capazes de elicitar 18 requisitos para um cenário hipotético, demonstrando a usabilidade e praticidade do Guia. Alguns participantes acharam o Guia extenso e amplo, sugerindo um estreitamento do escopo, e direcionamento para contextos específicos. Também notaram a necessidade da

disponibilização do Guia em outras línguas além do inglês.

Alguns possíveis trabalhos futuros sugeridos envolvem: a apresentação de exemplos de uso do RE4AI Ethical Guide, bem como das ferramentas, processos, frameworks e métodos disponíveis, em diferentes contextos; realização de outras avaliações do guia para identificar as percepções de um conjunto diversificado de profissionais em IA no uso das ferramentas, processos, frameworks e métodos do guia e propor melhorias; a emissão de um certificado de AI ético, através de auditorias oficiais feitas por órgãos públicos e/ou autorizados; apresentação de um catálogo ou uma base de dados de requisitos éticos em IA; avaliação da aplicação do Guia proposto quando deseja-se avaliar sistemas baseados em IA já implantados; a necessidade de mais trabalhos que foquem no ensino de ética em AI na formação dos futuros profissionais como parte do currículo adotado nos cursos relacionados ao desenvolvimento de sistemas baseados em IA, a fim de aumentar a consciência ética entre os alunos dos cursos de computação, bem como o treinamento dos profissionais de TI pelas organizações.

Esperamos contribuir no auxílio, tanto no desenvolvimento de pesquisas futuras no contexto de ética em IA ambos na academia e na indústria, quanto na escolha de ferramentas e processos que apoiam a implementação de ética em sistemas baseados em IA, além de conscientizar sobre as várias questões éticas envolvidas no uso de sistemas baseados em IA e seus desafios no processo de desenvolvimento. A contribuição principal deste trabalho foi a apresentação do Guia para Elicitação de Requisitos Éticos para IA.

Palavras-chave: Ética, Ética em Inteligência Artificial, Engenharia de Requisitos, Requisitos Éticos, Aprendizagem de Máquina, Desenvolvimento de Software.

Abstract

Context: The increasing use of Artificial Intelligence (AI) based systems in various areas of our daily lives provokes increased awareness about the ethical implications of their use. Most of the studies found focus on ethical guidelines and principles for AI, and these do not meet the demands of real-world software development projects, since these principles are high-level and abstract, implying little effort on how to operationalize ethics in AI. Thus, there is a gap in operationalizing these principles in practice. **Objective**: Therefore, the aim of this work is to propose a guide to support Product Owners and developers of AI-based systems in the elicitation of ethical requirements. **Methodology**: We used the Design Science Research methodology and in the awareness of the problem phase we conducted a systematic literature review. In addition, we developed an online guide and performed its validation through a survey and a focus group with professionals in the area. **Results**: We identified 33 primary studies related to guidelines and practical aspects of ethics in AI, among them, we highlighted 11 ethical principles and the ECCOLA method. We developed the Guide for Artificial Intelligence Ethical Requirements Elicitation (RE4AI Ethical Guide) as a web-based system, working as a Planning Poker, composed of 26 new cards, covering the 11 principles found in the first step. The evaluation of the Guide was performed through a survey with 40 undergraduate and graduate students, and a focus group with 5 AI professionals. Our findings suggest that the RE4AI Ethical Guide, according to the perception of experts in the area, is interesting and has the potential to facilitate the elicitation of ethical requirements. **Conclusion**: The proposed Guide is both useful and practical and can help in the elicitation of ethical requirements in the context of agile development. Thus, our preliminary results reveal that the Guide contributes to bridging the gap between high-level and abstract principles and practice by assisting developers and Product Owners, especially in agile development projects, to elicit ethical requirements and operationalise ethics in AI.

Keywords: Ethics, Ethics in Artificial Intelligence, Requirements Engineering, Ethical Requirements, Machine Learning, Software Development.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

There is an increasing amount of software development teams focusing on building Artificial Intelligence (AI) based systems, and they are gaining popularity in our society at a rapid pace [12] [13]. The use of AI techniques such as Machine Learning (ML) and Deep Learning (DL) in various fields such as medicine, surveillance systems, business, transportation and many other domains, has raised awareness of the ethical implications of using such systems [14] [13], becoming a subject of great interest to researchers, industry professionals and the general public [15], aggravated in the context of data privacy and model training in the pandemic situation of COVID-19 [16].

Although the popularization of AI is expanding, incidents related to AI-based systems are also becoming more common [14]. Some incidents have led to public discussions about the ethics of building applications in the context of AI. One such case is the Cambridge Analytic scandal, where data from users of Facebook was improperly obtained and used to influence the outcome of an election [17]. Another example is an ML algorithm biased against women in hiring new professionals at Amazon. This biased algorithm favoured hiring more male candidates [18]. There is evidence to suggest that crime prediction tools used for decision-making in the criminal justice system contain racial bias against blacks and minorities, for example, the COMPAS tool [19]. In addition, new threats are emerging regarding the ethical misuse of AI-based systems, such as fake news, with the use of deep-fake and AI-based voice technologies, where a person's face can be superimposed on videos and political leaders can be portrayed inciting violence and panic, for example, can be used to manipulate elections, change political opinions and spread disinformation in general [20] [21] [22].

Several ethical guidelines and principles for Artificial Intelligence have been proposed by organisations, commissions, institutes and industry. These proposals, however, do not meet the requirements of developing AI-based ethical systems in the real world, as these ethical principles are often too abstract and general [11] [23] [24] and do not constitute real evidence that they can influence ethical decision-making [25]. Consequently, those responsible for developing these AI-based systems, who are also concerned with the ethical questions that arise, are frustrated by the limited support that the abstract texts provided by the available guidelines and principles offer [13].

Most studies found in the literature focus on theoretical principles and conceptual guidelines, not providing an effective and realistic framework on how to implement ethics in AI [14] [13]. Therefore, a deeper focus on the technological details of various methods and technologies in the field of AI and ML is needed; in other words, there is currently a need to bridge the gap between abstract values (principles, guidelines and codes) and technical implementations [11] [13].

Morley et al. [13] found tools and methods on how to implement ethics in ML, however many of them focus on specific parts of the software development process, have little documentation and present lack of usability, besides extra work in the implementation stage – because they are not off-the-shelf software (ready to use) –, need adaptation to the scenario/context in which they will be applied. It is also worth noting that researchers have used as a central axis in their research the development of tools to explain or interpret ML-based software, that is, after the software is ready [13]. However, addressing the issue of ethics in AI from its design or development phase is cheaper and simpler than at the implementation phase [14]. The vast majority of these tools do not assist software development teams in the design and development processes in their entirety [6]. Thus, developing AI ethics is an extremely challenging and complicated task [20].

Some authors present work on ethical issue in Requirements Engineering (RE) [24] [26], but not in the specific context of AI. It is observed that while there are approaches on how to use AI and consequently ML to improve Requirements Engineering tasks, there are not many works on Requirements Engineering for ML or AI-based systems [12]. Vogelsang and Borg [12] defined characteristics and challenges of RE to meet the needs of the ML context, indicating possible modifications in the RE process arising from the ML development paradigm. The authors state that requirements engineers need, among other aspects, to include ethical principles in the requirements elicitation phase during the software development process in the AI context. Vakkuri et al. [6] presented ECCOLA, a card-based method to implement AI ethics at project level. Through questions investigated while performing requirements elicitation, the authors argue that it is possible to increase ethical awareness in agile development teams. It was not found in the literature, the application of the method presented by the authors in a real context.

Some research opportunities were pointed out by Morley et al. [13], such as: testing the tools, methods or frameworks and providing examples of use in a real context. Thus, there is a need to identify, propose or improve methods, frameworks and tools to support the implementation of ethics principles in the context of AI during the software development process. In this context, the focus of this work is to investigate how the implementation of ethical principles in AI-based systems can be performed during the software development process in requirements activities. Therefore, based on ethical principles and guidelines, as well as tools, methods and frameworks identified in the literature, we will propose a guide to support the implementation of these principles along with its evaluation.

1.1 Research Problem

Several studies available in literature address and investigate AI Ethics theoretical aspects, such as principles and guidelines [27] [11] [28] [29]. At least 84 public-private initiatives have emerged globally to define guidelines, values, principles and concepts for the development and implementation of ethics in AI-based systems [30]. AI ethics principles – despite promising to be easy to implement, in practice they contain mostly vague and abstract principles [30]. Given this scenario, the problem is that there is, so far, no standardization of how these principles should be applied in real scenarios, nor the application of a tool, framework or guide that can help software development teams in implementing these principles at project level [13].

In order to minimize this problem, this work focuses on investigating in the literature principles of ethics in AI and the existing tools to support the implementation of these principles by professionals during the software development process. From identified guidelines, principles and tools, a guide will be proposed containing the main ethical principles in AI, to support the implementation of these principles during the software development process, assisting AI practitioners in operationalising AI ethics in practice in the requirements elicitation phase. There, to steer our goal, research questions were defined, as presented in Subsection 3.1.1.

1.2 Justification

The need for AI ethics arises from the fact that AI-based systems with a poorly presented design can have a very negative impact on society [13], there can be misuse or behave in unpredictable and potentially harmful ways [31], aggravated by the increasing presence of such software in our daily lives. Currently, research activity in the field of AI ethics has been focused on the creation and elaboration of principles and guidelines, containing theoretical guides of what AI ethics is. It has already been indicated that it is partly due to the lack of formal methods and tools that AI ethics is not being widely implemented [32]. Thus, it is a challenge in progress to close the gap between research and practice in this area. Therefore, given this scenario, there is a demand for practical methods and tools for implementing AI ethics [32]. Some issues present in the literature to be avoided in the task of implementing ethics in AI-based systems are [14]:

- Appoint a single individual to operationalise ethics in AI. The entire development team should be involved;
- Outsource ethics to an ethics committee. Ethics, as well as quality requirements, cannot be outsourced;
- Ethics cannot be implemented without being carried out in a systematic way. As such, delegating ethical issues solely to the developers to address is unlikely to work. Without a clear method for addressing AI ethics to assist developers, they will rely only on their own capabilities.

Improving investors trust and research support funds, as well as promoting interest in AI research and broadening the acceptance of AI-based systems are some of the benefits highlighted by recent research when performing this task [13]. Consequently, it is important, to find the tools and methods for performing – in a systematized way – tasks to operationalise AI ethics during the software development process, as well as providing the evaluation of these tools, and to propose improvements.

1.3 Objectives

1.3.1 General objective

The overall objective of this study is to investigate how the translation of ethical principles in AI-based systems can be carried out during the software development process, to propose a guide to support how to operationalise these principles by software development teams, its evaluation and propose refinements of the guide.

1.3.2 Specific Objectives

To achieve the general objective of this work, the following specific objectives were defined:

- To conduct a systematic literature review to identify studies that investigate ethics in AI-based systems, both in a theoretical approach: through its principles, challenges, approaches, contexts, cases and applications; and practical: that propose or define techniques, tools, methods, frameworks and processes;

- Select the ethical principles in the context of AI found in the literature that will be used in the proposal of a guide;
- To analyse the techniques, tools, methods, frameworks and processes found in the literature and select those to be used in the proposed ethical guide;
- To evaluate the proposed guide and analyse the perception of software development teams regarding the impact of the use of the guide on the tasks performed by the team;
- Propose improvements to the guide, if necessary, based on software development teams perceptions.

1.4 Expected Results

- Identifying the techniques, tools, methods, frameworks and processes for eliciting ethical requirements in AI;
- Development of a guide with ethical principles and guidelines for applications in the context of AI;
- Implementation of a guide to support the operationalisation of ethical principles during AI-based systems development.
- Evaluation of the proposed Guide through a survey and a focus group.

1.5 Research Methodology

The conduct of this research was guided by Design Science Research (DSR) [1]. This method guides the cycle of design, evaluation and refinement of an artifact allowing the proposed goal to be achieved. Design Science Research means inventing and bringing into existence, i.e. design creates artifacts that do not yet exist. If the knowledge required to create such an artifact already exists, then design is routine, if not, it is innovative; and innovative design calls for conducting research to fill knowledge gaps. DSR is a rapidly evolving field, even in the last decade the most accepted name for the field has been changed from Design Research (DR) to Design Science Research. DSR is research using design as a research method or technique [1]. The defining characteristic of DSR is learning/knowledge creation through the construction of artifacts [1]. Although with a degree of similarity and compatibility with Action Research (AR) – contributing to practice and research at the same time – DSR and AR are distinct in levels of research interest and activities [33], with AR being more context-specific dependent.

The adoption of this research method is justified by its ability to help researchers build an artifact and improve it through a continuous process of refinement and evaluation. The goal of DSR is to contribute in an area of interest by creating as final output a new and interesting Design Science Knowledge (DSK) [1]. The DSK can be in the form of artifacts, constructs, models, frameworks, architectures, methods and/or instantiations. Constructs are the conceptual vocabulary of a problem/solution domain, which arise during the conceptualisation of a problem and are refined throughout the DSR cycle. A model is a set of propositions or statements that express the relationships between constructs. A method is a set of steps to accomplish a task. Frameworks are conceptual or actual guides that serve as a support or guide. The artifact generated in this work will be a guide to support the implementation of ethics in AI. Vernadat [34] defines framework as "a collection of principles, methods or tools relevant to a particular application domain". The research questions investigated in this work guided the process of building, refining and evaluating the guide for implementing AI ethics by development teams.

What distinguishes the DSR cycle to other corresponding design process models is that the main focus of DSR should be the contribution of new knowledge [1]. In this work the Design Science Research cycle based on Vaishnavi et al. [1] was used, consisting of the following phases: Awareness of the problem, suggestion, development, evaluation and conclusion. Nonetheless, in this study only one cycle of the DSR method was performed.

Figure 1.1 presents the phases of Design Science Research used in this study. Moreover, it shows an overview of the organization of the methodology used in this work, with the respective outputs developed in each phase.

1. **Awareness of the Problem:** This phase involves the search for information about the problem to be investigated, however, without trying to solve it yet. This phase aims at understanding and describing the problem. In this phase, the main actors involved and affected by the problem are identified, in addition to the objectives to be achieved, the causes/reasons that originate the problem, the effects and the contributions when the objectives are achieved. Thus, in this stage a systematic literature review was conducted, in an empirical and exploratory process of search, analysis and description of knowledge, involving the approaches to operationalise AI Ethics. This stage has as output a proposal for a new research effort.

2. **Suggestion:** Closely related to the previous phase, suggestion is primarily a creative stage in which new configurations are conceived and settled on a novel configuration of new or previously existing elements. For this phase, an artifact was designed, considering the data obtained in the Awareness of the Problem. The proposed artifact was an implementation of a method to support the implementation of AI

ethics by development teams. This phase has as output a conceptual proposal for the proposed guide (pilot project).

3. **Development:** In this phase, the pilot project is refined and defined. The proposed guide will be important to define a collection of relevant principles, methods and tools to perform the implementation of AI ethics in the software development process. This phase has as output our proposed Guide for Artificial Intelligence Ethical Requirements Elicitation (prototype).

4. **Evaluation:** Evaluation of DSR artifacts may include appropriate methods that result in empirical evidence. Qualitative data collection techniques will be used for evaluation purposes. In this step, the proposed artifact will be evaluated through a survey with undergraduate and graduate students and a focus group with experts in the field of AI and requirements elicitation, who will receive a brief description of how to use the guide, place their perceptions and propose refinements. The proposed guide will be evaluated through the analysis of the survey and focus group results.

5. **Conclusion:** In the last stage, the analysis of the results of the whole process will be carried out, pointing out the contributions found, possible adjustments, and future works.

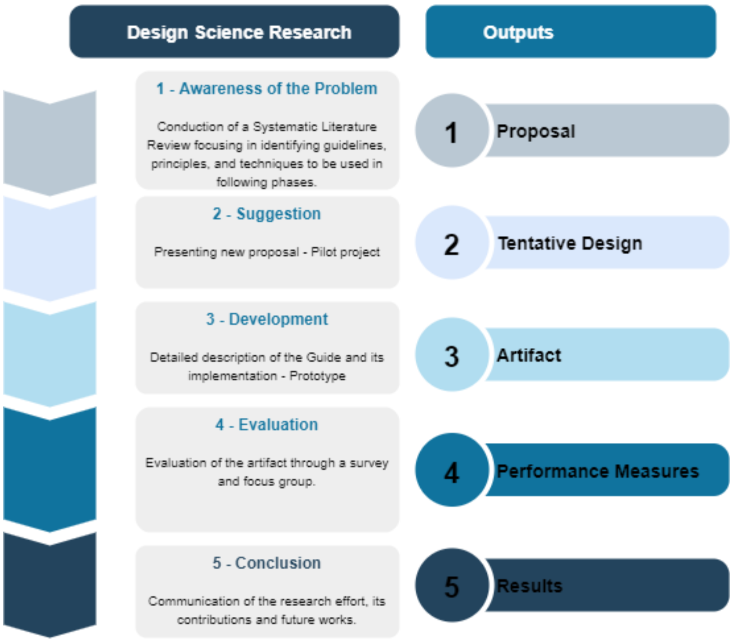

Design Science Research
Outputs
1 - Awareness of the Problem
Conduction of a Systematic Literature Review focusing in identifying guidelines, principles, and techniques to be used in following phases.
2 - Suggestion
Presenting new proposal - Pilot project
3 - Development
Detailed description of the Guide and its implementation - Prototype
4 - Evaluation
Evaluation of the artifact through a survey and focus group.
5 - Conclusion
Communication of the research effort, its contributions and future works.
1 Proposal
2 Tentative Design
3 Artifact
4 Performance Measures
5 Results

Figure 1.1: Design Science Research cycle with the addition of the outputs of each step. Adapted from [1].

1.6 Structure of the Study

This work is organized into five Chapters, in addition to this one, consisting of:

- **Chapter 2:** presents the theoretical background in relation to AI, AI Ethics in theory and practice, as well as requirements elicitation concepts. In addition, the related works identified in the literature review are presented.
- **Chapter 3:** presents the results of the Systematic Literature Review conducted to this study, answering the research questions that were defined.
- **Chapter 4:** presents a pilot project and a prototype as a Guide to support the implementation of ethics in AI-based systems development process.
- **Chapter 5:** presents the conduction and results of the evaluation of the proposed guide and suggests adjustments/improvements to be incorporated in the guide after its assessment by practitioners.
- **Chapter 6:** presents the main conclusions of this work and discusses future works.

Chapter 2

Theoretical Background

In this Chapter will be presented the main conceptual aspects necessary for a better understanding of this work. In Section 2.1 it will be presented a definition and contextualization of Artificial Intelligence and the essential features and characteristics of Machine Learning and Deep learning. Section 2.2 discusses the context of AI ethics, its research need, main features and challenges, then we unfold this Section in three adjacent ones: Guidelines, Principles and Practice. The first two present the theoretical aspects of AI ethics, identifying in Section 2.2.1 the multiple guidelines produced in the context of AI ethics by different authors, and in Section 2.2.2 the several principles that compose these guidelines; the third Section 2.2.3 will discuss the practical aspects of AI ethics, describing the practical approaches involved in AI ethics, dealing with the relations between the solutions that help the implementation of AI ethics and its principles and the phases of the Software Development Cycle. Then, Section 2.3 presents an overview about Requirements Elicitation within the field of Software Engineering. Next, Section 2.4 provides the related works of interest to our study. Finally, Section 2.5 presents a summary of what has been covered throughout this chapter.

2.1 Artificial Intelligence

Artificial Intelligence (AI) is a rapidly expanding area of development and application. After the Second World War, AI emerged as one of the newest fields of science and engineering, and in 1956 its name was coined [35]. Since then, it has been growing the interest of professionals and researchers from other fields of knowledge, such as medicine, microeconomics and biotechnology. There are several domains in which it can be applied, e.g., autonomous vehicles, legal area, education, transportation, chatbots, image recognition, content recommendation, fraud detection, machine translation [36]. AI is traditionally referred to as a branch of Computer Science dealing with the automation of intelligent behavior [3].

Artificial intelligence systems are software systems designed by humans that can perceive the environment through sensors or data acquisition, interpret the data or process the information, and from there decide on the best action to take to achieve a complex goal [37]. Some of the AI technologies developed and most employed today are Machine Learning (ML) – and Deep Learning (DL) [10]. It is worth pointing out that the field of AI – which has been a field in computer science for decades –, goes beyond Machine Learning and Deep learning, these being only the center of the current debate [10].

2.1.1 Machine Learning

An agent is learning if, after making observations about the world, it increases its performance on future tasks [35]. A learning algorithm is a method for learning from data, – usually from lots of data – which is used to train the algorithms until they are able to identify patterns correctly, and apply the knowledge to analyze new, similar situations [2]. A user separates the initial data set between data for training and data for testing, then trains a huge amount of data, until it obtains a model that is able to predict correctly in new situations, in possession of new data [35].

Training this huge amount of data means adjusting parameters of a mathematical structure of decision-making rules, and the analogy is made in that the algorithm is a box that applies this rule while adjusting these parameters (which can be millions). The opacity of these algorithms is defined by the difficulty of understanding how they work, and the term black box is devised to refer to this opacity [2]. Once the system has been trained, a test dataset is used to evaluate the accuracy and effectiveness of the obtained model [35].

The goal of Machine Learning is to create a trained model that is able to make generalizations without human intervention, being accurate not only for the examples in the training dataset, but also for future cases that have not been seen before [2]. The steps required for Machine Learning are presented in Figure 2.1.

Machine Learning is presented in several techniques, and the most widespread are supervised, unsupervised, and reinforcement learning [37]. Although these techniques differ in terms of approach and objectives, all share the main goal of giving computers the ability to act without being explicitly programmed.

During Supervised Learning, the algorithm learns from a pre-defined and pre-labelled dataset, making the link between the input data and the expected outputs by describing their relationship through a function. In other words, the goal of supervised learning is

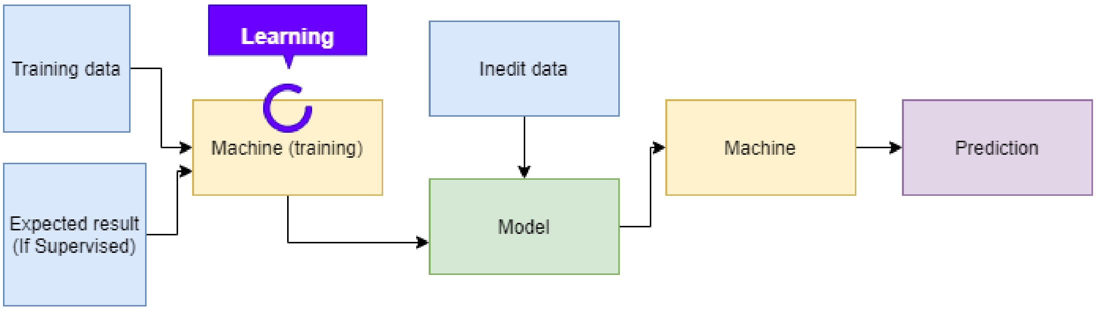


Figure 2.1: Machine Learning steps. Adapted from [2]

to learn a function that best describes the relationship between input values and their outputs. Thus, being able to predict outcomes from novel data [2].

The contrast between Supervised and Unsupervised Learning lies in the lack of explicit information about the output – the input data are not labelled –, and the aim is to identify structures or patterns present in the input data [35]. In Reinforcement Learning, the learning algorithm acts on the principle of reward maximization, that is, the algorithm learns from a series of reward or punishment reinforcement by observing states and taking specific actions [38].

2.1.2 Deep Learning

Deep Learning algorithms are one of the most evolved and successful of the types of learning approaches [39] [37]. Based on Neural Networks in multiple layers between the input and output, it allows the learning of the input-output relationship to occur in successive steps. Neural Networks are composed of a set of neurons, which, inspired by the biology of our brains, pass signal from one neuron to another – the nodes in Neural Networks [2]. By using Deep Learning it is possible to automatically learn, extract and translate features from datasets such as images, videos or text. Deep learning is a special form of Neural Networks, using thousands of layers and using a large amount of nodes in each layer, being able to recognize precise and extremely complex patterns in the data [2]. Figure 2.2 presents the topology of the Deep learning technique. Each neuron connects with that of the next layer, passing through multiple intermediate layers, reaching the output layer.

The opacity of learning algorithms, as defined in 2.1.1, is most evident in the case of Deep learning. This is because Deep learning algorithms are more complex, more precise, and more difficult to explain their inner workings [4]. In Figure 2.3 a black box is presented, representing this opacity, contrasting with the topology presented in Figure 2.2.

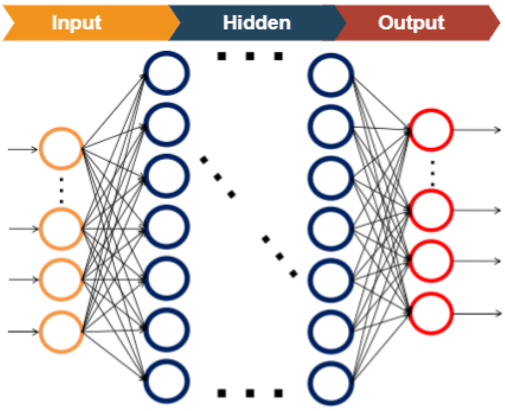


Figure 2.2: Topology of a multi-layer Neural Network. Adapted from [3]

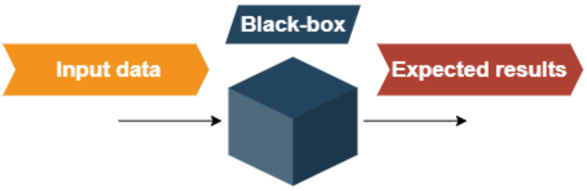


Figure 2.3: Representation of the black box of Learning algorithms. Own source

The areas in which this technique can be applied are in wide growth, some tasks that use it are: object identification in images and videos, face and identity recognition, natural language processing such as text and speech, as well as the creation of images and videos with interchanged faces (deepfake) [40], and of neural fake news, by generating natural language text, yielding large-scale realistic-looking videos [41]. Figure 2.4 illustrates the pertinence of each technique in its respective layer. Thus, it can be visualised that Deep learning is a subset of Machine Learning, being one of the ways to implement it.

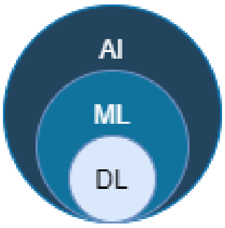


Figure 2.4: Diagram of the layers of AI techniques. Own source

Figure 2.5 presents the relationship between Learning techniques and explainability. Deep learning has a high degree of prediction effectiveness, however it is the technique that has the lowest explainability. Supervised learning is present in techniques such as Statistical models, Support Vector Machines (SVMs), and regression and classification techniques as Decision trees and Random forest, with a lower prediction efficiency but with an explainability slightly higher than the previous technique and; Learning by reinforcement in Markov Models and Markov Logic Networks (MLN) techniques [35], with an even lower prediction efficiency, however with an explainability relatively equal with the previous technique.

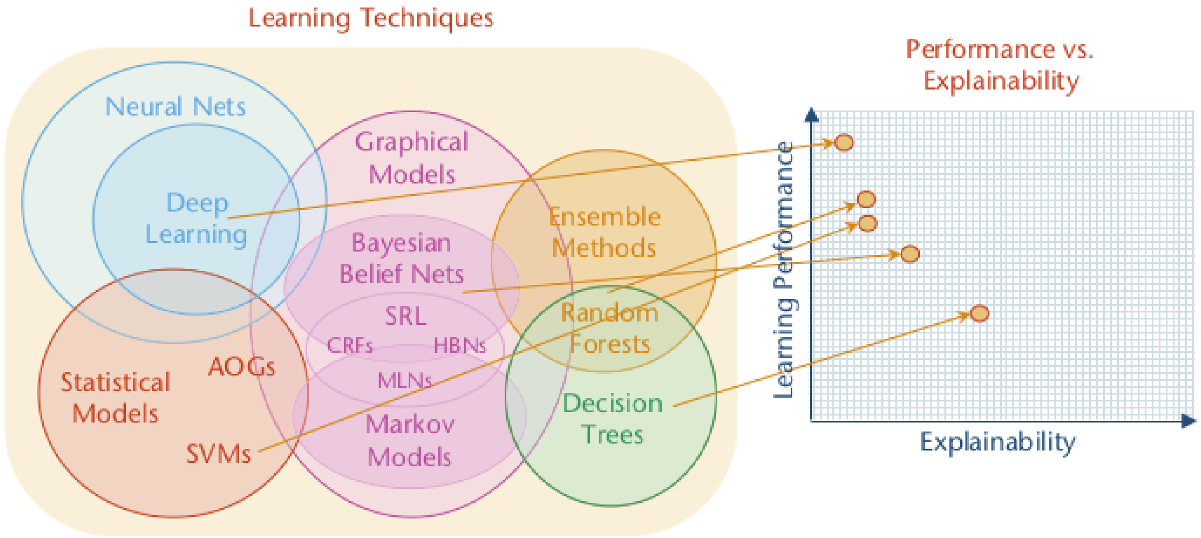


Figure 2.5: Prediction techniques versus explainability [4]

Artificial Intelligence offers several opportunities for individuals and society on a broad scale to dramatically improve and enhance their capabilities in performing complex tasks, being able to reinvent society by introducing the technologies, which, because they have such extraordinary and disruptive potential, also introduce proportional risks [29], beyond the business area, but also as global climate, legal system, medicine, global surveillance, transportation, wars, among others, from which arises the need to investigate AI ethics pointed out by several authors, including [36] [39] [14] [27] [23] and European institutions [42], and is currently a highly debated topic [10].

2.2 AI Ethics

Recently the spotlight on AI ethics is being increased and this has become apparent from the rising concern about the unintended negative impacts of the use of AI-based systems – which are present in our daily lives –, coupled with a lack of awareness, or even exclusion, of the ethical analysis of [43] engineering practice. In addition to scandals that have garnered global media attention, the ethical issues of using AI-based systems have an impact on society (privacy, human rights, dignity, bias, democracy), human psychology, the financial system, the legal system, the environment and the planet, among others [44]. Simultaneously, the professionals involved – engineers and developers – are not being trained, by the organizations where they work, to raise ethical issues during software development [11].

AI is pervading all areas of society's life, and to take advantage of the opportunities this technology delivers while limiting the associated risks of its use to individuals and society, a large number of guidelines for AI ethics have been emerging, developed by different stakeholders [39] [27]. There is an emerging convergence among the community regarding the values and principles present in these guidelines, such as transparency, fairness, non-maleficence, accountability and privacy [27]. However, many differ in the level of abstraction, and the higher this level the more interpretation is required to implement them [45], i.e. how ethical principles should be interpreted and implemented [27]. Furthermore, different countries are adopting strategies in different depths on the implementation and governance of AI ethics, differentiating themselves in terms of objectives and extent of investments [44].

That being said, it is not clear how to carry out the translation of often theoretical and abstract principles into practice [13] [14] [39]. AI-based systems are also software, that are developed by people – developers – therefore, they play an important role in AI ethics [14]. In the literature there are tools for implementing AI ethics, which are still in early stages of development, focus on small parts of the software development project and have a lack of usability, leading to frustration on the part of those responsible for implementing AI ethics [13]. In short, AI ethics has guidelines that govern how the practice should be carried out. Figure 2.6 presents the relationship of these concepts.

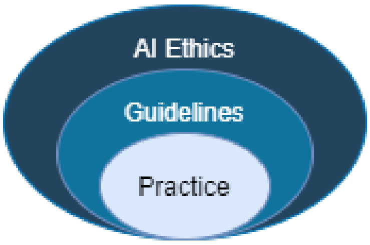


Figure 2.6: Diagram of AI ethics research layers. Own source

In addition, the literature presents the importance of the system being designed from its inception to satisfy ethical principles [13]. Moreover, ethical issues are cheaper to be addressed during the design phase [14]. Among the phases of the software development cycle, the tools listed in the literature focus mostly on explicability, i.e. in the implementation and maintenance phase; the existence of these tools are necessary but not sufficient [13]. No project-level tools were found in the literature [13]. Requirements engineering is an area within Software Engineering responsible for the elicitation and analysis of requirements for a system to be built – AI-based system, in our case – through the software development cycle [46].

Considering this motivation and our research objectives, Figure 2.7 presents an overview of the context of AI Ethics. From the figure we can identify that there are guidelines that address ethical issues in AI, which can be through the concepts of binding (e.g. GDPR and LGPD) and non-binding (e.g. AI HLEG and IEEE EADv1) laws and their respective ethical principles. Binding laws are "legally binding laws passed by legislators to define permitted or prohibited conduct" while ethical guidelines (non-binding) are not legally binding but of a persuasive nature [27]. Furthermore, it is possible to identify in which Software Development Life Cycle (SDLC) phases the AI ethics issues are implemented: 1. Requirements analysis; 2. Design; 3. Code; 4. Test; and 5. Implementing & maintenance. In this research, we will focus on its first phase – 1. Requirements analysis – in the context of AI ethics.

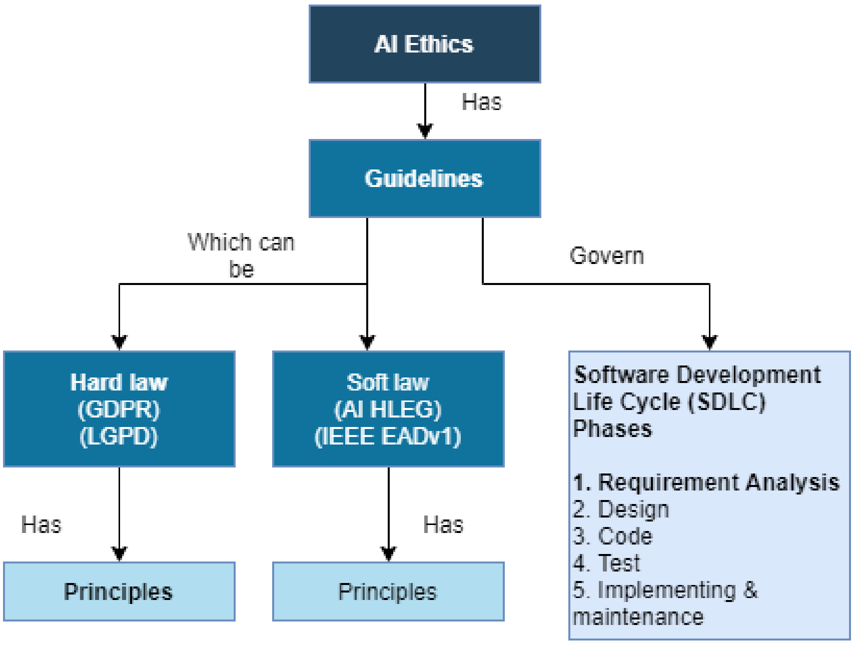


Figure 2.7: Diagram of exploration of ethics in AI, addressing theory and practice. Own source

2.2.1 Guidelines

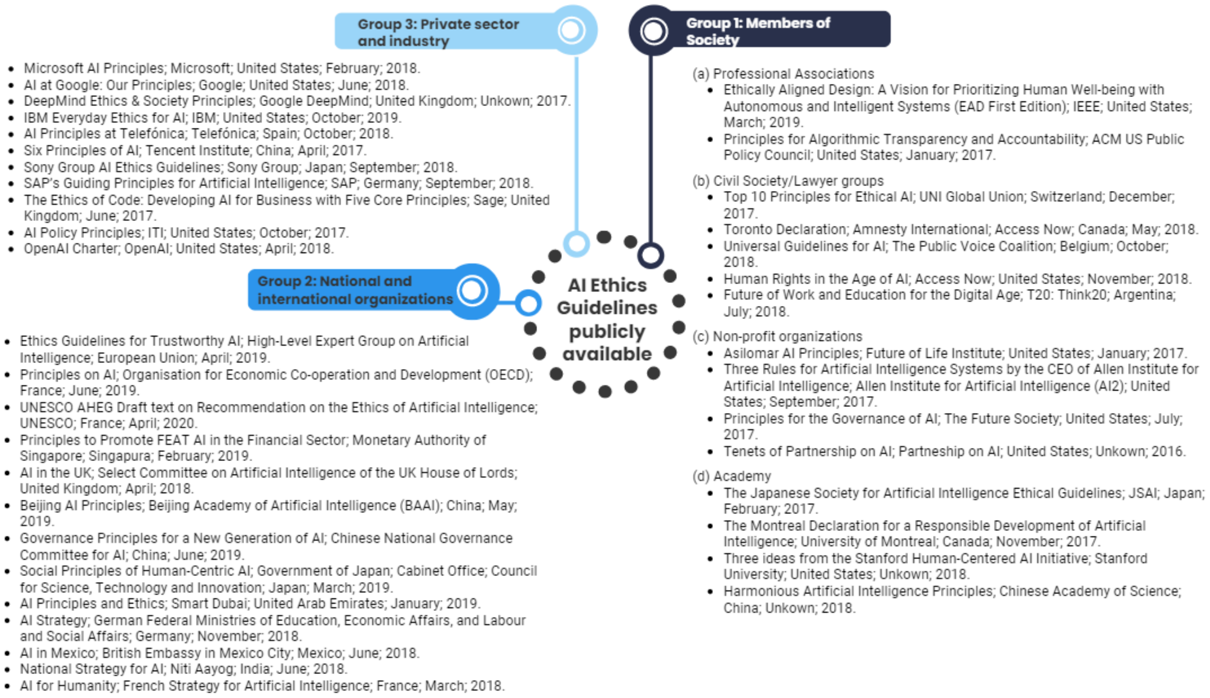
The past three years have seen a veritable proliferation of organisations publishing guidelines seeking to provide normative guides to AI ethics [47] [48]. As of November 2019, at least 84 of these initiatives have published reports describing ethical principles, values or other high-level abstract requirements for the development and deployment of AI [30]. Due to this high number of publications, sometimes the terms appear interchangeably in the papers, as in the Asilomar AI Principles, where principles are composed of guidelines. We assume throughout this paper that the guidelines – the guides – contain ethical principles in AI.

The issuers of AI ethics guidelines are conceived in 3 major groups, inspired by the classifications of [27], [2], [48], [49], [50]. Figure 2.8 presents each group and their respective descriptions:

1. **Group 1: Members of society** – consisting of actors drawn from groups within society.
    (a) **Professional associations** – their codes of ethics are intended to guide the work of professionals.
    (b) **Civil society/lawyers groups** – publishing documents that can serve to set an agenda for advocacy, activism or advocacy, as well as establishing a ground for ongoing debate.
    (c) **Nonprofit organisations** – often addressing sustainability issues.
    (d) **Academia** – exploring philosophical, legal, and technical aspects of how to use algorithms in an ethical way or bringing research and practice together [50].
2. **Group 2: National and international organisations** – presenting policies and regulations needed in different sectors.
3. **Group 3: Private sector and industry** – develop guidelines to guide the organization's internal development and use of AI technology, and communicate their objectives to other stakeholders, customers and regulators [48].

We found 39 AI ethics guidelines and performed an arrangement in order to make them belong to their respective group and enumerated them, as can be visualized in Tables A.1, A.2, A.3, A.4, A.5, A.6 of Appendix A and Figure 2.8. The Tables present information such as the Title of the guideline, the Source, the location, month and year of publication.

Most of the guidelines found are recent and were published between the years 2017 and 2020. The AI ethics guidelines presented are by no means final or exhaustive. Some collections of relevant documents are accessible to the general public. In particular, the Algorithm Watch's Global Inventory of AI Ethics Guidelines stands out, as it allows a search with the following criteria: sector/actor, document type, region and location. Founded in 2017 in Berlin, Germany, Algorithm Watch is a nonprofit organization that analyses the decision-making processes of Machine Learning class algorithms (these that make decisions automatically) and their possible consequences on human behavior. Charlotte Stix [51] conducted a compilation of the AI ecosystem in the European Union, pointing out which guidelines are being published by each European country.



Group 3: Private sector and industry

- Microsoft AI Principles; Microsoft; United States; February; 2018.
- AI at Google: Our Principles; Google; United States; June; 2018.
- DeepMind Ethics & Society Principles; Google DeepMind; United Kingdom; Unkown; 2017.
- IBM Everyday Ethics for AI; IBM; United States; October; 2019.
- AI Principles at Telefónica; Telefónica; Spain; October; 2018.
- Six Principles of AI; Tencent Institute; China; April; 2017.
- Sony Group AI Ethics Guidelines; Sony Group; Japan; September; 2018.
- SAP's Guiding Principles for Artificial Intelligence; SAP; Germany; September; 2018.
- The Ethics of Code: Developing AI for Business with Five Core Principles; Sage; United Kingdom; June; 2017.
- AI Policy Principles; ITI; United States; October; 2017.
- OpenAI Charter; OpenAI; United States; April; 2018.

Group 2: National and international organizations

- Ethics Guidelines for Trustworthy AI; High-Level Expert Group on Artificial Intelligence; European Union; April; 2019.
- Principles on AI; Organisation for Economic Co-operation and Development (OECD); France; June; 2019.
- UNESCO AHEG Draft text on Recommendation on the Ethics of Artificial Intelligence; UNESCO; France; April; 2020.
- Principles to Promote FEAT AI in the Financial Sector; Monetary Authority of Singapore; Singapura; February; 2019.
- AI in the UK; Select Committee on Artificial Intelligence of the UK House of Lords; United Kingdom; April; 2018.
- Beijing AI Principles; Beijing Academy of Artificial Intelligence (BAAI); China; May; 2019.
- Governance Principles for a New Generation of AI; Chinese National Governance Committee for AI; China; June; 2019.
- Social Principles of Human-Centric AI; Government of Japan; Cabinet Office; Council for Science, Technology and Innovation; Japan; March; 2019.
- AI Principles and Ethics; Smart Dubai; United Arab Emirates; January; 2019.
- AI Strategy; German Federal Ministries of Education, Economic Affairs, and Labour and Social Affairs; Germany; November; 2018.
- AI in Mexico; British Embassy in Mexico City; Mexico; June; 2018.
- National Strategy for AI; Niti Aayog; India; June; 2018.
- AI for Humanity; French Strategy for Artificial Intelligence; France; March; 2018.

Group 1: Members of Society

(a) Professional Associations

- Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (EAD First Edition); IEEE; United States; March; 2019.
- Principles for Algorithmic Transparency and Accountability; ACM US Public Policy Council; United States; January; 2017.

(b) Civil Society/Lawyer groups

- Top 10 Principles for Ethical AI; UNI Global Union; Switzerland; December; 2017.
- Toronto Declaration; Amnesty International; Access Now; Canada; May; 2018.
- Universal Guidelines for AI; The Public Voice Coalition; Belgium; October; 2018.
- Human Rights in the Age of AI; Access Now; United States; November; 2018.
- Future of Work and Education for the Digital Age; T20: Think20; Argentina; July; 2018.

(c) Non-profit organizations

- Asilomar AI Principles; Future of Life Institute; United States; January; 2017.
- Three Rules for Artificial Intelligence Systems by the CEO of Allen Institute for Artificial Intelligence; Allen Institute for Artificial Intelligence (AI2); United States; September; 2017.
- Principles for the Governance of AI; The Future Society; United States; July; 2017.
- Tenets of Partnership on AI; Partneship on AI; United States; Unkown; 2016.

(d) Academy

- The Japanese Society for Artificial Intelligence Ethical Guidelines; JSAI; Japan; February; 2017.
- The Montreal Declaration for a Responsible Development of Artificial Intelligence; University of Montreal; Canada; November; 2017.
- Three ideas from the Stanford Human-Centered AI Initiative; Stanford University; United States; Unkown; 2018.
- Harmonious Artificial Intelligence Principles; Chinese Academy of Science; China; Unkown; 2018.

Figure 2.8: Publicly available AI ethics guidelines issued by different stakeholders

In an attempt to discover the current ability of governments to exploit the innovative potential of AI, Oxford Insights created the Government Readiness Index in 2017, which sought to answer the following question: "how well placed are national governments to harness the benefits of AI in their operations and public service delivery?" And on 28 September 2020, the 2020 Government Readiness Index was published. The metrics used for this index are grouped into four broad pillars: government; infrastructure and data; technology sector.

In Table 2.1, we present the top 10 countries in this index, in addition to Brazil, which, in the 2019 edition was in 40th position, currently figures in 63rd position. Although China is an AI superpower, the researchers justify the country being in 19th position due to a lack of available data to correctly rank it.

Table 2.1: Index of Government Readiness towards Artificial Intelligence - 2020. Adapted from [9]

Country	Rank	Index
United States of America	1	85.479
United Kingdom	2	81.124
Finland	3	79.238
Germany	4	78.974
Sweden	5	78.772
Singapore	6	78.704

Republic of Korea	6	77.695
Denmark	8	75.618
Netherlands	9	75.297
Norway	10	74.430
...	...	...
Brazil	63	47.464

Oxford Insights' efforts to build these indices are extremely useful in assessing the ability of governments to exploit the potential of AI in the years to come, however, until 2019 this analysis did not take into account how robustly each country was contemplating the ethical issues involved in developing AI-based systems [44]. In 2020, a sub-index on the responsible use of AI was built, covering 34 countries, using the Organization for Economic Co-operation and Development (OECD) Principles on AI, published in June 2019, as pillars. However, this sub-index is a pilot project with several limitations, with few indicators and limited to factors such as number of surveillance companies based in each country. In Table 2.2 we see how the Scandinavian countries stand out in the top five positions, largely due to the data governance policies adopted by these countries, while Brazil is in 30th position among the 34 countries assessed.

Table 2.2: Responsible AI sub-index - 2020. Adapted from [9]

Country	Rank	Index
Estonia	1	79.852
Norway	2	77.201
Luxembourg	3	76.526
Finland	4	76.172
Sweden	5	72.975
Portugal	6	72.436
New Zealand	6	68.262
Denmark	8	66.873
Senegal	9	66.381
Uruguay	10	65.205
...	...	...
Brazil	30	42.358

Countries such as the United States of America and the United Kingdom are ready but not responsible countries, the authors present several factors, one of the most relevant being that the policies employed by these governments are largely favouring the interests of large technology companies over the interests of citizens. Russia and China are ranked 33rd and 34th respectively because, according to the authors, “both have developed a reputation for mass surveillance and restrictions on Internet freedoms”.

The European Union (EU) and the United States are leading the AI debate around the world – as can also be seen in Table 2.1, in addition to China, but each of them is conveying their own and different model of AI development and understanding [52], with the European Commission's Ethics Guidelines for Trustworthy AI, the Obama Administration's Report on the Future of Artificial Intelligence, and the Chinese Ministry of Science and Technology's Beijing AI Principles being guides published by these AI superpowers, respectively. The regions of the world that are marginalised regarding the debate of AI ethics regulation are: Central and South America, Eastern Europe, Africa and Central and Southeast Asia [52].

It is notable that several private organisations have seized on the public debate of AI ethics to indicate a commitment to ethics through value statements for responsible AI development. However, such statements raise more questions than answers, producing more conflicts, in an attempt to suggest how we should put it into practice [53]. Furthermore, according to Hagendorff [11] "ethics can also simply serve the purpose of quieting the critical voices of the public, while the criticised practices are maintained within the organisation." Most technology companies are not ensuring accountability for AI systems [54]. Commercial interests override public interests to the extent that the industry and its employees have duties towards their investors [30]. In short, private sector self-regulation will not be sufficient to address the ethical challenges of AI development [55].

There are no consequences for non-compliance with the various ethical codes [11], made explicit by the fact that the AI developer is not a formal profession [30] along with the absence of laws, occasioning AI-based systems that will not serve the public good [56]. In this sense, binding laws are paramount to effectively align public interests with practice in the development of applications in the AI context.

The documents presented in Appendix A and discussed in this section are termed **non-binding**. To clarify: "Reports and guidance documents for AI ethics are examples of what are called policy instruments of a non-binding or soft law" [27]. Among this group, the Ethics Guidelines for Trustworthy AI, published on April 8, 2019, by the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG) [42] – with 52 experts coming from different research fields, is worth mentioning. Aimed at providing practical and concrete guidance for developers of AI-based systems [57], taking into consideration aspects such as non-discrimination, dignity, privacy and protection of personal data, security and transparency [51] the AI HLEG document has influenced the AI ethics guidelines of other organisations, such as the Organization for Economic Cooperation and Development (OECD), which published its AI ethics guidelines one month after the release of the Ethics Guidelines for Trustworthy AI [57]. Another document we highlight is from the IEEE Global Initiative, which published in March 2019 the first edition of Ethically Aligned Design – A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (IEEE EADv1) [58]. In addition to being a high-profile professional technical organization, its document is one of the most significant published to date on AI ethics issues, proposing various means to mitigate them [44].

In contrast, documents that are **legally binding**, backed by legislation, provide the actors involved in the process with real responsibilities and binding rights. These types of documents are called binding or hard law. We will present the two most notorious binding laws. First, the General Data Protection Regulation – General Data Protection Rule (GDPR) [59] – of the European Union (EU), which came into force on 25 May 2018 and has been hugely influential in establishing safeguards for personal data protection in today's technological environment. Several countries outside the EU have adopted similar data protection rules, analogous to or inspired by GDPR, which is increasingly being recognized for its high standard of data protection, Brazil being one such nation. Aimed at robust EU citizens to have control over their data and protect them from data and privacy breaches, the GDPR applies to all relevant actors within the EU and those who process, monitor or store EU citizens' data outside the EU [51]. Second, the General Law on Personal Data Protection (LGPD) [60] in Brazil, which came into force on 18 September 2020, with the sanction of Law 14.058/2020, originating from Provisional Measure (MP) 959/20. The LGPD defines "rights of individuals in relation to their personal information and rules for those who collect and handle these records with the aim of protecting the fundamental rights of freedom and privacy and the free development of the personality of citizens." [61]. An effort towards harmonization between AI ethics guidelines (soft law) and legislation (hard law) is an important next step for the global community [27], however, it is beyond the scope of this paper.

2.2.2 Principles

While there is a profusion of ethical guidelines in AI, they remain separate and distinct from each other [10]. As presented in Figure 2.7, AI ethics guidelines contain ethical principles, and each published guideline contains its own set of principles. In the literature, most studies focus on the conceptual part of AI ethics, and one of them is the compilation, presentation and evaluation of ethical guidelines and their principles. Several authors have used different methodologies to explore sets of documents and extract the most recurrent principles and their definitions, usually concluding that they are too general, have high level of abstraction and degree of difficulty in applying them in real contexts, besides there is overlap between the principles [11], [27], [48], [49], [45], [62], [63], [10], [28]. Mittelstadt [30] and Whittlestone et al. [50] have criticized the principles as a way of approaching AI ethics.

Floridi and Cowls [62] initiated the debate on how ethical principles in AI can be synthesized into only 5, with 4 coming from classical biomedical principles – beneficence, non-maleficence, autonomy and justice –, added to a fifth principle, explicability, exhibited as a new enabling principle for AI. In order to address the established principles as a unified framework, the authors analysed only 6 guidelines. The European Commission High-Level Expert Group on Artificial Intelligence (AI HLEG) published the Ethics guidelines for Trustworthy AI [42], which has a great impact on AI ethics research and was influenced by Floridi and Cowls' unified framework [62]. Worth noting that, Floridi is a prolific author in the field of AI ethics, also with distinction in works that analyse how to bridge from theory to practice in AI ethics [13].

Smit et al. [45] besides extracting several principles from a collection of published AI ethics guidelines, mapped the results according to the principles presented by Floridi and Cowls [62], finding that all the identified principles could be mapped against at least one of the principles of this unified framework. The authors examined the principles at the design level, i.e. design principles, extending the understanding of design principles in relation to AI design and execution.

Ryan and Stahl [10] conducted a rigorous study with a robust methodology that reviewed a set of guidelines and compiled the detailed guidance that is available, presenting a list of principles aimed at developers and users. To the best of our knowledge, this is the study that makes use of a methodology that encompasses the most guidelines and definitions – as well as presenting a comprehensive taxonomy.

Both the study of Smit et al. [45] and that of Ryan and Stahl [10] aim to express the principles aligned with the goal of this work, the former being directed to principles at the design level, and the latter directed to principles for developers. As the amount of principles exposed by Smit et al. are 22 and by Ryan and Stahl are only 11, and because we understand that the latter study contains more accurate and comprehensive definitions, in this section we will present the definitions of the principles exposed by Ryan and Stahl [10].

The principles presented in: 1) Ethics guidelines for Trustworthy AI of the European Commission [42], 2) Smit et al. [45], 3) Floridi and Cowls [62], and 4) IEEE EAD 1st Version [58] are present in Appendix B.

We present below, the principles and a brief description (overview), followed by the synonyms or constituent ethical issues for each of the eleven principles, according to the work of Ryan and Stahl [10]:

Principle 1 - Transparency

One of the most widely discussed principles within the AI ethics debate, transparency is fundamental to enabling the other principles to exist. Organizations should be transparent with their goals for using AI, what the possible outcomes, benefits and harms are. In addition, be able to clearly interpret and demonstrate how their AI complies with current legislation, such as the General Data Protection Regulation (GDPR) or the General Data Protection Law (LGPD), and what measures are being taken to ensure compliance.

Teams developing AI-based systems must be able to intelligibly explain the incoming data, the outgoing data, what their algorithms do and their purpose for doing so, i.e. understand how AI works and explain the technical functionalities and decisions, ensuring traceability and explainability. It is on this principle that the degree of opacity of AI algorithms defined in 2.1.1 should be understood, and the black box, shown in Figure 2.3), interpreted. In addition, AI-based systems should be subject to active monitoring to ensure that they are producing accurate results, and be explainable to external auditing bodies to ensure their technical and ethical functionality.

End users should be clear that they are interacting with an AI-based system, rather than a human, and what the intentions and outcomes of the technology are, and be provided with accurate information to ensure they are not being manipulated, misled or coerced by AI. It is also noted that the AI-based system should be designed and used to retrieve little personal data, or that it should be anonymised, encrypted and securely processed.

Ethical issues related to Transparency principle: transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing.

Principle 2 - Justice and fairness

The second principle has been widely debated in the media and academia, as it brings issues such as discrimination and injustice to minority, vulnerable and underrepresented social groups. Organizations should promote the inclusion of women and minority groups in the development and design of AI-based systems – greater diversity in software development teams – in order to reduce issues of exclusion. In addition to promoting the equality, empowerment and benefits of individuals, finding ways to identify and mitigate unfair bias and discrimination, ensuring mechanisms for reversibility and redress of outcomes when harm occurs.

Developers of AI-based systems should identify levels of fairness and equality during the design phase, taking steps to ensure that the data used by these systems is not unfair, or contains errors that will corrupt the decisions made by the systems. Also, to promote equality, the system should be designed to fit this purpose, identifying impacts on different aspects of society and be designed to promote human well-being. Developers should implement mechanisms to prevent, remedy and reverse discriminatory outcomes arising from the use of AI, being designed for universal use and non-discrimination as to persons, group of persons, based on gender, race, culture, religion, age or ethnicity.

End users have the right of access to the data stored and used about them, and can also request that data be rectified or deleted, and are offered easily accessible explanations of decisions taken.

Ethical issues related to Justice and fairness principle: justice, fairness, consistency, inclusion, equality, equity, non-bias, non-discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution.

Principle 3 - Non-maleficence

This principle gained recognition with the 4 principles of biomedicine, and, in AI ethics with the examples of autonomous cars, unmanned aircraft and autonomous weapons, being outlined as not doing or preventing harm from occurring to others. Organisations must ensure cyber security – AI systems protected from attack – and conduct regular testing to deliberate that harm does not occur from their use.

Developers of AI-based systems must have the necessary skills to understand how systems work and their potential impacts, ensuring reliability and safety, implementing mechanisms to protect user safety, and ensuring that the AI system is safe and reliable through its life cycle. Developers must ensure that data is retrieved, analysed and used in such a way that the system being designed will not cause harm to people. Security should be implemented and tested in the design phase, while in the development phase, the objectives and expected impacts of the systems should be tested and precautionary measures documented. This is to say that, the AI-based system should pass quality assurance processes and be tested in real scenarios before, during and after its deployment.

Ethical issues related to Non-maleficence principle: non-maleficence, security, safety, harm, protection, precaution, prevention, integrity, non-subversion.

Principle 4 - Responsibility

Identification of a responsible and culpable party for incidents of the use of AI-based systems is often overshadowed, by the autonomy of these systems, creating a liability gap, making this identification difficult. A legal person should always be held accountable for damage caused by the use of AI-based systems, with blame not being placed on the systems that caused the damage, and implement ethical training to ensure responsible development and implementation of systems. Organizations should clearly and concisely assign responsibilities within the organization. Also, be held accountable for the use of poor data if there are adverse consequences of its use.

Developers need to be aware that they are responsible for the impact of systems on the world and are primarily responsible for the design and functionality of those systems and, if they discover bugs, security breaches or data leaks, they should report these issues to the appropriate authorities, stakeholders and, where relevant, the general public. One way to achieve accountability, where undesired effects exist, is through documentation and the use of recording systems .

Ethical issues related to Responsibility principle: responsibility, accountability, liability, acting with integrity

Principle 5 - Privacy

Due to the extensive use of data by AI-based systems, privacy and data protection is a key principle, being heavily debated after the GDPR [59] and LGPD [60] came into force.

Developers should ensure the privacy of end-users' personal data from the beginning of the design process. The collection and use of end-users' personal data should be kept to a minimum unless absolutely necessary and relevant. In addition, developers must process both personal data and data derived or created from end-users in a fair, legitimate and lawful way.

Ethical issues related to Privacy principle: privacy, personal or Private information.

Principle 6 - Beneficence

Another principle that has gained recognition through biomedical principles, beneficence denotes doing good, having the intention to benefit someone or society as a whole by performing an activity. Organisations developing AI solutions should use the data obtained for the benefit of customers and society, finding solutions to world problems, the prevention of damage to the environment and curing diseases, as well as promoting peace by ensuring that the AI-based systems are designed to benefit humans, outlining the benefits and the stakeholders who will benefit from them.

Ethical issues related to Beneficence principle: benefits, beneficence, well-being, peace, social good, common good.

Principle 7 - Freedom and autonomy

This principle addresses how to ensure the protection of freedom and the promotion of autonomy that users of AI-based systems should have. These systems should neither compromise human freedom and autonomy, nor illegitimately and covertly reduce citizens' options and knowledge. Organisations must ensure that end users are well informed, not misled or manipulated by AI systems and must be allowed to exercise their autonomy.

Developers of AI-based systems must ensure that they do not cause harm to end users through censorship (freedom of expression), tracking (freedom of movement) or surveillance (freedom of association).

The end user must consent to the use of personal data, which must be clearly articulated, prior to use. Their personal data cannot be processed in a way that they consider inappropriate or objectionable. In other terms, end users must be well-informed actors who have control over their decisions when interacting with AI systems.

Ethical issues related to Freedom and autonomy principle: freedom, autonomy, consent, choice, self-determination, liberty, empowerment.

Principle 8 - Trust

Trust is a principle that has gained notoriety recently following the publication of the AI HLEG in 2019, outlined as a fundamental precept for society to function. Organizations must prove that they are trustworthy and that their technologies are reliable. Developers can cultivate trust by ensuring accountability, transparency and safety of AI-based systems. End users must be able to trust the organisations developing AI solutions and that their systems work as desired.

Ethical issues related to Trust principle: trustworthiness.

Principle 9 - Sustainability

With less relevance in most guidelines, sustainability runs through all areas and disciplines, with more strength these days with global warming, and its importance in AI is no different. Organizations must ensure they are environmentally sustainable, incorporating environmental outcomes into their decision-making.

Software development teams should strive for energy efficiency, low resource consumption, protection of biodiversity and reduction of greenhouse gas emissions.

Ethical issues related to Sustainability principle: sustainability, environment (nature), energy, resources (energy).

Principle 10 - Dignity

This principle concerns the recognition of the value of individuals and the respect for their rights, and is a substantial principle. Developers of AI-based systems should take into consideration the intrinsic values of human beings during the design and use of AI, respecting, serving and protecting their physical and mental integrity, as well as their sense of personal and cultural identity, and the satisfaction of their basic needs. Furthermore, developers should make it clear that end users are interacting with an AI-based system, not another human being. This is the only principle where the related ethical issue is singular, this being homonymous to the principle.

Ethical issues related to Dignity principle: this is the only principle that ethical issues related to it, is it itself.

Principle 11 - Solidarity

The solidarity principle is about facilitating and promoting human development, social relations, security and social cohesion, and not undermining, obstructing or endangering democratic values, social relations and human development; and its use should contribute to global justice. Organisations should promote fair distribution of benefits from AI to ensure that social cohesion is not undermined.

Teams developing AI-based systems should not develop or use these systems so that they intentionally undermine democratic systems of government, carry out dissemination of fake news, or surveillance and invasion of privacy of individuals, which can lead to the weakening and compromising of social relations and solidarity.

Ethical issues related to Solidarity principle: solidarity, social security, cohesion.

Since AI ethics research has been mainly theoretical and conceptual, and given that this discussion of principles is still so active, it may explain the fact that few attempts to bring them into practice have been made [64]. There is a wide degree of convergence between the principles described by these guides [27], however, there is a large gap between the articulation of these high-level, abstract concepts and their real-world application [48], [10]. Vakkuri et al. [14], mentioned that organisations do not make use of formal tools or methods to implement ethics. Thus, in this practical context, there is a need for methods or tools created specifically for the context of AI ethics.

2.2.3 Practice

Ethical principles in AI are not automatically translated to practice, and there is also a lack of proven methods to make this translation in real development scenario [30]. These principles are often broad and abstract, and the higher the level of abstraction of ethical values, the more interpretation is required to perform the translation to practice [45]. It is noted that translation "involves the specification of high-level principles into medium-level standards and low-level requirements" [30]. In this Section, we will address the practical aspects of AI ethics, involving its challenges, drawbacks and possible solutions, and considering Figure 2.7 as the guide of this study, we will present the methods and tools in relation to the Software Development Life Cycle (SDLC) and AI ethical principles presented in Section 2.2.2.

In 2019 a greater focus has begun, among academia, on how to carry out this translation of ethical principles into practice, that is, the translation from "what" to "how" [65]. There is a clear difficulty in accomplishing this task. We raise some questions, with the principles presented in Section 2.2.2 as the guiding thread, how can we implement: beneficence, well-being, trust, dignity or solidarity in AI-based systems? Ethical principles serve as an aid to structure ethical impact analyses, however performing the translation of these principles into implementable strategies is a challenge [43].

Despite this visible and flagrant difficulty in translating theoretical and abstract principles into concrete practices, several tools and methods to implement ethics in AI exist, and surprisingly, the vast amount of available tools by itself is already a drawback [66]. In the literature, we found works that performed searches for publicly available methods and tools that assist the implementation of AI ethics in different phases of the Software Development Life Cycle. In Newman's study [65] 6 tools and 12 best practice frameworks were evidenced, and in Morley et al. [13] 106 tools were found, presenting a wide list of tools and methods and flag some significant challenges, such as: (a) the tools included are relatively immature, have little documentation and are not necessarily ready for use, resulting in little usability by developers and additional work to be put into practice; (b) difficulty in assessing their scope of use; (c) difficult to encourage their adoption by the practical mind of AI and ML developers; d) the vast amount of available tools makes it difficult to evaluate and choose, being difficult to compare one with another, despite there being articles or public repositories on GitHub, if the tool is not supported by a community, with active users – both developers and scholars – public availability of the source code and ample documentation, there will not be an adherence nor usefulness of it [66].

The Artificial Intelligence community, more specifically the Machine Learning community, has focused primarily on tools that seek to implement the Transparency principle. In other words, most tools and methods developed to implement ethics in AI focus on fulfilling the Transparency principle, on the ethical issue of **explicability**, that is, they are tools that seek to explain the operation of AI algorithms such as Machine Learning and Deep learning [13].

These tools are found in phases 4 and 5 of the SDLC, testing and implementation and maintenance, respectively. Morley et al. [13] point out that "the existence of these tools is necessary but not sufficient". The work of Arrieta et al. [67] introduces the term eXplainable AI (XAI), as a set of ML techniques that "produce more explainable models while maintaining the level of learning performance and allowing humans to understand, trust and manage emerging generations of AI". Some examples of XAI tools are interpretML [68], LIME [69] and SHAP [70].

Different authors and AI ethics guidelines point out that one solution to accomplish the translation of principles into practice are **checklists**, in which developers tick the boxes in which they perceive as ready regarding ethical principles or answer questions regarding ethical issues – focusing on the early phases of the Software Development Life Cycle –, with the perspective that professionals make ethical decisions along the SDLC [71]. However, they can be misused or even ignored if practitioners are not involved in their design or implementation [71], moreover, checklists should not be the only mechanism for ethics in AI [11].

Madaio et al. [71] put the imprint that checklists have the characteristic of being more related to the principle of justice and fairness in specific. However, their principle context is broader than the one mentioned by these authors. The highest profile checklist example is the HLEG Assessment List of the Ethics Guidelines for Trustworthy AI [42], in the public area. Other examples of such tools include: Deon [72] in the private area; Ethics & Algorithms Toolkit [73] by nonprofit organisation; private-public partnership (Microsoft and Carnegie Mellon University) [71]; and in the public area the Ethics Framework [74], from Digital Catapult, the innovation agency for the UK Government's digital and software industry. These applications of checklists have another disadvantage as they are often created for a particular context, public, or private, or for specific development teams [71].

While checklists are used in the early phases of the Software Development Cycle, impact assessment tools are employed in phases 4 and 5 of the SDLC: testing, implementation and maintenance. They can be used to monitor and test AI-based systems [75], being related to the principles of accountability [48] and explicability [76]. Several authors have elaborated impact assessment tools (e.g. IEEE 7010 [77], Well-being impact assessment [66], Enhanced Well-being Impact Assessment (GoodAI and Accenture) [75], Multi-layered explanations of GDPR Algorithmic Impact Assessments [76], Algorithmic Impact Assessments (AI Now Institute) [78]), sharing as a point of convergence the need for this tool, within or outside a specific framework, to be seen as a continuous, iterative process, and not just static or done once, but done throughout the life cycle of the system.

In the work of Schiff et al. [66], an impact assessment tool called well-being impact assessment for ethical AI-based systems was built on top of the IEEE 7010 standard [77]. This tool involves the iterative implementation of: “1) internal analysis, informed by user and stakeholder engagement; 2) development and refinement of a well-being indicator panel; 3) data planning and collection; 4) data analysis of evaluation results that could inform improvements to A/IS” [66]. The authors claim that this tool based on the IEEE 7010 standard approach has the potential to be broad, operationalisable, flexible, iterative, guided and participatory. Furthermore, impact assessment involves measurements, not just assumptions, about impacts, so data collection is key, extrapolating data collected related only to the development of AI-based systems, but being collected through surveys, focus groups or directly as output from systems [66].

With the need to clarify terminology, we note that often impact assessment tools have characteristics of checklists, but to differentiate them, we suggest that impact assessment tools can be used after the system is deployed, having data generated from the interaction with the user, aiming exactly at assessing the impact that this AI application exerts in relation to the ethical principles adopted when it is put into operation, while checklists are tools for the early stages of development, prior to deployment, thus the checklist-based approaches previously presented “do not provide for continuous monitoring over time, in particular, in the deployment phase of the Software Development Life Cycle of the AI-based system to be developed” [75].

Several companies have proposed practices involving training, hiring, algorithm development tools and frameworks, and governance strategies [66]. Telefónica, a large Spanish company, created their set of ethical principles in AI, and to implement their guidelines, they created a methodology called Responsible AI by Design [36], which consists of a series of tools to implement their principles. Some of the tools developed suggest considerations for the principles of accountability – auditing and mitigating bias in any existing ML rating model – transparency, privacy, and explainability (with their own XAI tool), which have applicability after the system has been deployed.

Google has also published efforts to carry out the translation of its AI ethics principles into practice. Regarding the Accountability principle, researchers at Google and the Partnership on AI have developed a framework focused on internal audit, with the claim that external audit only happens after system deployment, which would be after potential damage has occurred [79]. Importantly, the external audit would happen in phase 5 of the Software Development, Implementation and Maintenance Cycle, while the internal audit, according to the authors, would happen throughout the other phases. The authors then presented the SMACTR framework, serving to guide practical implementations, verifying whether the engineering processes involved in the creation and development of AI-based systems comply with the ethical principles in AI of the organization, having at its core an audit checklist. Internal auditing does not bring the other principles mentioned in Section 2.2.2 to society, in small or large scale, because Transparency – seen as the principle that allows the others to occur –, is compromised.

IBM also released its practical tool for implementing ethics in AI, AI Fairness 360 [80]. The authors presented an extensible and open-source toolkit, publicly available on the developers' GitHub. This tool aims to detect, understand and mitigate biases generated by algorithms, making it possible to examine datasets for biases through metrics, as well as techniques to mitigate these biases, in other words, they present: bias metrics, algorithms for bias mitigation, explanations for bias metrics, and industrial usability. Therefore, it has main focus on the principles Justice and fairness and Transparency (explainability), enabling the Beneficence principle; acting in the most advanced phases of the Software Development Cycle, such as 3, 4 and 5. Although IBM has published its own AI ethics guidelines, the authors have not made it clear on which guidelines they are basing their work.

Bellamy et al. [80] presented some important terminology for the fairness context, such as **protected attribute:** "partitions a population into groups that have parity in terms of benefits received", are not universal but application-specific (e.g., race, gender, religion); **privileged value of a protected attribute:** indicates a group that has a history of systematic advantage; **individual equity:** is the goal of similar individuals receiving similar treatments or outcomes; **group equity:** is the goal of groups, defined in the protected attributes, receiving similar treatments or outcomes; **bias:** a systematised error that places privileged groups at a systematic advantage and unprivileged groups at a systematic disadvantage; **equity metric:** is the quantification of unwanted bias in training data or models; **bias mitigation algorithm:** is a procedure to reduce unwanted bias in training data or models.

This last example from the private sector shows us a growing interest from industry and academia for tools that implement fairness. In fact, just like Transparency (explainability), there is a vast amount of effort to implement Principle 2 – Justice and fairness. This is due to the fact that principles like Transparency (in its ethical question of explainability) (1), Fairness (2), Non-maleficence (in its ethical question of safety) (3), Responsibility (in its ethical question of accountability) (4) and Privacy (5), share the characteristic that they "are more easily implemented mathematically and therefore tend to be implemented in terms of technical solutions" [11]. Other examples of tools and methods to implement fairness include: [81], [82], [83], [84], [85], [86]. While such tools focus on fairness, many include explicability techniques in their efforts (e.g., explaining why a dataset contains bias), again signalling how this principle (Transparency) allows other principles to occur. In addition to explainability, fairness tools can also assess the fairness of models through tools using audit tools for predictive models, as in the case of FairML [84]. This combination of tools focusing on specific principles but making use, for their final goal, of methods comprising other principles, is explained by the high level of abstraction and the existing overlap between the principles. This interweaving of principles such as fairness and accountability with XAI (explainability) is also discussed in the work of Arrieta et al. [67], where the authors point out that the proposed XAI tools can be used for bias detection, by facilitating the understanding and measurement of bias, as they can be used for the Accountability principle, by "helping explain AI-based systems from different profiles, including regulatory ones".

Sharma et al. [86] proposed CERTIFAI, a generic tool that can be applied to any black-box model and any type of input data, examining issues such as robustness, interpretability, transparency and fairness, illustrating this interweaving of tools with multiple principles. The main difference in techniques implementing fairness to those implementing explainability, is that the latter explicitly take into account the terminologies presented (e.g., protected attribute and bias), and may contain bias metrics and techniques to mitigate bias.

We present some technical solutions that several authors are working on to address issues such as fairness and explainability, however, their restricted scope is a problem, because by addressing issues related to discrimination, racial bias, inequality, among others, with an extremely technical and mathematical approach, the proposal may fail to consider how an AI-based system will be implemented in different social contexts or influence human behaviour, in other words, "there are many questions about ethics in AI that cannot be directly solved with a restricted technical lens" [66].

Mittelstadt [30] mentioned that the "practical normative requirements should somehow be incorporated into the development processes and functionally implemented into the design requirements". There are methods, such as Value-Sensitive Design, that incorporate values and ethical principles into the design and development cycle, introducing values and relevant stakeholders into the development process, allowing values to be introduced into system design, however, they are not very functional, "have difficulty capturing the degree to which the resulting artifact reflects particular values or specifications". Furthermore, due to the nature of the business of commercial organisations, Value-Sensitive Design would encounter resistance to be implemented, as it inhibits the efficiency and profit of these organisations. Some authors have presented studies on the applicability of Value-Sensitive Design as a method to implement ethics in AI-based systems, such as [87] [88] [43] [89] [53] [90], or as part of the Requirements Engineering process [91].

Practical frameworks for AI ethics should take into account the complexity of the development and deployment of AI-based systems and their socio-technical nature, in addition, the responsibilities and requirements of the different stakeholders involved in the development, deployment and evaluation processes [39]. Because of the difficulty of translating the various ethical principles into practice and into every application context, there is a demand for a framework acting as a process, in other words, "it is impossible to provide ethical principles that will be specific enough to provide answers in practice, and yet broad enough to be applied universally, but it is possible to provide a process" [43], which converges with the following statement: "ethics is a process, not a destination" [30]. Additionally, "it is crucial for any AI ethics framework to simplify, but not oversimplify, and provide guidance appropriate to the requirements of each stakeholder" [39].

Some examples of frameworks for implementing AI ethics include: [92] and [79] from the private sector; [39] from academia; [43] academia and public sector partnership; [93] from the UK government; and [94] from the public sector.

Schiff et al. [77] suggest 6 points that a framework for implementing ethics in AI should take into account, and we present an adaptation to our context:

1. **Broad:** should consider broad aspects of the impact of AI-based systems to be developed, from different principles and ethical issues to social and economic contexts. Software designers can identify which narrow-scope tools are appropriate, such as those for fairness and explainability, and use them as a sub-part of this framework;

2. **Operationalizable:** users of this framework should be able to articulate the ethical principles, in addition to other desirable requirements of the software to be developed, into specific strategies that can be implemented in AI-based systems, encompassing the identification of relevant actions and decisions assigned to the appropriate phase of the Software Development Life Cycle;

3. **Flexible:** the framework should be adaptable, adjusting to different use cases, implementation context, organizational settings, and different types of AI-based systems, with the intention of having greater applicability, as well as allowing satisfactory customization and sharing of language and learning. In other words, "there will always be ethical decisions and trade-offs that are not amenable to universal application, and need to be made with sensitivity to a specific context and stakeholders" [43];

4. **Iterative:** the framework can be applicable throughout the entire Development Cycle of the AI-based system to be developed, and in an iterative way. This is because implementing AI ethics is not done just once, changes may occur, e.g. in the system itself or in the implementation context, so there is a need to involve

the different stakeholders at each stage, and be re-evaluated over time, and as new issues arise [43].

5. **Guided:** must be easy to use for the users of the framework. Users must be able to easily access and understand the framework, as well as: apply, customise and solve possible problems, across different contexts. The framework must provide sufficient documentation for this;

6. **Participatory:** the framework should incorporate the perspectives of different stakeholders, especially those who may be impacted by the AI-based system developed, the public.

Using the 5 principles indicated by Floridi and Cowls [62] (beneficence, non-maleficence, autonomy, justice and explicability), Peters et al. [43] presented two frameworks for AI ethics that make use indirectly and to some extent of the points set out in the work of Schiff et al. [66], with a case study in the digital mental health context. We only present the first framework: Ethical Design Process. The authors conceived of this ethical development process by incorporating existing design processes – in particular anticipatory, reflective, inclusive, and responsive design processes – and supplementing these with methods for ethical analysis and design aimed at human well-being, and may involve activities used in Value-Sensitive Design. To present a roadmap for a responsible development process, in which the final product has ethical characteristics (contemplating adhered ethical principles), 5 development phases – research, insights, ideation, prototypes and evaluation – were adopted, and we present them adapted to the context of our study:

1. **Research:** Involves investigating the needs, preferences, contexts and lives of the people who will be served or impacted by the AI-based system to be developed. Possible approaches to conduct the research phase include: expert review, secondary research in relation to the specific domain, as well as Design Thinking methods, ethnographies, participatory workshops, among others;

2. **Insights**: Involves the analysis and synthesis of the data obtained in the previous phase into project-specific insights. Data analysis can be done to anticipate potential harms (e.g., biases, ethical risks) and support opportunities, through the adopted ethical principles;

3. **Ideation:** Involves the divergent generation of ideas for design solutions by brainstorming, in which domain-specific ethical concepts are introduced (e.g., concept of well-being in psychology, in a digital mental health system). The development team will focus on the ethical issues – domain-specific – that arise, integrating the ethical reflections raised and their possible solutions at this stage;

4. **Prototypes:** In this phase, the development team converges and builds various design solutions. This phase involves an extensive range of stakeholders, including the end-user, so that there is collaborative speculation on ethical well-being impact analysis, considering the ethical impacts (negative or positive) that a project may produce;

5. **Evaluation** (In Use): Involves the evaluation of the impact of using the AI-based system on the user, but also on society and the planet, during and after use.

In sum, a framework should be a systematic and rigorous process in which development teams can consciously make decisions and record the values and reasoning employed in making those decisions [43]. Such records can serve both to address the principle of Transparency with the public and to grant the development team confidence that the decisions were made through a systematic and professional approach. While frameworks – or the other ways of implementing AI ethics presented – cannot guarantee that an AI-based system will not have negative real-world consequences, they can help to "mitigate risks and provide practitioners with the assurance of having acted responsibly" [43]. In addition, the guide to be designed should be tested in different contexts and the evidence of effectiveness publicly disclosed [66].

Developers play a crucial role in implementing ethics in AI-based systems, for the reason that AI development is still software development [14], besides being fundamental in the phases of Software Engineering (i.e., requirements elicitation and specification, development, verification and validation, maintenance and evolution) [26], however, by the commercial logic of the organisations developing these systems, developers "are not systematically taught about ethical issues, nor empowered, for example by organizational structures to bring ethical concerns" [11]. In other words, developers of AI-based systems – prepared to deal with technical challenges – need direct guidance on how to deal with ethical dilemmas (i.e., how to implement ethical principles in AI) [39].

The purpose of our study is to develop a practical Guide to AI ethics for developers of AI-based systems, focusing on the requirements analysis phase, in this way, it should be taken into account the responsibilities and requirements of developers of AI-based systems, in addition to the usability provided by this guide to developers of these systems. In other words, beyond a technical practice, our Guide should take into account an adaptation of the organizational practice, through adjustments in the realization of requirements elicitation.

Largely, it is noted that the tools and solutions presented to undertake the implementation of ethics in AI do not have their focus on the early stages of the Software Development Cycle – they focus on small parts of the development process [14] –, and are often applicable only when the software is already deployed, as are most fairness and explainability tools [13]. Thus, as with quality requirements in traditional software development, e.g. privacy and security, ethical principles and issues in AI-based systems have a much lower cost when addressed and dealt with in the early stages of the Software Development Cycle, than after deployment [14]. Ethics in AI should be in the requirements, crafted so that they are understood by developers [6]. For this reason, we should place a focus on Requirements Engineering, as a process within Software Engineering.

The Requirements Engineering community focuses efforts on producing and using AI-based systems, which are tools that automate decision making, during the Requirements Engineering process [95]. The use of AI for Requirements Engineering is referred to in the literature as **AI4RE** [96], and some tasks that these tools automate include requirements elicitation, prioritization, requirements refinement into specifications, ambiguity detection, relevance analysis, interpretation and classification of requirements written in natural language into functional and non-functional [95] [97]. The most used Machine Learning techniques in AI4RE are Natural Language Processing techniques, because the requirements are described in textual form [97]. Several authors have published papers related to the use or presentation of AI techniques to mitigate Requirements Engineering issues [97] [98] [99] [100] [101] [102] [103] [104] [105].

While the use of AI for Requirements Engineering is being widely researched, a limited number of studies are investigating how the change of Requirements Engineering and Software Engineering processes occurs in the development and design of AI-based systems [95] [12]. The Requirements Engineering process for AI-based systems is called **RE4AI** [96].

Vogelsang and Borg [12] stated that requirements engineers should be aware of the new types of requirements introduced by the ML paradigm – ethical requirements – that contemplate the ethical principles presented. Due to its interdisciplinary nature, Requirements Engineering is an opportune place to address ethical issues, involving the development team, project managers, stakeholders and end users, and is performed in the first phase of the Software Development Life Cycle, as presented in Figure 2.7.

In Section 2.3, we present in more depth the Requirements Elicitation process, the ethical and legal requirements, and the challenges of Requirements Engineering and Software Engineering in the context of ethical AI-based systems and their possible solutions.

2.3 Requirements Elicitation

Requirements are a reflection of user needs for a system, which must serve a purpose (e.g., plot the fastest route on the map, not disclose user data to third parties). Requirements for a system are defined as the description of the services that a system must provide, plus the constraints on its operation [106]. This Section addresses the first phase of the Software Development Life Cycle, requirements analysis, as presented in Figure 2.7.

Requirements are defined by users, who often do not have the necessary technical knowledge to achieve a high level of detail of the requirement to be raised. Therefore, there are different levels of requirements description. Sommerville [106] distinguishes between user requirements and system requirements in order to highlight this fact. User requirements are high-level, abstract, natural language descriptions of what the user expects the system services to be able to provide, and constraints on which it must operate [106]. System requirements are more detailed descriptions, and should define exactly what is to be implemented, such as functions, services and operational constraints [106]. This need for different levels of detail when describing requirements is due to the fact that different types of users use them in different ways [107].

In addition to the distinction between the levels of detail of requirements, they are classified into two groups: functional requirements and non-functional requirements. Functional requirements are statements of what the system should provide, and how it should react and behave to particular situations, as well as what it should not do [106]. Non-functional requirements generally describe constraints on a system's functionality or services (e.g., the maximum query time must not exceed 5 seconds, the system cannot be down more than 5 minutes in a day) [106].

There is no precise separation between the types of requirements, revealing that the requirements are not independent of each other, and that one can generate or limit other requirements. In the case of a user requirement related to security, such as limiting access only to authorized users, which may appear as a non-functional requirement, when developed with a greater degree of detail can generate requirements that are functional, such as the need for the inclusion of the user authentication service in the system [106].

To refer to the processes of identifying, analysing, documenting and verifying these services and constraints, is to define Requirements Engineering (RE). Requirements Engineering is generally seen as the first stage of the software engineering process [106]. There are three main activities involved in RE: elicitation and analysis – discovery of requirements by interacting with customers; specification – conversion of the requirements into a standard form; validation – checking that the requirements actually define the system the customer wants. These processes are iterative and interleaved, producing at the end the system requirements document [106]. In this dissertation, we focus on the activity of requirements elicitation in the context of ethics in software based on artificial intelligence.

The goals of requirements elicitation are: a) to understand the work that users perform; and b) how a new system can assist this work [106]. Requirements elicitation comprises the set of activities that enable the discovery, understanding and documentation of the objectives as well as the reasons for developing a software system [108]. It is during this phase – of requirements elicitation – that software engineers work together with users in order to understand issues related to the application domain, work activities, the services and features of the system that are desired by stakeholders, in addition to necessary aspects such as performance, privacy, among others [106].

Stakeholders generally do not know what they want from a software system, because they are mostly from non-technical background, and usually describe in general terms and with difficulties in expressing what they expect. There is a hindrance for the software engineer in understanding the user domain, due to the stakeholders' expression in their own terms, in which there is no (nor is expected) a previous expertise (e.g., a jurist describing a judicial system). As there are different stakeholders and different requirements, the software engineer needs to identify the different sources of requirements, and their similarities and differences. Due to the dynamic nature, political and economic factors influence requirements, sometimes by specific requirements from managers or by unavoidable changes in requirements that may occur during this phase. Thus, there are several difficulties perceived and faced during this process [106, 109, 110] [111] [112, 46, 113].

Requirements Engineering in Agile Development

AI-based systems are increasingly being built using Agile Software Development methods [114], as occurs at Microsoft [115]. The most widespread Agile Software Development (ASD) methods are Scrum and Extreme Programming [116], having as a common base a rapid iteration of the entire software development process, going through the entire Software Development Cycle [117]. While our report on Requirements Engineering has stayed on the traditional way of performing it (i.e., iterative execution of activities such as elicitation and analysis, specification and validation), with system requirements documents as the basis of communication, Requirements Engineering in ASD is not well defined, and there is also no certainty about how much Requirements Engineering in ASD differs from traditional [116], also seen as informal and based on the skills and knowledge of individuals, and more flexible and reactive than traditional [118]. In sum, Requirements Engineering in ASD is still imprecise, both for academia and developers [119].

Curcio et al. [119] performed a systematic mapping of Requirements Engineering in DSA, classifying the studies found according to 13 SWEBOK sub-topics, one of them being Requirements Elicitation. The authors found 7 articles related to this topic, and the techniques explored include: aspect-oriented approach, JAD, prototyping, mind-maps, query-based requirements engineering, simulations and gamification. The authors claim that these techniques can help motivate Requirements Engineering activities, provide documents and improve the quality of communication between team members.

Regarding Scrum, it is open to developers to choose the methods, techniques and practices of software development [120]. Only one person is responsible for eliciting requirements and prioritizing them, that is, the Product Owner (PO). As well summarized by Heikkila et al. [116]: “Requirements in Scrum reside in a product backlog, which is a prioritized list of all the work items envisioned for the software, which may also include technical improvements. The work items in the product backlog are called backlog items. Only the PO can add new items to the backlog. The PO works with a development team of five to nine cross-functional software developers”. This means that requirements in ASD are items (or user stories) that make up the product backlog, which are worked on by the development team in iterations, called sprints, that last from 1 to 4 weeks. The items in the product backlog – requirements – are discussed, better understood, and reprioritized in meetings [119]. In addition, it is important to note that the conduct of the requirements analysis, specification and validation phases are performed by the PO and the development team informally and collaboratively [116].

The most used techniques for Requirements Elicitation in ASD are [120]: interviews, brainstorming, ethnography and use case analysis. One of the biggest disadvantages of agile requirements engineering is the lack of documentation, justified also by the minimalist format of user stories, in which there is a difficulty in validating their consistency and verifiability [119].

2.3.1 Ethical Requirements

The **ethical requirements** are “requirements for AI-based systems derived from ethical principles or ethical codes (norms)” [23]. In our study, we call ethical codes ethical guidelines, so ethical requirements are derived from non-binding laws (e.g., AI HLEG, IEEE EADv1). These requirements are similar to **legal requirements** [121]), as they are requirements derived from laws and standards, in this way, legal requirements are derived from binding laws (e.g., LGPD, GDPR). The relationships between binding and non-binding laws and their principles are presented in Figure 2.7.

Ethical issues should be included in the Requirements Elicitation phase, since ethical requirements allow ethical issues to be considered from the beginning of the development process of AI-based systems, ensuring a focus on ethical aspects during requirements validation [23]. To elicit ethical requirements, Guizzardi et al. [23] extend the concept of users from traditional Requirements Engineering to **runtime stakeholders**, which includes stakeholders who are using, are affected, or influenced by the results of a running AI-based system. In the context of a digital mental health system, runtime stakeholders comprise patients, their families, physicians, and psychologists. In addition, ethical requirements are functional and non-functional requirements elicited from the runtime stakeholders and according to a set of adopted ethical principles [23].

2.3.2 Requirements Engineering for AI

The process of Requirements Engineering for AI-based systems (**RE4AI**) is different from traditional systems [95], and there is an additional complexity to the development of AI-based systems [122], because there is a dependency between the large amount of data and algorithms [46], being observed in some cases the use and extension of already well-established approaches, principles and tools in Software Engineering for the development of AI-based systems [114]. Some authors have explored the challenges of Requirements Engineering, as well as Software Engineering, for AI-based systems (e.g., [46] [122] [114] [123]).

Nguyen-Duc et al. [122] conducted a case study, in seven organisations developing AI software, in order to understand how Software Engineering processes and practices can be applied to the development of AI-based systems. The authors argue that the understanding of these processes and practices is limited, and in many cases the development of AI-based systems in these organisations is exploratory and experimental. Regarding ethical requirements, seen as non-functional requirements, the authors observed that there is no specific non-functional requirement addressing the ethical issue of explainability of AI models in the investigated cases, however, some cases give attention to ethical requirements of privacy and data security.

Belani et al. [46] investigated the challenges in developing AI-based systems and presented a taxonomy for RE4AI. Challenges in the development of AI-based systems present themselves throughout the Software Development Cycle, not only in its early stages. In relation to the ethical principle of privacy, the authors argue that there is a violation of requirements traceability when black-boxes (subsystems) are introduced in Software Engineering processes, this occurs because it is not possible to find the source of the requirement, nor to indicate if there was a success of the elicited requirements, seen as fundamental to: analyze impacts when requirements changes occur, trace which user requirement is related to which system requirement, besides the stakeholder that motivated it, and perform the verification phase within Requirements Engineering.

The last two studies presented converge with the discussions presented in the work of Raji et al. [79], where, in the context of AI-based systems there is a difficulty in tracing the output of a model back to system requirements, as they may not be explicitly documented, and possible issues arise only when the system is deployed [79]. Furthermore, “there is a lack of a formalization of a standard development or practice model, or process guidelines for when and in what context it is appropriate to implement certain recommendations" [79].

Cysneiros and Leite [124] propose the use of Softgoal Interdependencies Goals (SIG) catalogues in order to obtain an improved set of non-functional requirements. The authors state that "software engineers must begin to investigate how to operationalise the requirements of ethics, security, safety, and privacy", with SIG catalogues being a way to address this issue by reusing knowledge bases of non-functional requirements that have a social responsibility purpose in the AI context.

2.4 Related works

Morley et al. [13] presented a study that proposes to bridge gaps between ethical principles and practice by creating a typology of applied ethical AI, identifying several tools for implementing ethics in AI. This study is one of the most comprehensive on tools for implementing ethics in AI, where the tools found are framed in the established typology, besides addressing the same principles proposed in the unified framework, proposed by Floridi and Cowls [62] (Floridi is one of the co-authors of the paper in question), that is, the 5 principles (i.e., beneficence, non-maleficence, autonomy, justice, explicability) presented in Section 2.2.2. The tools found are not homogeneously distributed along the typology, this goes in line with what is discussed in Section 2.2.3, where most of the tools focus on the ethical principles of transparency, justice and fairness, non-maleficence, responsibility and privacy. The authors concluded that most of the tools found are in their early stages of development, and have difficulties in knowing where and how to apply them.

Krafft et al. [39] presented a practical framework for making the transition from "what" to "how", i.e., from principles to practice. The tool presented follows the VCIO (values, criteria, observables, indicators) approach to perform the ethical assessment, synthesizing this assessment into an Ethical AI Label. This framework, which has as main strength a simplified visualization of the evaluation of an AI on ethical principles, is also inspired by the energy efficiency label, just like an household appliance, an AI-based system would be conferred an Ethical AI Label. Despite being called a framework, the authors present a checklist or an impact assessment list, as discussed in Section 2.2.3. Thus, this tool would have its greatest practical utility in the final stages of the Software Development Life Cycle, after the system has been deployed. It is also noted that the authors leave out the developers in the list of stakeholders that are benefited by the proposed framework.

Schiff et al. [66] discussed the gap between principles and practice in AI ethics through 5 aspects: 1) the complex impact of AI, 2) how to distribute the responsibility of AI impact on human well-being, 3) the plurality of professions in AI ethics (e.g., engineers, computer scientists, policy makers, sociologists, ethicists), 4) the abundance of tools, 5) the division of labour into technical and non-technical teams. The authors suggest that interdisciplinary teams in both higher education institutions and organizations, with humanities and social science students partnering with engineering and computer science students, and developers partnering with non-technical teams (e.g., social scientists, lawyers, ethicists), respectively, should aim to learn each other's languages and work together. Then, they listed 6 criteria for a framework for responsible AI development (broad, operationalizable, flexible, iterative, guided and participatory) in order to propose a framework, based on the IEEE 7010 standard. Although the authors' proposal is designated framework, they end up introducing an impact assessment list, reflecting in our criticism exposed in Section 2.2.3 and also carried out in the work of Krafft et al. [39], that is, this tool would also have its greatest practical utility in the final phases of the Software Development Life Cycle. Moreover, the authors do not present a very detailed definition of the framework nor how to follow it, and the use case presented includes few technical details on how to actually implement the framework, presenting broad and generic recommendations.

Guizzardi et al. [23] presented in the paper Ethical Requirements for AI Systems a discussion on ethical requirements, pointing out that well-established techniques used in Requirements Engineering can also be used to develop AI-based systems in compliance with ethical principles and guidelines. The authors explained that the system built based on their proposal is not an ethical agent – they rationalise and make ethical decisions – but rather an AI system with qualities established in ethical requirements and other requirements that they must comply with, going in line with the proposal of our study. Furthermore, they reported a case study in the area of autonomous vehicles, listing some functional and non-functional requirements by applying Floridi and Cowls' 5 principles [62]. Although they argue that traditional Requirements Engineering techniques can be used for such task, they do not detail which technique is used, neither define its steps, and the context of autonomous cars is the only AI-based system addressed.

Vakkuri et al. [6] presented a method to implement ethics in AI-based systems, called ECCOLA. This method consists of a set of 21 cards, divided into 8 themes, with questions to be answered by the PO and developers. The ethical principles laid out in the AI HLEG and IEEE EADv1 guidelines (available in Appendix B) served as a basis for the cards classification into themes and in designing the questions shown in the cards. The authors argue that the use of cards is not new in Software Engineering, and there exist methods for performing Requirements Engineering in ASD, such as Planning Poker [125], and the user stories present in the product backlog items in Kanban boards. This set of cards is modular, the relevant cards for the iteration (sprint) in question can be used, i.e., it is suitable for agile development. Although the aim of this study is to implement ethics in AI, it is stated that the ECCOLA method only helps to increase the ethical awareness of the development team, providing no means of measuring the impact of the use of the tool, nor do they include an example of the use of the method in practice.

The mentioned works relate in the area of AI ethics, both in principles and practice. Although the thematic area of presented studies is similar to the area of our study, the objective of this study and those presented in this section are different. Our aim is to investigate how the implementation of ethical principles in AI-based systems can be performed during the software development process and to propose a guide to support the implementation of these principles. To achieve this goal, we will follow the Design Science Research steps presented in Section 1.5, investigating in the literature the existing tools, proposing a new one, focusing on developers in the first phase of the Software Development Life Cycle, including extensive documentation and informative resource to the users of the tool, besides the creation of evaluation criteria and the realization of the evaluation of its impact in an AI-based system practitioners.

2.5 Chapter Summary

This chapter presented the main concepts necessary for the understanding of this work. In addition, the works related to the theme of this research were presented. First, we briefly presented what is Artificial Intelligence and its best-known techniques, then we presented our strategy to explore ethics in AI (Figure 2.7). We presented a variety of published ethical guidelines and their discussions, in addition to the concepts of binding (hard law) and non-binding laws (soft law), pointing out that AI ethics guidelines, such as the 39 listed in Appendix A, are non-binding. Next, we presented the discussion regarding the ethical principles present in these guidelines, listing the principles: 1) Transparency; 2) Justice and fairness; 3) Non-maleficence; 4) Responsibility; 5) Privacy; 6) Beneficence; 7) Freedom and autonomy; 8) Trust; 9) Sustainability; 10) Dignity; 11) Solidarity, and their associated ethical issues. We then addressed the context of practice in AI ethics, its challenges, proposed means to address and implement the principles, such as checklists, impact assessment lists, tools and frameworks. We follow, bringing this debate to the Software Engineering field, showing techniques for requirements elicitation, the ethical requirements and the RE4AI, in order to address this issue in the first phase of the Software Development Life Cycle, the requirements analysis. Finally, we presented the works related to our study, pointing out the main differences existing in relation to our work. In Chapter 3, we will present the Systematic Literature Review performed, result of the first stage of the Design Science Research presented in Section 1.5.

Chapter 3

Systematic Literature Review

In this Chapter we present the Systematic Literature Review (SLR), as well as the results found in its execution, as proposed in Section 1.5, which comprises the first stage of Design Science Research – 1. Awareness of Problem.

3.1 Systematic Literature Review

We conducted a Systematic Literature Review (SLR) to identify works that investigate ethics in the context of Artificial Intelligence systems and that propose or define techniques, tools, methods, frameworks and processes that support the use of Ethics in the requirements elicitation stage in software development. An SLR aims to identify, analyse and interpret the available evidence related to a particular research topic or phenomenon of interest [5]. The SLR was conducted using the three phases proposed by Kitchenham [5]: Planning, Conducting and Publication of results. The phases and the respective activities conducted in this research are presented in Figure 3.1.

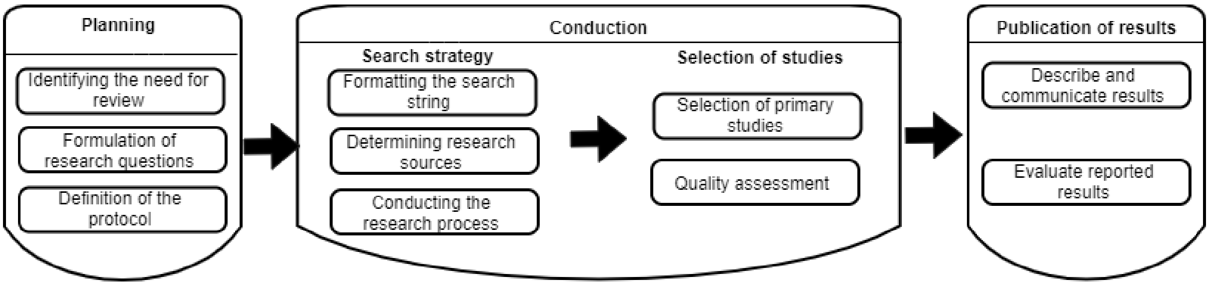

Planning
Identifying the need for review
Formulation of research questions
Definition of the protocol
Conduction
Search strategy
Formatting the search string
Determining research sources
Conducting the research process
Selection of studies
Selection of primary studies
Quality assessment
Publication of results
Describe and communicate results
Evaluate reported results

Figure 3.1: Phases and activities carried out in the SLR [5]

- **Planning**: aims to identify the real need to conduct an SLR, that is, the motivation for the execution of a research [126]. This phase is composed of the main activities of defining the objective and verifying the need for revision, formulating the research questions and preparing the protocol that will guide the RSL, aiming to minimize biases that may be committed by the researcher.

- **Conduction**: this phase of the SLR was subdivided into two stages, the first relates to the definition of the search string to be applied in the automatic search sources. The string needs to be tested and adapted before conducting the search to identify works of interest to the research. The second stage consists of conducting the SLR, reading the titles of the papers found, reading the abstracts, and applying the inclusion and exclusion criteria of the papers, after reading the introduction, conclusion and methodology of the pre-selected articles. From this set of steps is performed the selection, collection and synthesization of data aiming to answer the research questions and thus facilitate the analysis and synthesis for the creation of results [126].

- **Publication of Results**: the last phase of the SLR is related to the documentation and description of the results, preparation of answers to the research questions [127] and evaluation of the results found.

The research protocol of this work was developed to meet the objective of identifying techniques, methodologies, methods, frameworks, processes and tools to support Ethics in the requirements elicitation stage of software development, as well as to identify the works that investigate ethics in requirements elicitation for applications in the context of Artificial Intelligence and Machine Learning, as the latter is contained in the former.

3.1.1 Research Questions

To meet the objectives set out in Section 1.3.1, we set the following Research Questions (RQ) to guide the execution of this research.

RQ.1: What techniques, methodologies, methods, frameworks, processes and tools exist in the literature to support the operationalisation of ethical requirements in AI?

RQ.2: How can we enable the implementation of AI ethics during the software development process?

RQ.3: What ethical principles and guidelines exist in literature and industry in the context of Artificial Intelligence?

3.1.2 Search String

The automatic search string used in this research was adapted according to the possibility of using the digital libraries' connectors: Scopus, DBLP-Computer Science Bibliography, ACM Digital Library, Google Scholar and European Parliament Think Tank. The basic string structure used was: (“ethic” OR “ethics” OR “ethical” OR “ethically” OR “applied ethics” OR “ethical values” OR “responsible ai” OR “ai ethics”) AND (“design” OR “development” OR “governance” OR “method” OR “framework” OR “tool” OR “process” OR “implementing” OR “implementation” OR “practices” OR “guidelines” OR “principles”) AND (“artificial intelligence” OR “machine learning” OR “AI” OR “ML”).

3.1.3 Selection Criteria (Inclusion and Exclusion)

The following selection criteria were defined to identify the primary studies to be included in the research:

1. The study must be available in the previously defined digital databases;
2. The publication year of the studies must be between 2018 and 2021;
3. The study should relate to ethical guidelines or principles in the context of Artificial Intelligence;
4. The study should address the practical issues of ethics in Artificial Intelligence or Machine Learning;
5. The study should address the issue of Requirements Engineering in Artificial Intelligence.

The exclusion criteria for the studies were the non-fulfilment of any of the inclusion criteria, as well as:

1. Do not address ethics in AI;
2. Address the issue of artificial moral agents or ethical decision making.

3.1.4 Conduction of the SLR

In order to obtain diverse studies that can encompass the theoretical and practical aspects of AI ethics to ensure a broad scope for the review, the following databases were used: Scopus; DBLP; Google Scholar; ACM Digital Library; European Parliament Think Tank.

The Scopus, DBLP and ACM databases were used because they represent solid and international databases with publications related to computer science. Google Scholar was selected to complement the articles found by random selection, while the European Parliament Think Tank base was selected due to the European Union being an influential

AI ethics research hub, with recent contributions in AI ethics regulation and legislation, such as the AI HLEG [42] and the GDPR [59]. As a way to validate the string, we used the Scopus database, and the final string:

> (TITLE-ABS-KEY ((“ethic” OR “ethics” OR “ethical” OR “ethically” OR “applied ethics” OR “ethical values” OR “responsible ai” OR “ai ethics”)) AND TITLE-ABS-KEY ((“design” OR “development” OR “governance” OR “method” OR “framework” OR “tool” OR “process” OR “implementing” OR “implementation” OR “practices” OR “guidelines” OR “principles”)) AND TITLE-ABS-KEY((“artificial intelligence” OR “machine learning” OR “AI” OR “ML”))) AND (LIMIT-TO(SUBJAREA, “COMP”)) AND (LIMIT-TO(PUBYEAR, 2021) OR LIMIT-TO(PUBYEAR, 2020) OR LIMIT-TO (PUBYEAR, 2019) OR LIMIT-TO(PUBYEAR, 2018))

The subject area was limited to Computer Science only, as it resulted in a very high number of articles to be analysed (over 2000), besides returning articles from diverse areas (e.g., medicine, business, social sciences, humanities, communication), so the area was filtered in order to find articles more relevant to the scope of the study, which resulted in 691 articles.

With the purpose of filtering recent articles on the topic, the years 2018, 2019, 2020 and 2021 were used as a time delimiter. This reduced time interval is due to the level of maturity of the international debate on AI ethics. Prior to the year 2018, the debate still remained strictly in the theoretical field and proved to be outdated, as most of the most significant (88%) AI ethics guidelines documents were published after 2016 [27], e.g. Ethically Aligned Design (IEEE EADv1) and Ethics Guidelines for Trustworthy AI by the High-Level Expert Group on Artificial Intelligence (AI HLEG) in March and April 2019, respectively. In addition, the European Union’s General Data Protection Regulation (GDPR) came into force in May 2018, and Brazil’s General Law on Personal Data Protection (LGPD) came into force on 18 September 2020, both of which are binding laws, i.e. hard laws. We also note, that the AI ethics debate also mostly addresses the issue of ethical or moral artificial agents (i.e., that perform ethical decision making), which do not belong to the scope of our study.

In order to highlight the increasing attention given by academia in recent years, we highlight in Figure 3.2 milestones associated with the volume of publications per year on this topic. These milestones are the binding laws – hard laws – GDPR [59] and LGPD [60], and the non-binding laws (guidelines) – soft laws – IEEE EADv1 [58] and AI HLEG [42].

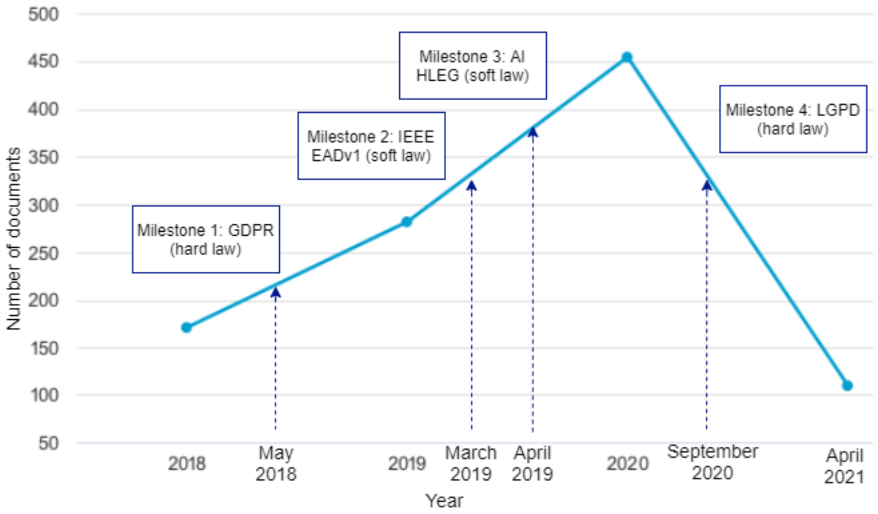


Figure 3.2: Milestones associated with the trends presented in the evolution of the volume of publications in the literature in the area

Figure 3.3 presents the phases of article selection in the databases, the quantities found and rejected in each phase, and the final selection after applying the selection and exclusion criteria. After the search in the selected digital databases, using the search string, a set of 1018 initial articles was returned to be analyzed. In the first phase of exclusion of the studies, the titles, abstracts and key words of the articles were read, where 734 articles that were not in the scope of the work were discarded. After this initial exclusion phase, a second phase was employed, where we applied the selection and exclusion criteria, and 158 papers were rejected. Next, a complete reading of the filtered articles was performed. We also discarded articles that did not bring additional information – redundant in relation to others considered more complete and current. We noticed that a large number of articles addressing the issue of ethics in AI from the perspective of artificial moral agents, i.e., ethical decision making, were rejected. In the end, 33 primary studies were selected.

3.2 SLR Results

In the RSL, 33 primary studies were selected, as presented in Table 3.1. The "ID" column represents the identifier of the study. The "Title/Reference" column represents the title and the bibliographic reference of the paper. The T/M/Me/F/P/Tx column represents whether the paper investigated/used/proposed any technique (T), methodology (M), method (Me), framework (F), process (P) or taxonomy (Tx). The Tools column indicates whether the work used or developed a tool (Y) or not (N). The NE column represents the context in which the work was applied, which can be: in the Academic context (A), in Industry (I) or if the authors did only an illustration of the proposal (IL).

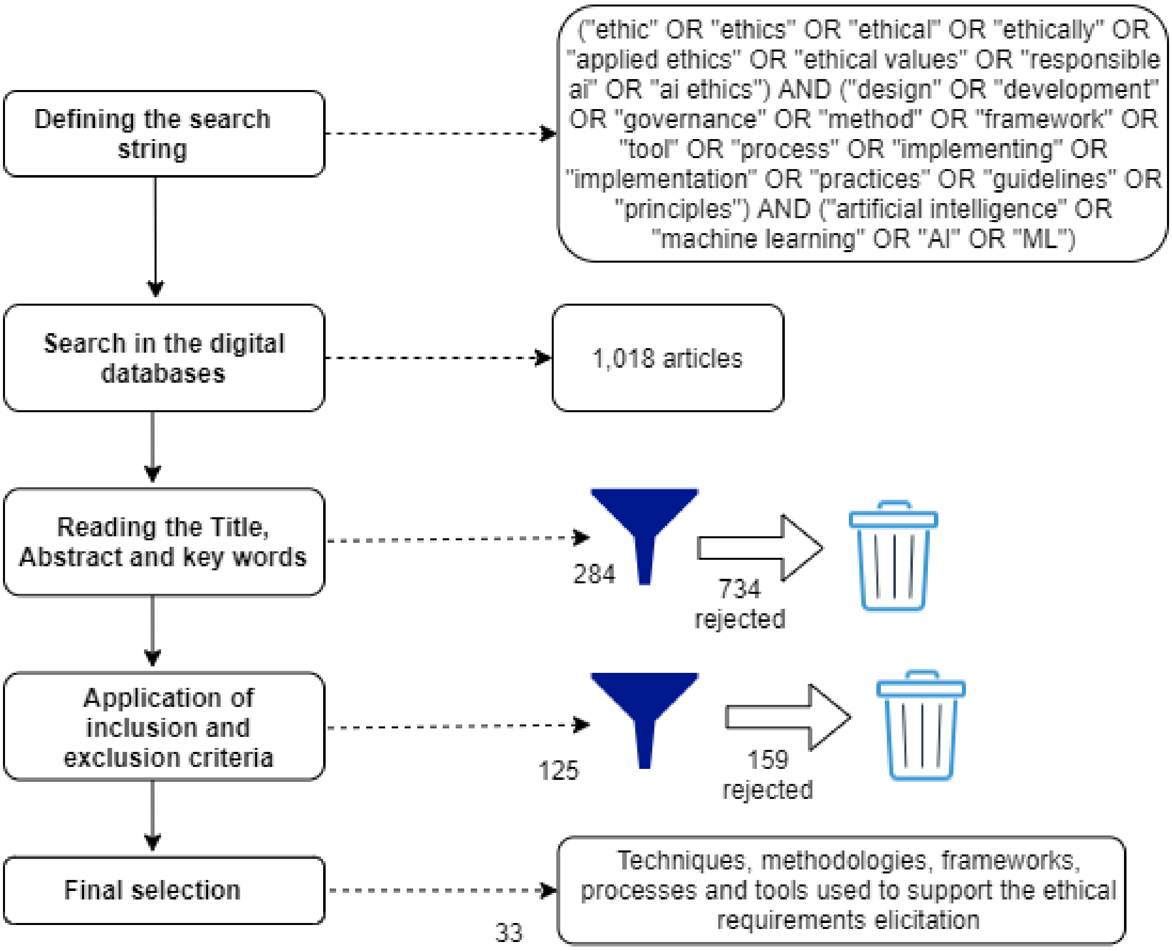

Defining the search string
("ethic" OR "ethics" OR "ethical" OR "ethically" OR "applied ethics" OR "ethical values" OR "responsible ai" OR "ai ethics") AND ("design" OR "development" OR "governance" OR "method" OR "framework" OR "tool" OR "process" OR "implementing" OR "implementation" OR "practices" OR "guidelines" OR "principles") AND ("artificial intelligence" OR "machine learning" OR "AI" OR "ML")
Search in the digital databases
1,018 articles
Reading the Title, Abstract and key words
284
734 rejected
Application of inclusion and exclusion criteria
125
159 rejected
Final selection
33
Techniques, methodologies, frameworks, processes and tools used to support the ethical requirements elicitation

Figure 3.3: Phases of works selection in the databases

The Phases (SDLC) or Guidelines column represents the phases of the Software Development Life Cycle: 1. (R) Requirement Analysis; 2. (D) Designing; 3. (C) Coding; 4. (T) Testing; and 5. (I or M) Implementing & Maintenance, or whether the article addresses mainly ethical guidelines. Finally, the RQ column represents the research questions that the primary study answered.

Table 3.1: Primary studies selected from the SLR

ID	Title/Reference	T/M/Me/F/P/Tx	Tools (Y/N)	NE (A,I,IL)	Phases SDLC or Guidelines	RQ

S1	A Roadmap for Ethics-Aware Software Engineering [26]	Proposes a Framework for Ethics-aware Software Engineering	N	A	Todas	1,2
S2	AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations [29]	Proposes the AI4People Framework	N	A	Guidelines	1, 3
S3	An Ethical Framework for the Design Development Implementation and Assessment of Drones Used in Public Healthcare [88]	Proposes a Framework for Designing Drones in healthcare	N	IL	Designing, Implementing & Maintenance	2
S4	Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI [20]	Proposes a Framework for building Ethical AI	N	IL	Guidelines	2
S5	Does ACM's Code of Ethics Change Ethical Decision Making in Software Development? [25]	Not Applicable	N	A and I	Guidelines	2
S6	ECCOLA - a Method for Implementing Ethically Aligned AI Systems [6]	Proposes the ECCOLA Method	N	IL	Requirement Analysis	1, 2
S7	Enhanced well-being assessment as basis for the practical implementation of ethical and rights-based normative principles for AI [75]	Proposes the Enhanced Well-Being Impact Assessment (EWIA) Framework	N	IL	Designing, Implementing or Maintenance	1
S8	Ethical Framework for Designing Autonomous Intelligent Systems [90]	Proposes a Framework to Analyze and Discuss ethical issues	Y - User Stories	IL	Designing	1, 3
S9	Ethical Requirements for AI Systems [23]	Proposes the use of traditional RE techniques to elicit and analyze ethical requirements	N	IL	Requirement Analysis	1

S10	Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI [67]	Proposes a Taxonomy for XAI	N	A	Designing, Implementing & Maintenance	2
S11	From What to How An Initial Review of Publicly Available AI Ethics Tools Methods and Research To Translate Principles into Practicess [13]	Presents a catalog with several Tools and Methods	Y	IL	All	1
S12	Global AI Ethics: A Review of the Social Impacts and Ethical Implications of Artificial Intelligence [15]	Not Applicable	N	Not Applicable	Not Applicable	2
S13	On the Interplay between Requirements, Engineering, and Artificial Intelligence [95]	Not Applicable	N	Not Applicable	Requirement Analysis	2
S14	Progressing Towards Responsible AI [128]	Not Applicable	N	Not Applicable	Not Applicable	2
S15	Relevance of Ethical Guidelines for AI - A Survey and Evaluation [28]	Not Applicable	N	Not Applicable	Guidelines	3
S16	Requirements Engineering Challenges in Building AI-Based Complex Systems [46]	Proposes a Taxonomy for RE4AI	N	A	Requirement Analysis	2
S17	Requirements Engineering for Machine Learning: Perspectives from Data Scientists [12]	Proposes a Process for RE of ML systems	N	A, I	Requirement Analysis	2
S18	Responsible AI—Two Frameworks for Ethical Design Practice [43]	Proposes the Frameworks The Responsible Design Process and The Spheres of Technology Experience	N	IL	Designing, Coding, Implementing & Maintenance	1
S19	Review of AI principles in practice [45]	Not Applicable	N	Not Applicable	Guidelines	3
S20	The Current State of Industrial Practice in Artificial Intelligence Ethics [14]	Not Applicable	N	I	Not Applicable	2
S21	The Ethics of AI Ethics An Evaluation of Guidelines [11]	Not Applicable	N	Not Applicable	Guidelines	3
S22	The Global Landscape of AI Ethics Guidelines [27]	Not Applicable	N	Not Applicable	Guidelines	3

S23	This is Just a Prototype How Ethics Are Ignored in Software Startup-Like Environment [32]	Not Applicable	N	I	Not Applicable	2
S24	Trustworthy AI as a Future Driver for Competitiveness and Social Change in the EU [129]	Not Applicable	N	Not Applicable	Guidelines	2, 3
S25	Principles to Practices for Responsible AI: Closing the Gap [66]	Proposes the Well-being Impact Assessment Framework	N	I, IL	Designing, Implementing & Maintenance	1
S26	From Principles to Practice: An interdisciplinary framework to operationalise AI ethics [39]	Proposes the AI Ethics Label Framework	N	IL	Designing, Implementing & Maintenance	1
S27	The impact of the GDPR on artificial intelligence [130]	Not Applicable	N	Not Applicable	Guidelines	3
S28	Understanding Artificial Intelligence ethics and safety - The Alan Turing Institute [94]	Proposes the Process-Based Governance Framework (PBG)	N	IL	All	1
S29	Responsible AI by Design in Practice [36]	Proposes a company-wide Methodology	Y - Luca Ethics, Spectra, Luca Comms	I	Designing, Implementing & Maintenance	1, 3
S30	Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications [10]	Not Applicable	N	N	Guidelines	3
S31	Time for AI (Ethics) Maturity Model Is Now [131]	Proposes the development of an AI ethics Maturity Model	N	IL	Not Applicable	1
S32	Ethics as a service: a pragmatic operationalisation of AI Ethics [132]	Proposes the concept of Ethics as a Service	N	IL	Not Applicable	2

S33	How to Write Ethical User Stories? Impacts of the ECCOLA Method [133]	Evaluates the ECCOLA Method	N	A	Requirement Analysis	1, 2

3.2.1 RQ.1. What techniques, methodologies, methods, frameworks, processes and tools exist in the literature to support the operationalisation of ethical requirements in AI?

Recent research indicates a growing interest in studies that assist in identifying mechanisms that can support the elicitation of ethical requirements in the context of AI-based systems. Aydemir and Dalpiaz [26] proposed an analytical framework that assists stakeholders of the software engineering process in the analysis of ethical issues in relation to the Software Development Life Cycle, highlighting the importance of preserving some ethical aspects, such as diversity, privacy, autonomy, eco-sustainability and discrimination. The proposed framework method is agnostic in relation to the different development methodologies, being organised in five phases (Articulation, Specification, Implementation, Verification and Validation) arranged in cycles, trying to reinforce the fact that ethics is not a single activity, but requires continuous effort while the software artifact is being built, maintained and used.

The main opportunities and risks of projects that link AI to society are presented through the AI4People framework [29], which presents a synthesis of five ethical principles that should underpin the development of this type of project, highlighting the need to adopt 20 concrete recommendations organized into actions that involve activities to assess, develop, encourage and support good AI practices. These recommendations can be classified into actions that can be carried out directly by national or supranational policy makers and actions that can be disseminated and used by stakeholders in the development cycle.

The framework proposed by Leikas et al. [90] used as a baseline the relevant ethical principles for a software development project and a framework with an iterative and multidisciplinary perspective, which can be used in different phases of the project to discuss and analyse ethical issues in AI. The authors argued that at the beginning of the project, project goals are defined and interpreted as project requirements, and iteratively, when reaching a more detailed project level, the framework can be applied again. Finally, the final project can be evaluated with the proposed framework. For this, the authors propose the use of scenarios as a tool to capture the specific qualitative user or key stakeholder information that is needed for a systematic analysis of the ethical issues in the specific project case.

The framework proposed by Schiff et al. [66] assists in the implementation of ethics in AI-based systems, presenting the following characteristics: broad, operationalisable, flexible, iterative, guided, and participatory. Building on these characteristics and on the IEEE 7010 standard, the authors developed a wellbeing impact assessment to ensure that organisations can understand and address the various impacts that developed AI-based systems may have on human well-being. In an iterative manner, not just during the design phase, the strategy embarks on internal analysis, in conjunction with stakeholder and user engagement, both to accomplish the task of determining the context of use and where the impact on human well-being of using the AI-based system will be. Finally, the authors signalled that technical and non-technical based teams should use a common language to work together, besides suggesting the adoption of new educational practices both in higher education institutions and in organisations.

The framework presented by Krafft et al. [39] proposed the implementation of ethics in AI based on the VCIO (values, criteria, indicators, observables) model. The authors state that the framework can bring principles that can assist in organizational practice, supporting the dissemination of values, being simple to understand.

AI-based systems are software, and only a portion comprises AI code such as Machine Learning [131]. While ethical requirements such as transparency, explainability, fairness, have unique meanings in the context of AI, they have not been sufficiently addressed in existing software models [131]. Vakkuri et al. [131] proposed a maturity model designed to address technical and ethical requirements in AI, in a similar way to existing maturity models in Software Engineering, such as Capability Maturity Model Integration (CMMI). However, the authors did not detail how the model should be developed.

Peters et al. [43] presented two frameworks to support the integration of ethical analysis into engineering practice to mitigate the challenges of bringing principles into practice in AI ethics. The first framework addressed the responsible development process, involving the phases of research, insights, ideation, prototyping and in-use evaluation, providing an overview in which it is possible to research, develop and situate new methods and tools that support each of the phases. The second framework addressed the impact on the experience of AI-based systems, presenting in six spheres the experience of technology use: adoption, interface, task, behavior, life and society. Although focused on autonomy and well-being, the spheres presented can be used to assess the impact in relation to other ethical principles, and at any stage of the responsible development process.

Leslie [94] has defined through the PBG Framework a guide to outline values, principles and guidelines to assist UK public sector departments in ensuring the development and deployment of ethical, safe and accountable AI. The PBG Framework provides an overview of the governance procedures and protocols that organize the control of project workflow structures, producing an outline of the relevant team members in each governance action, the relevant workflow stages for the governance goals, the timescales for actions, re-evaluations and ongoing monitoring, the defined logging protocols to ensure auditability. The PBG is composed of the phases of problem formulation; data extraction and acquisition; data preprocessing; modeling, testing and validation; deploy, monitor and reassess.

Havrda and Rakova [75] propose the practical application of a framework for well-being impact assessment of the use of Autonomous and Intelligent Systems. This process can enable a human-centred algorithm-based approach to understanding the impacts of AI use on systems. The infrastructure provided by this work, aims to enable AI-based systems runtime stakeholders a form of cooperation with a focus on implementing enhanced well-being impact assessments through the use of the Well-Being Impact Assessment (EWIA). EWIA can be executed through the establishment of joint monitoring and testing systems, allowing the collective implementation of AI principles in practice and thus enriching the tool options of practitioners to guarantee positive results in the development of Autonomous and Intelligent Systems.

Guizzardi et al. [23] investigated on how classical techniques developed in Requirements Engineering can be used to develop AI-based systems that are in accordance with ethical principles, explaining how ethical requirements can be seen as ecological requirements, being derived from value and risk assessments, positive and negative contributions, respectively. A technique was presented that works through value and risk assessments by stakeholders. The authors indicated that these activities should be integrated into the elicitation of functional and non-functional requirements (ethical requirements).

Benjamins et al. [36] defined a methodology to apply AI ethics in organizations with the use of some tools. To implement the methodology, the organization should first start an awareness campaign on AI ethics by introducing the adopted ethical principles, the methodology, the training program and the tools. Next, a training program (i.e., online course) should be initiated with the people involved with the design and development of products and services that use AI, in addition to the people in charge with the acquisition of third-party technologies. After that, with the use of an agile governance model, the responsibilities should be delegated to the people related to the products and services. The authors state that by using this methodology it is possible to cover the ethical principles of fairness, explainability, transparency and data privacy.

The ECCOLA method [6] makes considerations that allow organisations to analyse various ethical issues present in AI systems, making high level AI ethics principles more practical, enabling developers to implement them in practice more easily. In practice, ECCOLA takes the form of a deck of cards with 21 cards covering 7 ethical principles in addition to a Stakeholder Analysis card, i.e., 8 themes, showing from 1 to 6 cards for each theme. Each card in ECCOLA is divided into three parts: (1) motivation (i.e. why this is important), (2) what to do (to solve this problem) and (3) a practical example of the topic (to make the problems more tangible). ECCOLA is based on the AI HLEG and IEEE EADv1 guidelines and its purpose is to help developers and Product Owners, especially in agile development projects, to implement ethics as part of user stories.

Halme et al. [133] made an evaluation of the ECCOLA method with 15 master's level student projects, where 9 teams utilized ECCOLA and 6 did not. 298 user stories were produced by using only 4 cards presented in the method regarding system security and privacy & data, where 179 were user stories produced using the ECCOLA method. The authors discuss that ECCOLA method seems to result in more human-centric user stories, assists in writing non-functional user stories, helps teams producing higher quality user stories and in producing user stories with a wider perspective than just of its functionalities. However, the use of its cards (the themes presented) did not affect how teams wrote user stories. Authors argue that this study provided an introductory view on writing ethical user stories, nevertheless, lack to provide such user stories.

Morley et al. [13] presented a list of tools and methodologies that aid the implementation of AI ethics, considered one of the most comprehensive available in the literature. To this end, they created a typology of applied ethical AI, where developers can search for appropriate tools and methodologies given a context. The authors argued that tools and methods are not equally distributed throughout the typology, and that the existence of such tools is necessary but not sufficient. Finally, they point out that most of the tools found are in their early stages, and there is a lack of usability of the tools and methods found, that is, there is a lack of documentation leading to a need for further work before they are put into production.

Our findings reveal that the studies found in the literature that investigate, use, or propose some practical means of implementing ethics in AI are in their early stages of development, addressing in different ways the challenge of practicing ethics in AI. These studies do not yet present in a clear way how to apply it, thus implying additional work for developers to put it into practice during the Software Development Life Cycle. In addition, there is no substantial evidence nor have they been sufficiently tested to make it possible to state that they can effectively operationalise ethics in AI. However, despite their limitations, they have demonstrated their usefulness in increasing ethical awareness among developers of AI-based systems.

3.2.2 RQ.2. How can we enable the implementation of AI ethics during the software development process?

In the context of Artificial Intelligence, it is noticeable that recent initiatives have been established to deal with ethical issues of the developed products, as occurs in the industry of autonomous cars and war robots [26]. Vakkuri et al. [14] stated that the implementation of ethics in AI is still in its early stages, evidencing that even if there are several guidelines to support this area it is necessary to make them more practical for developers, project teams, Product Owners and other stakeholders. Moreover, it is important to emphasize that this theme cannot be outsourced in the projects, but should be treated in the totality of those involved and that the implementation should occur in a systemic way [14].

Antonov and Kerikmäe [129] highlighted the importance of finding a common language between developers, designers, relevant stakeholders and legislators of ethical and legal guidelines for the application of ethics in software projects to occur fully. An example of the European Union's (EU) efforts to address AI ethics in an attempt to regulate the use of AI-based systems has three parts: 1) leveraging the EU's industrial and technological capacity and assimilation of AI across the economy; 2) preparing for socio-economic changes and; 3) ensuring an appropriate ethical and legal framework [129].

Explainability in Machine Learning, also referred to as eXplainable AI (XAI), enables humans to understand, trust and manage prediction models, and produce more explainable models while maintaining high prediction accuracy, where the authors rephrased this definition to: "Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its operation clear or easy to understand" [67]. Arrieta et al. [67] presented a literature review on explainability in Machine Learning, in addition to a detailed taxonomy related to XAI, starting from the literature review, identifying trends for explainability techniques related to different Machine Learning techniques. Although explainability is strongly related to post-hoc explainability (after the AI-based system has been deployed), the authors consider explainability as a design goal, because this is considered a broad concept. Post-hoc explainability is implemented through techniques that convert non-interpretable models into explainable models. In the proposed taxonomy it is possible to observe that the branch of post-hoc explicability in XAI is, in fact, considerably larger than the other branch, that of transparent models. Finally, after presenting a number of challenges in XAI (e.g., interpretability versus performance, explainability in Deep Learning, concepts and metrics in XAI), they turn to responsible AI, presenting ethical principles interrelated with explainability, such as fairness and accountability.

The implementation of ethics in software projects, according to Scantamburlo et al. [128], involves multidisciplinary work, and it is indicated to engage experts from different areas (e.g., software engineers, lawyers, data protection analysts, philosophers, journalists, sociologists, academic researchers, entrepreneurs, marketing analysts). With this variety of profiles it is expected to generate a wide range of opinions and experiences around ethical topics. The existence of such a working group can bridge some gaps between the ethical debate and engineering practice in organisations, involving: support for researchers and practitioners to navigate the ethical challenges that arise in different real-world AI applications; aiding interdisciplinary dialogue involving people from different backgrounds; promoting cross-fertilisation between different sectors, including academia, business and public institutions; and generating inspiration for future responsible practices to be applied in the AI field [128].

Siau and Wang [20] mentioned that when AI is developed or used, it is necessary to make the application of ethical principles in different phases of the project, from design, development and up to its application and use. The authors state that in order to be able to implement AI ethics in software development projects, it should be reflected on at least three groups of factors:

1. AI characteristics that may originate ethical problems: Transparency, Data security and privacy, Autonomy, Intentionality and responsibility.

2. Human factors that may cause ethical risks: Accountability, Ethical standards, and Human rights laws.

3. Social impacts stemming from AI applications: Automation and job replacement, Accessibility, Democracy and civil rights.

The relevance of the wider and more appropriate use of ethics in AI comes from the fact that ethical decisions in software development can substantially impact end users, organisations and the environment [25]. The use of a practical tool can facilitate the consideration of ethics and human values in technology design in software projects, and is especially useful in the practical application of ethics by people who have limited experience in this area [88]. However, practical tools can have the opposite effect, as they can lead to systematic neglect of some principles, neither providing sufficient practical guidance nor making clear who is responsible for decisions made [132].

Importantly, implementing AI ethics in the early phases of the Software Development Life Cycle is cheaper than if it is started during later phases of the project, such as at deployment, and AI ethics can be treated as a non-functional requirement of an AI-based system [14]. Nevertheless, implementing AI ethics in practice is still a constant challenge. There are high-level guidelines guiding how to perform such a task, drafted by governments and private organisations, but lack practicality for developers [6] [132]. Furthermore, only adopting the process of explicitly instructing stakeholders in the use of codes of ethics, such as that of ACM, for decision making cannot be considered sufficient for the development of systems that make use of ethics in their routines and applications [25].

Furthermore, for AI ethics to be operationalised by developers and be effective about protecting individuals, society and the environment, this operationalisation must happen at the appropriate level of abstraction; it should not consist of a single checklist done only at the beginning of the development process [132].

Morley et al. [132] devised the term ethics as a service, where the authors compare a way to enable the implementation of AI ethics during the software development process with PaaS (Platform as a Service – Platform as a Service), a Cloud Computing term, where tools are neither too flexible nor too strict, and governance neither too centralized nor too decentralized. In this way, responsibility would be distributed between the actors, i.e. independent multi-disciplinary ethics advisory board, and AI-based systems developers. Although they argue that in this way the operationalisation of AI ethics would overcome many of its limitations, they do not yet know whether Ethics as a service works in practice.

Even with government and large technology companies endeavors, it is perceived that there is still a gap between the research and practice of applying ethics in AI, as many of the goals outlined by the ethical guidelines are not widely adopted in practice. One of the factors leading to the existence of this gap is the lack of formal methods and tools to implement AI ethics in a way that meets the needs of stakeholders [32]. Despite the lack of consolidated mechanisms, several established Software Engineering practices can be used to implement AI ethics, such as documentation, version control and project management practices. This type of use for practices allows system development to occur in a transparent way, enabling the tracking of actions and decision-making, in addition to the use of software quality practices to deal with problems in the context of AI ethics [32].

Belani et al. [46] addressed the challenges in Requirements Engineering for developing AI-based systems (RE4AI), proposing a taxonomy. The authors argue that the challenges in developing AI-based systems goes beyond the first phases of the Development Cycle, and list some Software Engineering challenges for Deep Learning, including: limited transparency, troubleshooting, testing, limited resources, monitoring, data privacy and protection, cultural differences. In addition, black-box elements hinder requirements traceability. After that, they perform a mapping of the challenges into a RE4AI taxonomy, including the Requirements Engineering activities: elicitation, analysis, specification, validation, management, and documentation. From the proposed taxonomy, some of the challenges presented for the elicitation and analysis and specification activities are: unclear ethical standards and unbalanced dataset. Finally, they signal that goal-oriented

RE (GORE) is an area that can be applied to address the listed challenges; however, they emphasize that further analysis of the GORE frameworks and methods is required to enable its applicability.

Social aspects of AI ethics and their impacts should also be analysed in software development projects, so that understanding culture provides a better understanding of ethics and vice versa. The unique vocabulary of society can negatively influence the understanding and expectations of the use of ethical AI-based systems. Some important pointers in this regard are: regional differences are significant as perceptions and understanding of AI are shaped by social contexts and cultural differences; AI may exacerbate social inequality, particularly in marginalized social groups; more actions should be explored in this context, such as conducting rigorous and independent ethnographic research, enabling the identification of the ethical and social implications of the use of AI-based systems in different cultures [15].

It is a role of requirements engineers to relate the application and results of the methods and tools that data scientists have available to balance, clean, validate and explain data to the context and needs of users [12]. Vogelsang and Borg [12] addressed the definition of Requirements Engineering characteristics and challenges for Machine Learning based systems. Although not explicitly dealing with ethics in AI, the authors consider legal requirements as a challenge for Requirements Engineering in ML. The authors state that there are few works on Requirements Engineering for Machine Learning systems. In addition, the authors presented some challenges and requirements for these systems from conducting interviews, including:

- Explainability: understanding how the program works is essential for the developer; furthermore, requirements engineers must explicitly elicit explainability requirements aimed at the users of these systems;

- Freedom from Discrimination: although Machine Learning-based systems are designed to discriminate (by identifying recurrent patterns in data), discrimination on the basis of race, gender, among others, is unacceptable to society or by law; furthermore, requirements engineers must elicit and identify protected characteristics not to be used by the algorithms;

- Legal and Regulatory Requirements: requirements engineers must comply with legal requirements and demonstrate that no illegal characteristics have influenced the final data set used to train the models;

- Data requirements: Requirements engineers must identify and specify requirements related to data collection, format and range, as well as searching for, and being critical of, additional data sources.

Furthermore, the Requirements Engineering processes for ML-based systems are summarised, according to the results of the interviews and the challenges encountered, in the elicitation, analysis, specification, verification and validation steps:

- Elicitation: participation of data scientists and legal experts, determination of restrictions with respect to laws (such as GDPR) and protected characteristics;
- Analysis: discussion of performance measures and expected results by which the systems will be evaluated. The requirements engineer must elicit, analyse and discuss the conditions for data preparation, definition of outliers and data obtained;
- Specification: focusing on data requirements and the quality of requirements, requirements engineers must specify the quantity and quality of data, and specify requirements with respect to explainability and protected characteristics;
- Verification and validation: The requirements engineer should detect biases in the data, re-train the models, detect anomalies, and analyse operational data.

Kostova et al. [95] although not exploring AI ethics explicitly, address the relationship between Requirements Engineering and Artificial Intelligence in both directions. While the academic community produces efforts to create AI-based tools to support the Requirements Engineering phase (e.g., automated tools to elicit requirements, prioritize requirements, refine requirements into specifications, interpret and classify requirements), referred to as AI4RE, few efforts currently exist to explore the changes to Requirements Engineering process practices that the introduction of AI and ML components cause, this one termed RE4AI. To a large extent, this is due to the fact that there is no clarity about the requirements pipeline and the Requirements Engineering process for integrating AI or ML components into software systems, with data management being identified as the main difference between traditional and data-driven software engineering. The authors argued that those responsible for the design of systems and algorithms are humans, and their role in the development of AI-based systems can lead to biases in prediction models, delineating people to certain behaviours, such as recommendations of what to watch, buy or read, influencing users' actions in a limited way. Finally, they state that "Requirement engineering is the only place to address this problem due to its interdisciplinary nature, with a strong technical emphasis".

We identified an inclination, among the reviewed papers, to use or adapt the Requirements Engineering process in the context of artificial intelligence and machine learning to address ethical issues, despite the existing challenges in the development process of AI-based systems (e.g., large amounts of data and the lack of a common language among the actors involved in this process). Table 3.1 can assist researchers in understanding this issue, by presenting which phase of the Software Development Life Cycle is addressed in the primary studies that were selected during this systematic literature review.

3.2.3 RQ.3. What ethical principles and guidelines exist in literature and industry in the context of Artificial Intelligence?

It is critical to design and develop AI-based systems based on ethical values and principles [90]. A high number of AI ethics guidelines have been published by different sectors, however, most of the guidelines were published after the year 2016, mostly coming from more economically developed countries such as the United States, United Kingdom, Japan, Germany, France and Finland [27]. The debate on AI ethics has remained primarily in the theoretical field, and, several of the studies found aim to explore the various AI ethics guidelines published, often presenting a synthesis of the principles found.

Smit et al. [45] presented a review of 30 documents that bring ethical principles in AI in order to find which ethical principles of AI design are recognised by governments and international organisations. From these documents, 316 principles were identified, which were mapped into 22 ethical principles: Human Augmentation, Do Good, Trustworthy, Human Centric, Autonomy, Equality (design), Equality (excursion), Traceability, Human dignity, Human Rights, Transparency (design), Democrability, Privacy, Security, Safety (design), Safety (excursion), Collaboration, Accountability, Understandability, Responsible use of data, Accuracy, Education and Promotion. Regarding the frequency of the principles, the five most prevalent, as well as their definitions, are:

- Do Good: AI-based systems should be designed and used to enhance financial, manufactured, intellectual, human, social or natural capital;
- Accountability: A person or organisation is responsible for the design and execution of an AI-based system;
- Equality: A designed AI system should treat all people equally;
- Privacy: An AI system should be designed and executed in such a way that it anonymises, or runs on top of anonymised data, and preserves users' power over the access and use of their data;
- Education: The design and implementation of AI should be guided by public engagement and democratic debate.

The study by Jobin et al. [27] – seen as one of the most comprehensive and significant on this topic by the literature – with the intention of mapping a global overview of existing ethical guidelines, analysed 84 guidelines, from non-binding laws only, compiling the results into 11 ethical principles. The following principles were compiled, and we also present the frequencies in the associated documents in parentheses: Transparency (73/84), Justice and fairness (68/84), Non-maleficence (60/84), Responsibility (60/84), Privacy (47/84), Beneficence (41/84), Freedom and autonomy (34/84), Trust (28/84), Sustainability (14/84), Dignity (13/84), Solidarity (6/84).

Leikas et al. [90] performed a synthesis of 6 publicly available AI ethics guidelines into 14 ethical values: 1) Integrity and human dignity; 2) Autonomy; 3) Human control; 4) Responsibility; 5) Justice, equality, fairness and solidarity; 6) Transparency; 7) Privacy; 8) Reliability; 9) Safety; 10) Security; 11) Accountability; 12) Explicability; 13) Sustainability; 14) Role of technology in society.

It is not uncommon for conflicts and trade-offs to exist between principles, such as privacy and transparency, where the decision in choosing one principle occurs at the detriment of another. To address this issue, "a practical set of guidelines that developers and users of AI can apply in practice needs to be aware of such conflicts and provide mechanisms for identifying them and dealing with them in an appropriate way" [10]. Furthermore, in order to enable developers and organisations to adopt the ethical principles and guidelines in practice, there is a need to map which tools relate to which ethical guidelines.

Ryan and Stahl [10] analysed a set of 91 AI ethical guidelines and condensed them into 11 ethical principles, aimed at developers and users of AI-based systems, providing a taxonomy of the main principles with a set of ethical issues that constitute each principle, as well as their descriptions. This work is strongly influenced by the study by Jobin et al. jobin2019global. The ethical principles listed are: 1) Transparency; 2) Justice and fairness; 3) Non-maleficence; 4) Responsibility; 5) Privacy; 6) Beneficence; 7) Freedom and autonomy; 8) Trust; 9) Sustainability; 10) Dignity; 11) Solidarity.

Considered one of the most high-profile guidelines, the EU's Ethics Guidelines for Trustworthy AI is developed on top of four principles: 1) Respect for human autonomy; 2) Prevention of harm; 3) Fairness; 4) Explicability [129].

Rothenberger et al. [28] conducted an exploratory study, where they condensed 5 AI ethics guidelines from industry, academia, governments and other institutions into just 6 principles: Transparency, Responsibility, Protection of Data Privacy, bias minimisation, AI purpose, and robustness. The principles were ranked according to responses obtained by interviews and questionnaires with experts and a wide audience, respectively. Of these, responsibility ranked first, where respondents presented doubts regarding who would be held accountable: the organization, the developer, or the user; and data privacy ranked second. As future work, the authors mentioned how a global AI ethics guideline can be developed fulfilling an ethical pluralism.

The AI4People [29] framework, presents a synthesis of several AI ethics guidelines in only five ethical principles that should underlie the development of this type of project. The first four derive from principles already existing in bioethics, and the fifth is an innovation brought by the authors to the field of Artificial Intelligence. The five principles are:

- Beneficence: promoting well-being, preserving the dignity and sustainability of the planet;
- Non-maleficence: providing privacy, safety and caring with capability;
- Autonomy: enabling decision-making;
- Justice: promoting prosperity and preserving solidarity;
- Explicability: enabling the other principles through intelligibility and accountability.

Benjamins et al. [36] displays the set of principles published by Telefónica, a large communications company in Spain, and used to guide this company's strategy in developing a methodology for implementing AI ethics. Called Principles of AI, they include the following principles: 1) Fair AI; 2) Transparent and explainable AI; 3) Human-centric AI; 4) Privacy and Security by Design.

Sartor's study [130], published by the European Parliamentary Research Service, explored the impact of the European Union's General Data Protection Regulation (GDPR) – a hard law, in force since 25 May 2018, on AI-based systems. It is noted that the GDPR generally provides significant indications for data protection in relation to AI applications, and apparently does not require expressive changes to address AI, although the GDPR prescriptions are generally vague and do not mention AI explicitly. However, a specification of regulatory and technological requirements regarding AI projects is needed. The author argues that more guidance is needed regarding the generation of prediction models on personal data and the logic of the operation of AI-based systems. Explanations, even high-level ones, should be provided to users so that they can challenge the results. This guidance requires a multi-stakeholder approach, encompassing civil society, specialised agencies and all stakeholders.

There is an overlap between the principles that are listed by the various publicly available guidelines [11]. Several are the criticisms regarding the available AI ethics guidelines [11]: the multiple AI ethics guidelines published by the private sector serve mainly as a marketing strategy, as there are no consequences for not complying with these guidelines. Moreover, the lack of a sense of accountability and the distribution of responsibility, the lack of prior knowledge of the ethical impact, and above all, the financial incentives (e.g., companies expect to produce more in less time), hinder the commitment to the ethical principles present in the guidelines during the development and application of AI-based systems.

3.3 Threats to validity

This Section describes the threats to validity from this systematic literature review, and respective mitigation strategies.

One of the difficulties encountered in carrying out this study is the interrelationship of papers found with the research questions we sought to answer. For example, the work of [88] enumerates ethical principles (answering RQ.3), proposes a framework (answering RQ.1) and addresses how to implement AI ethics during the software development process through design requirements (answering RQ.2). We adopted an approach that could satisfactorily answer the stated research questions, striving for a regular distribution of the studies found in relation to the RQs that were defined. To answer the RQ.3 we gave preference to studies that address reviews of ethical guidelines synthesizing them in a few principles. To answer RQ.1 we chose studies that explore or present some practical means to implement ethical principles. And to answer RQ.2 we prioritized studies that explore the relationship of the software development process with AI ethics or requirements engineering directed to AI or ML based systems, which enable the implementation of ethics in these systems.

In addition, other threats to validity in relation to conducting the SLR are:

- **Research Questions**: the defined questions may not have covered the whole area of ethics in Artificial Intelligence. Therefore, it is not possible to find answers to the questions not defined in this paper. As this factor is considered a real threat, several discussion meetings with the research team were held to fine tune the research questions of the systematic literature review;

- **Subjectivity in study selection**: it is not possible to guarantee that all existing relevant primary studies have been selected. It is possible that relevant documents were not selected. To mitigate this risk, the automatic search strategy and the manual search were performed to try to collect all primary studies of the scope that was defined;

- **Subjectivity in data extraction**: during the data extraction process the primary studies were classified based on the judgment of the researchers. To reduce possible impacts of this problem, the process of study classification was conducted through peer review;

- **Replicability of the systematic process**: there is a risk involving the ability to replicate or extend this systematic literature review. This threat is mitigated through the detailed description of the systematic process of this work, as all details of the systematic literature review protocol have been described. In addition, publication of research findings in conferences and journals was sought to make the findings available in additional sources of information.

3.4 Chapter Summary

In this Chapter a Systematic Literature Review was conducted in order to further deepen and understand the subject of ethics in AI and its different approaches. One of the challenges encountered in performing this study is the interrelationship of the works found with the research questions we sought to answer. For example, the work by [88] enumerates ethical principles (answering RQ.3), proposes a framework (answering RQ.1) and addresses how to implement ethics in AI during the software development process through design requirements (answering RQ.2). We adopted an approach that could satisfactorily answer the stated research questions, seeking an even distribution of the studies found in relation to the RQs that were defined. To answer the RQ.3 we gave preference to studies that address reviews of ethical guidelines synthesizing them in a few principles. To answer RQ.1 we chose studies that explore or present some practical means to implement ethical principles. And to answer RQ.2 we prioritized studies that explore the relationship of the software development process with AI ethics or requirements engineering aimed at AI or ML based systems.

In particular, we highlight the ECCOLA method, proposed by Vakkuri et al. [6], among the studies answering RQ.1 – which address practical means of implementing ethics in AI – as the method whose purpose comes closest to the purpose of our study. This is due to the relationship between the applicability of the method in an agile context and assistance to developers and product owners in eliciting ethical requirements in the development process of AI-based systems. Such a method collaborates with the study of Kostova et al. [95] by making use of the Requirements Engineering phase to address ethical issues (RQ.2). Furthermore, the debate on AI ethics remained in the theoretical context, being addressed primarily through guidelines and principles. To answer RQ.3, we highlight the relevant ethical principles for developers and users [27] [10]: 1) Transparency; 2) Justice and fairness; 3) Non-maleficence; 4) Responsibility; 5) Privacy; 6) Beneficence; 7) Freedom and autonomy; 8) Trust; 9) Sustainability; 10) Dignity; 11) Solidarity.

In Chapter 4, the guide proposal, its definition, objectives, main elements, a pilot project and a prototype aimed at supporting the elicitation of ethical requirements in the first phase of the Software Development Life Cycle, in the context of AI-based systems, will be presented, as a result of the second and third stage of the Design Science Research presented in Section 1.5.

Chapter 4

Guide for Artificial Intelligence Ethical Requirements Elicitation

In this Chapter it will be presented the main aspects involving the decision-making in the development of the suggested Guide, its conception and presentation. It will be presented the phases 2 and 3 of the methodology adopted for the development of this work, Suggestion and Development, respectively. In other words, we present the suggestion as a pilot project, besides reporting the steps for the development of the artifact and presenting the Guide as a prototype.

4.1 Guide Definition

The evolution of the emergence of software that makes use of AI techniques, mostly ML, amplifies the manifestations of accidents and the awareness of the associated ethical issues [132]. In general, ethics in AI has been addressed, in the literature, in its theoretical field, through ethical guidelines [30]. While the existence of guidelines and principles is necessary, little practical direction exists for developers – those responsible for implementing ethics in AI-based systems – to apply in real contexts, even more with the market delivery demands [30], where often the ethical considerations involved is a quality to be considered in the software only after its deployment [6]. Furthermore, developers do not receive adequate training within development projects, nor during their academic studies. There are no legal consequences for not implementing AI ethics, as the guidelines present in the literature, and proposed by organisations, are often non-binding laws (soft law). Thus, there is neither motivation nor punishment for developers in the area of AI ethics.

During the requirements elicitation phase there is a greater interaction between different actors involved in software development and its use, providing a fertile environment for debate on ethical issues [95], and there is a reduction in additional work by considering ethical issues in the early stages of software development, rather than as an afterthought [6].

In the first phase of Design Science Research, as presented in Section 1.5 – Awareness of the problem – a Systematic Literature Review was conducted (Chapter 3. For the main objectives of this work, the first phase had two central usefulness: to identify which method is best suited to our focus, i.e., performing ethical requirements elicitation for AI-based systems during the first phase of software development, requirements analysis; and which ethical principles will be present in the proposed guide. The second phase of the DSR – Suggestion –, and the third phase – Development – will be conducted using primarily as a baseline the information obtained in this first phase. Based on the method and the principles found, a guide will be created to be used by Product Owners and developers – in an agile software development context –, as well as support material for the proposed guide.

4.2 Concept Proposal

In the Suggestion phase, we outline the initial configurations, in a design attempt, in order to elaborate a conceptual proposal, besides its basic criteria, presenting a pilot project of the proposed guide. The Guide will take into consideration the principles elicited by Ryan and Stahl [10] – for condensing a larger amount of ethical guidelines (91), as well as considering developers and users of AI-based systems in their work – and also the ECCOLA method. Vakkuri et al. [6] devised the ECCOLA method employing the AI ethics guidelines HLEG [42] and IEEE EaD v1 [58]. This method was chosen because it allows developers to implement ethical principles through a deck of cards, with themes on AI ethics. We consider the set of principles presented by Ryan and Stahl [10] to be broader and more comprehensive, besides encompassing the guidelines AI HLEG [42] and IEEE EaD v1 [58].

A number of tools that support the implementation of ethical principles in distinct stages of software development were identified in our previous study [134] available in Appendix D. In this study a survey of the tools was conducted in the open source repositories on GitHub. From the identified tools, we mapped them with the addressed ethical principles and present them in the body of the cards, as a possible suggestion of a practical tool for the implementation of the ethical principle addressed in given card. Thus, we adapt the ECCOLA cards with the principles listed by Ryan and Stahl and the tools found in our previous study [134]. These issues are further detailed in Section 4.3.

The following points present in ECCOLA that meet the objectives of this study are presented below:

- Provides developers with a practical tool to implement ethics in AI;
- Uses distinct AI ethics guidelines in practice;
- Support for iterative development;
- Method agnostic – it is possible to use it with any in-house Software Engineering process/method.

We propose an interactive Guide as a web based system that can be accessed by the Product Owner when eliciting the ethical requirements with the development team. The elaboration of an appropriate elicitation technique, providing requirements analysts with adequate tools for ethical requirements elicitation are challenges to be overcome in order to achieve an ethically aware software engineering [26].

Users of the guide should be able to access the system, obtain information regarding its usage and supporting materials, and then select the cards for the process of elicitation of ethical requirements through filters. The ethical principles are the filters contained in the guide, therefore, only corresponding cards to the filter selected by the user will be displayed. In addition, a complementary functionality is implemented, where users can select two or more cards for comparison, and only those displayed side by side, independently from principle at hand.

ECCOLA is a method based on a deck of cards, for the elicitation of ethical requirements in AI, with a particular focus on the context of agile development teams. The use of deck of cards for requirements elicitation in agile development teams is not new in Software Engineering [64], there are methods for performing Requirements Engineering in agile software development, such as Planning Poker [125]. This method comprises of a set of 21 cards, covering 7 principles, with questions to be addressed by the Product Owner and developers, acting as a Planning Poker. From this initial concept, new cards and contents are devised, i.e. the number of cards, their content and principles are altered to adapt them to our context.

In sum, we push forward the ECCOLA method [6], employing: the principles of Ryan and Stahl [10]; tools found in our previous work [134]; as well as modifications that we deem pertinent that have emerged throughout the development of this work. Our guide will be aimed at creating ethical requirements or user stories to serve as items in the product backlog.

4.2.1 Guide Criteria

In this Section it is presented which criteria will be used to propose the guide to perform the implementation of AI ethics in the early phase of the Software Development Cycle.

It was presented in Section 2.2.3 a discussion about the several ways to implement ethics in AI. For the guide proposal, we used an adaptation of the criteria proposed by Schiff et al. [66] for framework, however, with proper adjustments for our work, i.e., aimed at a guide:

1. **Broad:** should consider broad aspects of the impact of AI-based systems to be developed, from different principles and ethical issues to social and economic contexts. Software designers can identify which narrow-scope tools are appropriate, such as those for fairness and explainability, and use them as a sub-part of this guide;

2. **Operationalizable:** users of this guide should be able to articulate the ethical principles, in addition to other desirable requirements of the software to be developed, into specific strategies that can be implemented in AI-based systems, encompassing the identification of relevant actions and decisions assigned to the appropriate phase of the Software Development Cycle;

3. **Flexible:** the guide should be adaptable, adjusting to different use cases, implementation context, organisational settings and different types of AI-based systems, with the intention of having greater applicability, as well as allowing satisfactory customisation and sharing of language and learning. In other words, "there will always be ethical decisions and trade-offs that are not amenable to universally applicable specifications, and that need to be made with sensitivity to specific context and stakeholders" [43];

4. **Iterative:** the guide can be applicable throughout the development cycle of the AI-based software to be developed, and in an iterative way. This is because implementing AI ethics is not done just once, changes may occur, for example in the system itself or in the implementation context, so there is a need to involve the different stakeholders at each stage, and to be re-evaluated over time, and as new issues arise [43].

5. **Guided:** should be easy for users of the guide to use. Users should find it easy to access and understand the guide. The guide should provide sufficient documentation for this;

6. **Participatory:** the guide should incorporate the perspectives of different stakeholders, especially those who may be impacted by the AI-based system developed, the public.

Therefore, the proposed Guide will be broad, by considering different ethical principles, besides indicating possible tools of restricted scope with usability for a specific principle, such as XAI tools for the Transparency principle (e.g., InterpretML [135] and TransparentAI [136]). The Guide will be operationalizable by allowing users to elicit requirements and include them in their Sprint backlogs in the form of user stories, being part of a larger context in which are included the functional and non-functional requirements that the system must fulfill, in addition to the ethical ones.

This criterion is aligned with the purpose of our guide, to help the creation of ethical user stories to serve as items of the product backlog in an agile software development context. In other words, the system requirements are the user stories, which are present in the product backlog, i.e., a list of requirements, that will be worked on by the development team in iterations, called sprints, lasting from 1 to 4 weeks [116].

The guide will be flexible, since the cards have open questions and there are no single answers or only one context for applying the guide. The Guide will be iterative, as users are free to decide the best moment to use the cards, and they can be reused, with the inclusion of different stakeholders in the process.

The Guide will be guided as there will be user documentation, providing the user with a prompt familiarization with the system in a simple and intuitive way. The guide will be participatory, i.e., different stakeholders of the organization can be part of the ethical requirements elicitation activity, including users participation in software development meetings.

4.2.2 Guide pilot project

As a product of the second phase of Design Science Research – Suggestion (Figure 1.1), besides the conceptual proposal, a pilot project was developed. This initial pilot project is a proof of concept, to test the feasibility of the functionalities and practical application of the proposed guide. As a first contribution of this work, we present a pilot project, which implements the ECCOLA method, that is, the entire deck of cards and the explanatory text for its use are present, as detailed in the work presented by the authors of the ECCOLA method [6]. The purpose of this pilot project was to validate the interface and its functionalities, so that in the next stage, it would be possible to modify the content to the objective of this study. In addition, there is no evidence in the literature of the application of this proposal in practice, and it was pointed out by Morley et al. [13] that there is a need to evaluate and test the currently existing tools, in order to identify what works, what can be improved, and what needs to be developed. Thus, we seek with the construction of the pilot project only the reproduction of a method identified to support the elicitation of ethical requirements, however, enabling it in digital form, with graphical interface, and making its access and source code openly available, in https://josesiqueira.github.io/eccola/index.html, https://github.com/josesiqueira/eccola, respectively.

The system was implemented through the use of Hypertext Markup Language (HTML) [137], Cascading Style Sheets (CSS) [138] and JavaScript (JS) [139]. HTML serves as the structure of the system, to indicate where each element of the system will be disposed; CSS for presentation and appearance; and JS for dynamism and action of the system's functionalities. Pure CSS was used, without the use of frameworks, to facilitate the maintainability, because there is no need to learn a framework, besides the scalability, where using just pure CSS it is possible to make improvements in the system, when desired.

The system is also responsive, making it possible to use it on different devices, such as mobile phones, tablets, notebooks, and personal computers. This is possible through the use of Media Queries of the employed CSS, which allows the presentation of the content adapted to a diversity of devices, not requiring to change its content for each device.

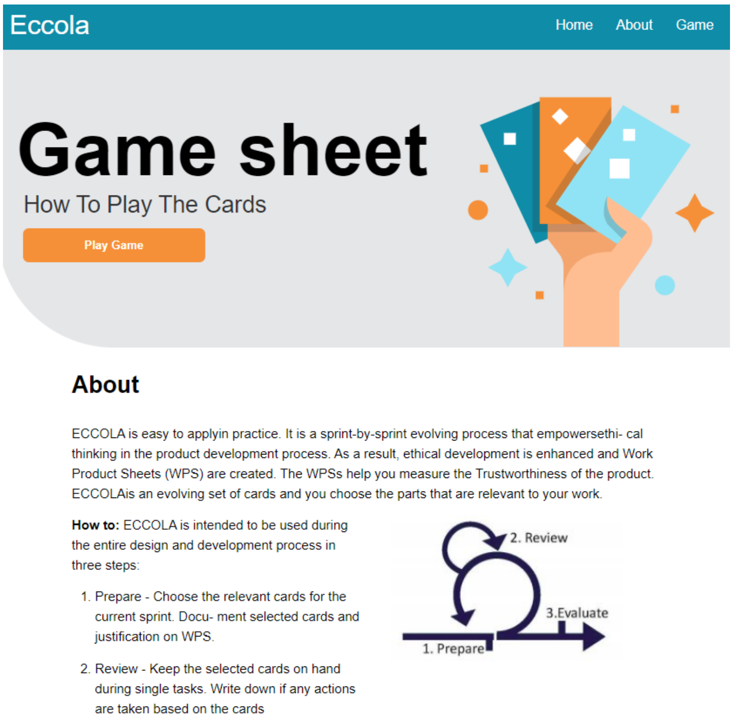
Throughout the development of the system, some measures were observed, such as the HTML implementation in order to enable readability and comprehension through the correct use of semantic HTML (i.e., semantic tags describe the meaning of the content present in the system files), making the reading more straightforward. Furthermore, this is an accessibility enabling element, as our system will also be accessible to visually impaired users, through screen reading tools, such as the NVDA (Non Visual Desktop Access).

In addition to providing a static website with the use of HTML and CSS, JS was used to perform the dynamism and interactivity of users with the system. In this way, users can select cards and compare them, as well as filter cards according to the ethical principle they choose to explore. Also, the dynamism implemented makes it easy to make changes in the system, such as: modify and add cards and ethical principles.

Although our system is not AI based, we aim to contemplate some ethical principles. Through the public availability of the system's source code, the understandability of the code, and the instructions for use, we aim to contemplate the ethical principle of Transparency, in the ethical issues of Explainability, Understandability, Interpretability, Communication, Disclosure and Showing. By allowing the use of screen readers, we contemplate the principles of Beneficence and Dignity.

Immediately when accessing the system, a user can read how to use the cards in an agile development context, and select some options, for example, Play Game, Home and About. The initial screen of the developed pilot project is presented in Figure 4.1.

When the user wishes to start the game, and selects Play Game, a set of cards will be presented, arranged along the screen, where it is possible to read all its contents. However, to filter based on the ethical principles present in the cards, the user must select



Game sheet

How To Play The Cards

About

ECCOLA is easy to applyin practice. It is a sprint-by-sprint evolving process that empowersethi- cal thinking in the product development process. As a result, ethical development is enhanced and Work Product Sheets (WPS) are created. The WPSs help you measure the Trustworthiness of the product. ECCOLAis an evolving set of cards and you choose the parts that are relevant to your work.

How to: ECCOLA is intended to be used during the entire design and development process in three steps:

1. Prepare - Choose the relevant cards for the current sprint. Docu- ment selected cards and justification on WPS.
2. Review - Keep the selected cards on hand during single tasks. Write down if any actions are taken based on the cards

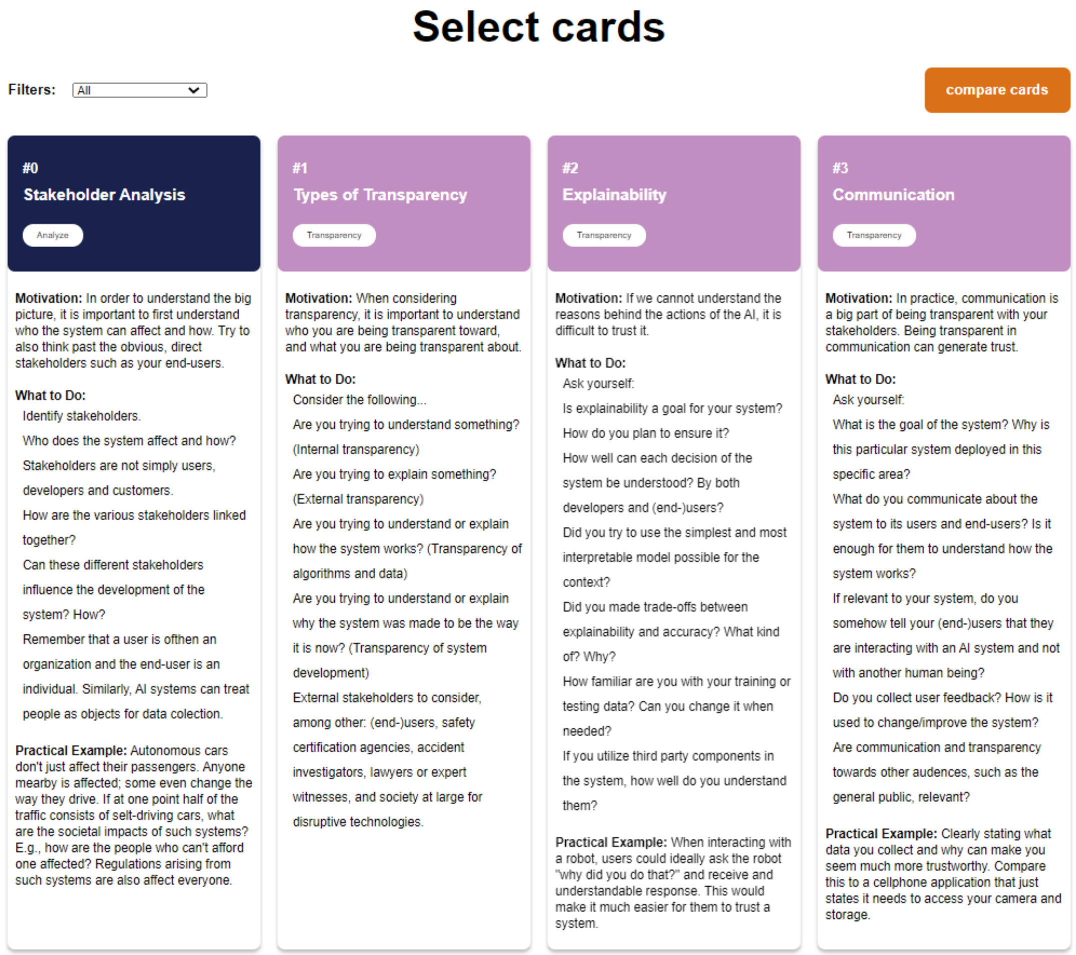
Figure 4.1: Initial screen of the pilot project – Implementation of the ECCOLA method [6].

the "Filter" dropdown list, and choose the desired filter. In addition, it is possible to select 2 or more cards, and compare them by selecting the "Compare" button, regardless of whether they belong to the same principle or not. Figure 4.2 presents the game screen where the user can select the cards in the pilot project.

In this Section, we described the initial proposal, the Guide criteria, and presented the initial pilot project of the system, implementing the ECCOLA method in practice, in order to obtain a proof of concept, as a practical model to implement the concepts established in our study. In the next Section, we present the third stage of Design Science Research – Development (Figure 1.1) –, where we modify and evolve the initial pilot project, presenting our arguments, and at the end, the prototype of our guide.

4.3 Guide Development

After the end of the second phase, we begin the third phase of the DSR – Development – where we address in a more in-depth way which steps were fulfilled to achieve the objectives of our proposal in the system design, conceptually defined in Section 4.2, in order to improve the pilot project and define the prototype. Our artifact, therefore, will be a guide that will serve as an assistant to support the elicitation of ethical requirements in the context of AI-based systems. In this phase, we will follow a set of steps for the



Select cards

Filters: All

compare cards

#0

Stakeholder Analysis

Analyze

Motivation: In order to understand the big picture, it is important to first understand who the system can affect and how. Try to also think past the obvious, direct stakeholders such as your end-users.

What to Do:

Identify stakeholders.

Who does the system affect and how?

Stakeholders are not simply users, developers and customers.

How are the various stakeholders linked together?

Can these different stakeholders influence the development of the system? How?

Remember that a user is ofthen an organization and the end-user is an individual. Similarly, AI systems can treat people as objects for data colection.

Practical Example: Autonomous cars don't just affect their passengers. Anyone mearby is affected; some even change the way they drive. If at one point half of the traffic consists of selt-driving cars, what are the societal impacts of such systems? E.g., how are the people who can't afford one affected? Regulations arising from such systems are also affect everyone.

#1

Types of Transparency

Transparency

Motivation: When considering transparency, it is important to understand who you are being transparent toward, and what you are being transparent about.

What to Do:

Consider the following...

Are you trying to understand something? (Internal transparency)

Are you trying to explain something? (External transparency)

Are you trying to understand or explain how the system works? (Transparency of algorithms and data)

Are you trying to understand or explain why the system was made to be the way it is now? (Transparency of system development)

External stakeholders to consider, among other: (end-)users, safety certification agencies, accident investigators, lawyers or expert witnesses, and society at large for disruptive technologies.

#2

Explainability

Transparency

Motivation: If we cannot understand the reasons behind the actions of the AI, it is difficult to trust it.

What to Do:

Ask yourself:

Is explainability a goal for your system? How do you plan to ensure it?

How well can each decision of the system be understood? By both developers and (end-)users?

Did you try to use the simplest and most interpretable model possible for the context?

Did you made trade-offs between explainability and accuracy? What kind of? Why?

How familiar are you with your training or testing data? Can you change it when needed?

If you utilize third party components in the system, how well do you understand them?

Practical Example: When interacting with a robot, users could ideally ask the robot "why did you do that?" and receive and understandable response. This would make it much easier for them to trust a system.

#3

Communication

Transparency

Motivation: In practice, communication is a big part of being transparent with your stakeholders. Being transparent in communication can generate trust.

What to Do:

Ask yourself:

What is the goal of the system? Why is this particular system deployed in this specific area?

What do you communicate about the system to its users and end-users? Is it enough for them to understand how the system works?

If relevant to your system, do you somehow tell your (end-)users that they are interacting with an AI system and not with another human being?

Do you collect user feedback? How is it used to change/improve the system?

Are communication and transparency towards other audences, such as the general public, relevant?

Practical Example: Clearly stating what data you collect and why can make you seem much more trustworthy. Compare this to a cellphone application that just states it needs to access your camera and storage.

Figure 4.2: Cards selection screen in the pilot project. Implementation of the ECCOLA Method [6].

development of the AI Ethical Requirements Elicitation Guide. First, we will delimit the set of ethical principles to be used, then we will define the set of possible tools that will serve as a suggestion to developers, finally, we will present the cards that will compose the deck, where each card will comprise a tool and a principle, besides other pertinent information, with the goal of creating a prototype of the artifact and presenting it. Figure 4.3 presents an overview of the steps that will guide the process of developing the Guide content. The following sections describe the process of the development steps in more depth.

4.3.1 Delimitation of Ethical Principles

There are several guidelines containing ethical principles serving as normative guides for AI ethics, and as of November 2019, at least 84 organisations – public, private, government, the academia and civil society – have been publishing reports describing ethical

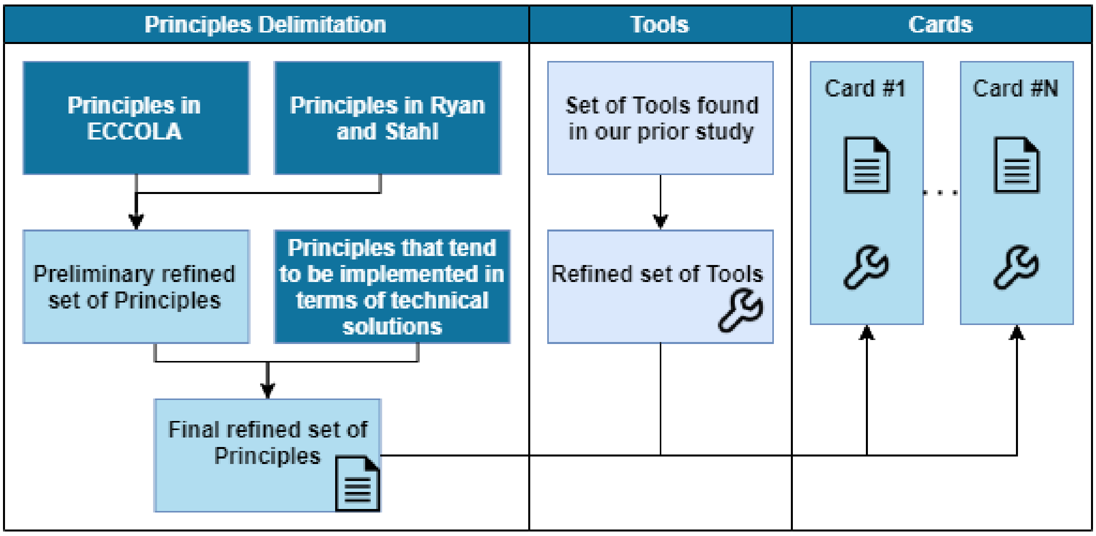

Principles Delimitation
Tools
Cards
Principles in ECCOLA
Principles in Ryan and Stahl
Set of Tools found in our prior study
Card #1
Card #N
Preliminary refined set of Principles
Principles that tend to be implemented in terms of technical solutions
Refined set of Tools
Final refined set of Principles

Figure 4.3: Exploration diagram of the development of the Guide for Elicitation of Ethical Requirements in AI. Own source

principles, values or other abstract high-level requirements for the development and deployment of AI [30]. Therefore, there is an initial challenge in choosing which ethical principles will be used in the development of the guide proposed in this work. Throughout this study, a set of principles useful to our goal have been identified through the Systematic Literature Review presented in section 3.2.3, and set out in more depth in section 2.2.2. However, it is a challenge to include all the principles identified in our Systematic Literature Review – for example: how to elicit requirements related to the principles of Solidarity, or Dignity? Moreover, the ethical principles present in the ECCOLA method cards are not part of the same set of principles identified in our Systematic Literature Review, and are often named in different ways, but with similar content. Next, we describe the decisions to choose the set of principles to be used in the Guide.

In the development of the Guide, it will be taken as central axis the principles listed by Ryan and Stahl [10]. The principles present in the pilot project created are the same selected by Vakkuri et al. [6]. We will present these principles, and then those of Ryan and Stahl [10]. Subsequently, we will perform a mapping between these sets of principles in order to explore how they are related, displaying a preliminary set of refined principles. Finally, we introduce a subset of principles that are more easily implementable in terms of mathematical solutions (presented by Hagendorff [11]) in order to provide the practical tool suggestions, relating them to the first set of principles found, presenting a final refined set of Principles. In short, we standardized both the principles in Vakkuri et al. [6] and Hagendorff [11] with the principles in Ryan and Stahl [10] through mappings presented in tables. In Figure 4.3 one can visualize the steps we adopted for the delimitation of the ethical principles chosen for the prototype of the proposed guide, in the lane called

Principles Delimitation.

First, there are seven principles present in the ECCOLA method – based primarily on the AI HLEG [42] and the IEEE EADv1 [58] – as presented in Table 4.1, and the developed pilot project. Ryan and Stahl's principles are presented in Table 4.2.

Table 4.1: Ethical principles present in the work of Vakkuri et al. [6]

#	Principle
1	Transparency
2	Data
3	Agency & Oversight
4	Safety and Security
5	Fairness
6	Wellbeing
7	Accountability

Table 4.2: Principles and their ethical issues presented in the work of Ryan and Stahl [10]

#	Principles	Ethical issues
1	Transparency	Transparency, Explainability; Explicability; Understandability; Interpretability; Communication; Disclosure; Showing
2	Justice and fairness	Justice; Fairness; Consistency; Inclusion; Equality; Equity; Non-bias; Non-discrimination; Diversity; Plurality; Accessibility; Reversibility; Remedy; Redress; Challenge; Access and distribution
3	Non-maleficence	Non-maleficence; Security; Safety; Harm; Protection; Precaution; Prevention; Integrity; Non-subversion
4	Responsibility	Responsibility; Accountability; Liability; Acting with integrity
5	Privacy	Privacy; Personal or private information
6	Beneficence	Benefits; Beneficence; Well-being; Peace; Social good; Common good
7	Freedom and autonomy	Freedom; Autonomy; Consent; Choice; Self-determination; Liberty; Empowerment
8	Trust	Trustworthiness
9	Sustainability	Sustainability; Environment (nature); Energy; Resources (energy)
10	Dignity	Dignity
11	Solidarity	Solidarity; Social security; Cohesion

In the ECCOLA method, for each principle, there are a number of cards that may contain an ethical issue distinct from another in its same set. In other words, within a set of cards of the same principle, there is more than one ethical issue associated. For this reason, we relate each ECCOLA card to a principle and an ethical issue in Ryan and

Stahl, and present this relationship in Table 4.3.

Table 4.3: Mapping the cards in ECCOLA with Ryan and Stahl [10]

#	Principle in Vakkuri et al. [6]	Card title	Principle in Ryan and Stahl [10]	Ethical issue
1	Transparency	Types of Transparency	Transparency	Transparency, Explainability
2	Transparency	Explainability	Transparency	Explainability, Explicability, Understandability
3	Transparency	Communication	Transparency	Communication, Disclosure, Showing
4	Transparency	Documenting Trade-offs	Transparency	Communication
5	Transparency	Traceability	Transparency	Explicability
6	Transparency	System Reliability	Transparency	Explainability, Explicability
7	Data	Privacy and Data	Privacy	Privacy, Personal or private information
8	Data	Data Quality	Responsibility	Acting with integrity
9	Data	Access to Data	Privacy	Personal or private information
10	Agency & Oversight	Human Agency	Transparency	Interpretability, Showing
11	Agency & Oversight	Human Oversight	Freedom and autonomy / Justice and fairness	Self-determination / Reversibility, Remedy, Redress
12	Safety & Security	System Security	Non-maleficence	Non-maleficence, Security, Safety
13	Safety & Security	System Safety	Non-maleficence	Harm, Protection, Precaution, Prevention
14	Fairness	Accessbility	Justice and fairness	Inclusion, Equality, Equity
15	Fairness	Stakeholder Participation	Justice and fairness	Diversity, Plurality
16	Wellbeing	Environmental Impacts	Sustainability	Sustainability, Environment (nature), Energy, Resources (energy)
17	Wellbeing	Societal Effects	Beneficence	Social good, Common good
18	Accountability	Auditability	Responsibility	Accountability
19	Accountability	Ability to Redress	Responsibility	Responsibility, Liability
20	Accountability	Minimizing Negative Impacts	Responsibility	Acting with integrity

After obtaining the preliminary refined set of principles by mapping each ECCOLA card with an ethical principle and issue in Ryan and Stahl, we employ a last criterion to analyse the principles used in our proposal: we will use the principles that share the characteristic of "being more easily implemented mathematically and therefore tend to be implemented in terms of technical solutions" [11]. According to Hagendorff [11], these principles are: "Accountability; Explainability; Privacy; Fairness; but also other values such as robustness or Safety". In order to present them in our context, we perform the mapping of these principles with those defined by Ryan and Stahl (the preliminary refined set of principles) and their respective ethical issues. For instance, Accountability in Hagendorff is an ethical issue in the principle of Responsibility in Ryan and Stahl. As such, the principles in Hagendorff are a subset of the principles defined by Ryan and Stahl [10] in Table 4.2, and we present this mapping in Table 4.4. This subset of principles will be useful for mapping the tools found in Siqueira et al. [134] with the principles that will be present in our guide, as can be seen in the next Section. Moreover, it will serve for future research or evolution of this work, in the task of including more practical tools to operationalise AI ethics along with the related principles.

Table 4.4: Principles more easily implementable in terms of technical solutions [11]

#	Principle in Hagendorff	Principle in Ryan and Stahl	Ethical issues
1	Explainability	Transparency	Explainability; Explicability; Understandability; Interpretability
2	Justice	Justice and fairness	Consistency; Inclusion; Equality; Equity; Non-bias; Non-discrimination; Diversity; Plurality; Accessibility; Reversibility; Remedy; Redress; Challenge; Access; Distribution
3	Safety	Non-maleficence	Security; Safety
4	Accountability	Responsibility	Accountability
5	Privacy	Privacy	Privacy; Personal or private information

Finally, conducting the process of delimitation of the principles to be used in the guide, explained in Figure 4.3, in the first lane – Principles Delimitation –, and carried on throughout this section, allowed us to standardize the principles indicated by Vakkuri et al. [6] and Hagendorff [11] with the principles elicited by Ryan and Stahl [10], presenting the final refined set of Principles. Standardizing the principles presented in Table 4.1 to Ryan and Stahl's principles (Table 4.2), accomplished in Table 4.3 – allows us to reuse the ECCOLA method cards, as well as being an initial starting point for our guide; and the standardization of the principles in Hagendorff to Ryan and Stahl's principles, performed in Table 4.4, allows us to facilitate the association of the refined set of tools, or new tools, with the principles arranged in the final refined set of Principles, reducing the complexity of the evaluation of the set of 11 principles.

In sum, the principles and ethical issues that will compose our guide are those present in Table 4.2, listed by Ryan and Stahl [10]. Several authors have used different methodological approaches to analyse sets of documents and extract the most recurrent principles and their definitions, generally concluding that they are too general, have a high level of abstraction and a degree of difficulty in their application in real contexts, as well as an overlapping among the principles [11] [27] [48] [49] [45] [62] [63] [10] [28]. Amongst these studies, Jobin et al. [27] drew the attention of the global community for analysing a set of 84 AI ethics guidelines, extracting 11 ethical principles. The work of Ryan and Stahl, builds on the work of Jobin et al. [27] and their 11 principles, extends the set of ethical guidelines analysed to 91, furthermore, they present the ethical issues of each principle directed at developers and users of AI-based systems, aligning with the context of this work. To the best of our knowledge, this is the study that makes use of a methodology that encompasses the largest amount of guidelines and definitions, presenting a comprehensive and concise taxonomy.

4.3.2 Tool set Analysis

In this next step, we identify the tools that will be present in the content of the cards. We explore the tools identified in our previous work [134] available in Appendix D. In this work, we identified only 21 tools that assist the implementation of AI ethics. The tools were mapped with the principles and ethical issues presented in Table 4.4, and we present the result in Table 4.5.

Table 4.5: Mapping of tools found with principles

ID	Tool	Principle	Ethical issue	Justification
T1	DALEX	Transparency	Explainability; Explicability; Understandability; Interpretability	The DALEX package takes an X-ray of any model and helps to explore and explain its behavior, helps to understand how complex models are working.

T2	Melusine	N/A	N/A	Apart. A high-level Python library for email classification and feature extraction with a focus on the French language. Contains Ethical Guidelines for evaluating AI design based on the AI HLEG [42] in French.
T3	Interpretable AI	Transparency	Explainability; Explicability; Understandability; Interpretability	Apart. A list of Interpretability techniques for building robust AI applications and examples of AI propagating biases.
T4	InterpretML	Transparency	Explainability; Explicability; Understandability; Interpretability	A Microsoft open source package that incorporates machine learning techniques where it is possible to train interpretable models and explain black box systems, supporting global understanding of models or the reasons behind predictions.
T5	Deon	Transparency / Justice and fairness / Non-maleficence / Responsibility / Privacy / Freedom and autonomy	Explainability; Explicability; Understandability; Interpretability / Consistency; Inclusion; Non-bias; Non-discrimination; Diversity; Plurality; Reversibility; Remedy; Redress / Security; Safety; Harm / Accountability ; Personal or private information / Consent; Choice; Self-determination	Apart. A command line tool that allows adding an ethical checklist to data science projects.

T6	Fooling LIME and SHAP	Transparency	Explainability; Explicability; Understandability; Interpretability	Apart. Code from an article [140] where the authors aim to deceive LIME and SHAP (two XAI tools)
T7	TransparentAI	Transparency / Justice and fairness / Non-maleficence / Responsibility / Sustainability	Explainability; Explicability; Understandability; Interpretability; Showing / Non-bias; Redress / Security, Safety; Harm / Responsibility, Accountability; Acting with integrity / Energy, Resources (energy)	A toolbox in Python to know if an AI-based system is ethical, based on the AI HLEG [42].
T8	CALIMOCHO	Transparency	Explainability; Explicability; Understandability; Interpretability	An implementation of Explanatory Active Learning (XAL) based on Self-explanatory Neural Networks.
T9	social and Ethics in ML	Justice and fairness / Privacy	fairness, Non-bias, Consistency / Personal or Private information	Apart. It shows how privacy and equity was achieved in a project.
T10	SWED	N/A	N/A	Apart. An educational argument diagramming tool for the domain of Software Engineering ethics, with a specific version to discuss AI ethics.
T11	AI collaboratory	N/A	N/A	Apart. A project that performs analysis, evaluation, comparison and classification on some pre-defined datasets.
T12	Variational Fair Autoencoders (VFAE)	Justice and fairness	Non-bias	A tool that enables training models and obtaining predictions that are less biased by the sensitive properties of people on a pre-defined dataset.

T13	The Impartial Machines Project	Justice and fairness	Non-bias	A tool that attempts to eliminate potential influences/biases in news.
T14	Fairness-Aware-Ranking in Search & Recommendation Systems	Justice and fairness	Non-bias	A tool that attempts to eliminate potential influences/ biases in ranked lists generated by recommender systems.
T15	Fair-ML-4-Ethical-AI	Justice and fairness	Non-bias	Pedagogical resources for bias detection and elimination in datasets using R. In French.
T16	Deon-feedstock	N/A	N/A	Discarded. A repository containing continuous integration support and configuration scripts for Deon.
T17	Multi Accuracy Boost	Transparency / Justice and fairness	Explainability; Explicability; Understandability; Interpretability / Non-bias	UA tool for auditing and post-processing black-box algorithms to ensure accurate predictions in datasets with protected attributes [141].
T18	Fair-Forest	Justice and fairness	Non-bias	A Java library that attempts to eliminate potential influences/ biases in decision trees and random forests [142].
T19	ABOD3	Transparency	Explainability; Explicability; Understandability; Interpretability	ABOD3 is an integrated development environment (IDE) for Behavior Oriented Design (BOD) that allows you to visualize, develop and debug AI in real time [143].
T20	Scruples	N/A	N/A	Discarded. Addresses ethical dilemmas.
T21	Freelance Developer Toolbox	N/A	N/A	Discarded. A curated list of tools.

Repositories were identified that have no usefulness to our goal (T16, T20, T21) – Discarded – because they contain supporting code, curated list (e.g., a list on a given topic that has been carefully compiled, usually by a survey) of tools outside our scope, and address ethical dilemmas in AI. In addition, some tools will be present in the supporting material of the Guide, as an extra reference, however, they will not be in the body of any card (T2, T3, T5, T6, T9, T10, T11) – Apart –, because they do not present a practical tool, but include information pertinent to developers about implementing ethical principles in AI in some way (e.g., educational tool, privacy implementation in a project).

The tools included in the content cards of our guide (T1, T4, T7, T8, T12, T13, T14, T15, T17, T18, T19) – that actually assist in implementing ethical principles in AI-based systems – total in 11 tools, composing our refined set of tools, visible in the Tools lane of Figure 4.3.

We have found, through our interpretation of the README files of the repositories and associating them with principles (Table 4.2), that there are tools that extrapolate the subset of tools that are more easily implementable through mathematical solutions, presented in Table 4.4. One of these is TransparentAI, a Python toolbox that tests an AI-based system against ethical principles. Besides Transparency, Justice and fairness, Non-maleficence and Responsibility, this tool also operationalises Sustainability. This is due to the fact that the tool is based on the principles listed in the AI HLEG [42], where the authors do not present technical solutions for all principles, claiming that "Some aspects do not have technical implementation in this tool because it requires legal or other knowledge." Therefore, this tool proves to be conducive to our context.

The second tool is Deon, a command-line tool that allows the addition of customizable ethical checklists in an AI-based project. By default, this checklist contemplates, in addition to Transparency, Justice and fairness, Non-maleficence and Responsibility, it also operationalises the principles of Privacy, Freedom and Autonomy. In a distinct way from the first tool, Deon only raises questions to be answered by developers, as a checklist to be fulfilled, realized as a Python technical tool, to be added to a project. However, checklists should not be the only mechanism for ethics in AI [11], and can be misused or even ignored if practitioners are not involved in their design or implementation [71]. Furthermore, the implementation of AI ethics "must not consist solely of a one-off tick-box exercise completed only at the beginning of the Design process" [132]. Both tools are promising, however, we separate Deon as complementary material (for its checklist feature), and only TransparentAI, between these two, will be part of the content of the cards.

From the initial set of tools (21), we filtered 11 as the refined set of tools, and based on their relationships with our final refined set of principles, we note that there is a trend in tools operationalising the principles of Transparency and Justice and fairness. Some tools operationalise more than one principle.

1. Transparency - T1, T4, T7, T8, T17, T19 - Total: 6

2. Justice and fairness - T7, T12, T13, T14, T15, T17, T18. Total: 7

3. Non-maleficence - T7 - Total: 1

4. Responsibility - T7 - Total: 1

5. Sustainability - T7 - Total: 1

Thus, we conclude the tool selection step by presenting the refined set of tools and their rationale (Table 4.5).

4.3.3 Definition of the Content of the Cards

The Guide for Artificial Intelligence Ethical Requirements Elicitation will consist of a deck of cards. The cards will be separated by principles and ethical issues. Each principle may have more than one ethical issue, i.e. more than one card will be available for each principle. Each card is composed of four parts:

1. **Preamble** – why this is important;

2. **Issues to be addressed** – to tackle this issue;

3. **Illustration** of this topic – to further exemplify the issue;

4. **Tool Suggestion** – tools available on GitHub that support the implementation of the ethical issue.

Items 1, 2 and 3 are adapted from Vakkuri et al. [6] and are available in our pilot project. In **Preamble**, there is an observation of why it is important to address this issue, as something positive to be achieved that reflects on the user in the end, or provides an overview of that topic. It is noticeable that, in **Issues to be addressed**, there is no single, direct and objective indication of what developers should do, but there are questions, which users of the guide need to discuss, in order to operationalise ethics in AI. In this way, ethical awareness among the development team is increased. In **Illustration** we further illustrate the issue by offering a case where ethical requirements were not considered and led to incidents, or the illustration of the topic in a specific context. In **Tool Suggestion** we offer the options of available tools in the refined set of Tools, however, it was seen that this set of Tools does not cover all the principles in the Guide, i.e. this field is not mandatory and will not appear in all the cards, as there are no tools available for all ethical issues.

Two more cards compose the guide, related to stakeholder analysis and assessment. Important to note that, in the ECCOLA method originally conceived, and present in our pilot project, there is an initial card, which must be addressed before the user explores other cards, called Stakeholder Analysis. This card motivates the developer to answer questions related to stakeholder assessment, who they are, how they are affected and how they are related. Stakeholder analysis, as described by Vakkuri et al. [6] converges with the concept of Runtime Stakeholders introduced by Guizzardi et al. [23]. The latter work, states that a key concept for obtaining/eliciting ethical requirements is that of Runtime Stakeholders: “These include those stakeholders that are using, affected by, or influencing the outcomes of a system as it is operating”. An example of Runtime Stakeholders in the context of AI-based systems for healthcare are: patients and their families, the doctors, nurses, x-ray operators and other healthcare professionals. Thus, this card shows to be crucial to allow the other cards to be applied.

The second additional card, deals exclusively with the evaluation of the AI-based system being developed. Ethical evaluation must become an integral part of the operation of a system, or there is no guarantee that tools – such as this guide – will have any positive impact on the ethical implications of AI systems [132]. There is a caution to insert on some cards throughout the deck, under “Issues to be addressed”, whether the system allows for evaluation (e.g., internal, or external, and to what extent), however, it is observed the urgency of inserting a card exclusively for this purpose, in that oversight, at the evaluation stage, is “concerned with whether the algorithmic system is continuing to operate in the right way once deployed, needs to be revised, or can be improved” [132]. In other words, even after the system is deployed, the development team should define a time interval between one evaluation and another. The need for the inclusion of this card also emerges from the distribution of responsibility between the components defined by Morley et al. [132] on AI ethics governance: an independent multi-disciplinary ethics board; and the AI professionals themselves. According to Morley et al. [132], positive ethical qualities are susceptible to progressive increase, that is, “an algorithm can be increasingly fair, and fairer than another algorithm or a previous version, but makes no sense to say that it is fair or unfair in absolute terms”.

In sum, the Stakeholders’ assessment card will be card # 0, as in Vakkuri et al. [6], to be addressed at the beginning (before the other cards), and we have added a card dealing with system assessment, as a last card, which should be periodically revisited.

4.3.4 Defining the logo

The purpose of defining a logo is to assign an identity to the Guide, in addition, we will define the acronym that will be used. The name of the artifact created as a prototype in this work is **Guide for Artificial Intelligence Ethical Requirements Elicitation**. An acronym that captures the essence of our proposal has been defined as **RE4AI Ethical Guide**. The term “Guide” is meant to be a document providing information on the subject, helping people form opinions or make decisions, directing or influencing software development behavior, showing or pointing the way for the development team. Thus, Guide for Artificial Intelligence Ethical Requirements Elicitation in this context aims to: guide software development teams through open-ended question cards to elicit requirements that conform to ethical principles for AI-based systems. Figure 4.4 shows the logo of the RE4AI Ethical Guide.

Figure 4.4: RE4AI Ethical Guide logo. Own source

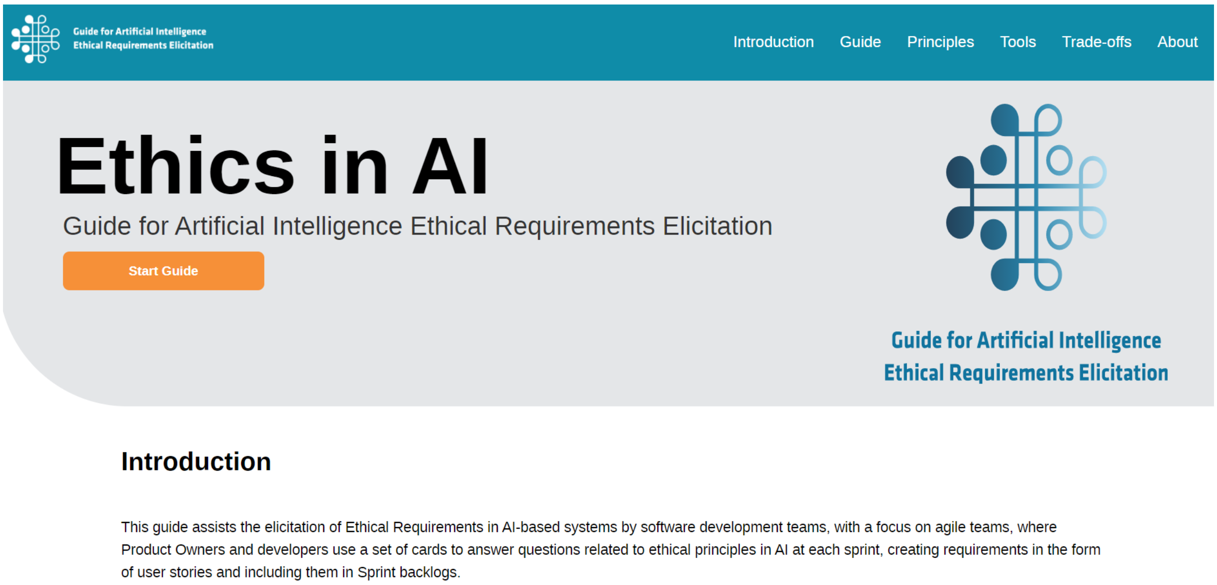
The logo is a synthesis of the two main ideas of the project: Ethics and Artificial Intelligence. The symbol refers to a brain formed by circuits, the balance of the design is broken by the filled circles on the left and the open ones on the right, representing the dichotomy of ethics, where one must choose the correct path. The Klavika font was chosen for the name for being a typographic family without serifs and contemporary, in addition, it was chosen the same colour scheme used throughout this study, in order to maintain regularity and preserve consistency.

4.3.5 Guide Overview

In this Section we present an overview of the guide devised. The Guide for Artificial Intelligence Ethical Requirements Elicitation was implemented as a web-based system and is divided into: Introduction, presenting a brief introduction and how to use it; Guide, presenting the set of cards; Principles, presenting all the principles present in the guide; Tools, presenting which tools are present in the guide related to the principles; Trade-offs, presenting which trade-offs may occur when developing AI-based systems that take ethical issues into consideration; and About, briefly presenting the authors, information about the guide and references used. In Figure 4.5 we present the initial screen of the guide, where its subdivisions are present.

By clicking Start Guide, the user will be presented by default with all the cards, and the options to filter or compare cards, as illustrated in Figure 4.6. These guide features are the same as described in our pilot project in Section 4.2.2.

In order to display only the cards related to a specific principle, the user must select the desired principle from the Filters menu (on the left). At the top of the card there is the card number, its ethical issues, and the ethical principle, besides that, the principles are related to different colours. The user can click on the tool provided in the Tool Suggestion field, where a new tab will open in the browser, displaying the source code repository on


Guide for Artificial Intelligence Ethical Requirements Elicitation
Introduction
Guide
Principles
Tools
Trade-offs
About
Start Guide
Guide for Artificial Intelligence Ethical Requirements Elicitation

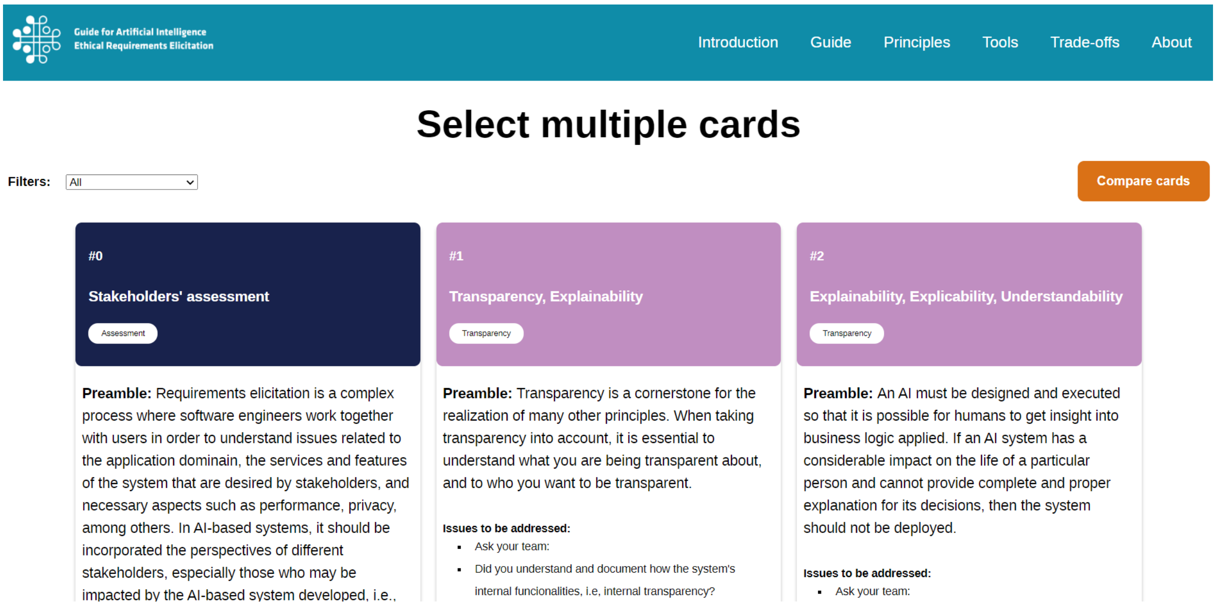
Ethics in AI

Guide for Artificial Intelligence Ethical Requirements Elicitation

Introduction

This guide assists the elicitation of Ethical Requirements in AI-based systems by software development teams, with a focus on agile teams, where Product Owners and developers use a set of cards to answer questions related to ethical principles in AI at each sprint, creating requirements in the form of user stories and including them in Sprint backlogs.

Figure 4.5: Home page of the guide. Own source



Select multiple cards

Filters: All

#0

Stakeholders' assessment

Preamble: Requirements elicitation is a complex process where software engineers work together with users in order to understand issues related to the application domainn, the services and features of the system that are desired by stakeholders, and necessary aspects such as performance, privacy, among others. In AI-based systems, it should be incorporated the perspectives of different stakeholders, especially those who may be impacted by the AI-based system developed, i.e.,

#1

Transparency, Explainability

Preamble: Transparency is a cornerstone for the realization of many other principles. When taking transparency into account, it is essential to understand what you are being transparent about, and to who you want to be transparent.

Issues to be addressed:

- Ask your team:
- Did you understand and document how the system's internal funcionalities, i.e, internal transparency?

#2

Explainability, Explicability, Understandability

Preamble: An AI must be designed and executed so that it is possible for humans to get insight into business logic applied. If an AI system has a considerable impact on the life of a particular person and cannot provide complete and proper explanation for its decisions, then the system should not be deployed.

Issues to be addressed:

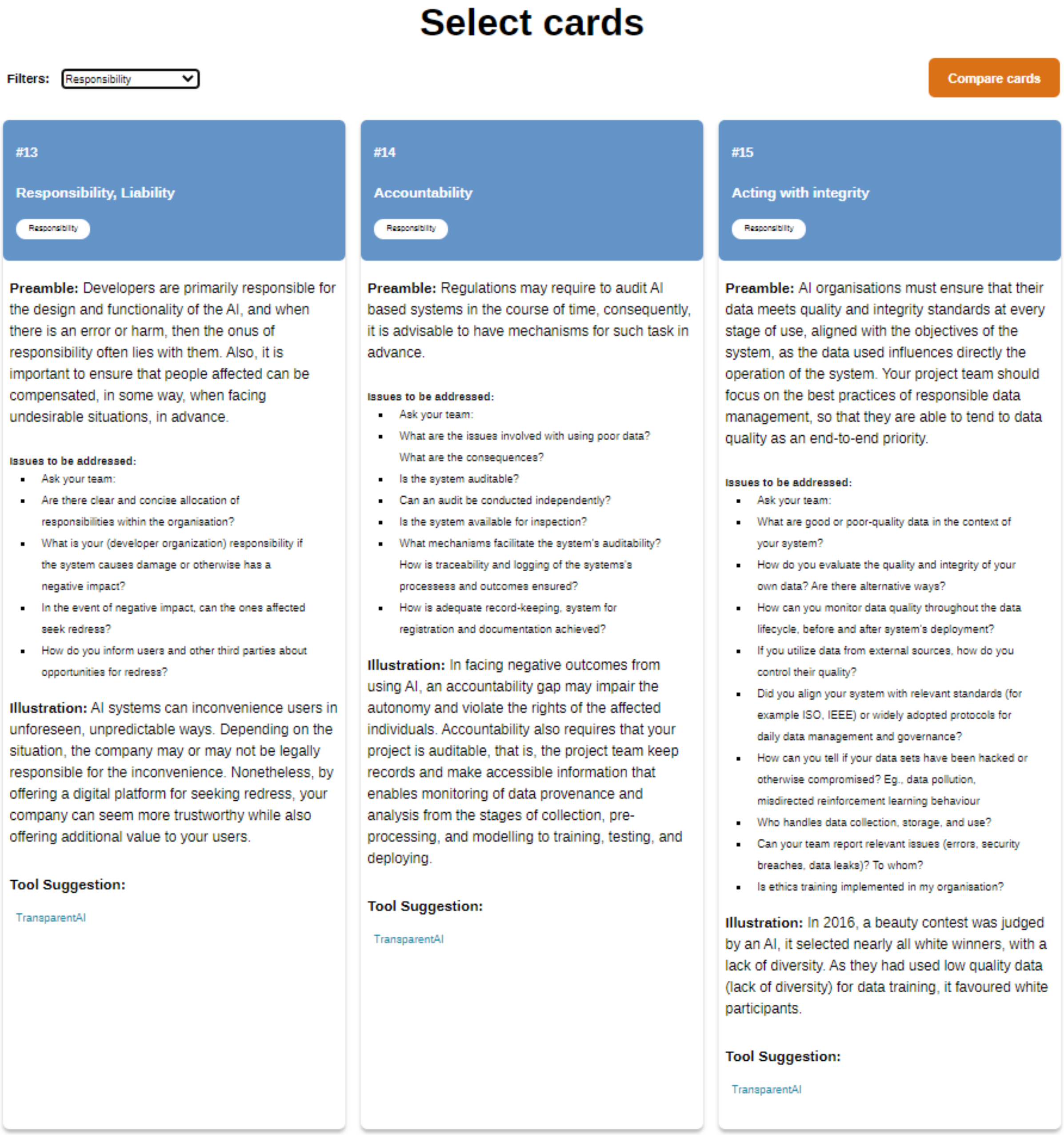
- Ask your team:

Figure 4.6: Card selection page. Own source

GitHub of the respective tool. We illustrate in Figure 4.7 the scenario in which the user selects the principle Responsibility.

If 2 or more cards are selected, the user can click Compare cards (on the right), where only those cards will be displayed. The user can then click Start again to return to the previous screen where all cards are displayed. In Figure 4.8 we illustrate the scenario where 4 cards of different principles are selected and compared.

In the footnotes is available the address of the source code of the system, in addition to the license Creative Commons 4.0 International (CC BY 4.0), in order to allow the



#13

Responsibility, Liability

Responsibility

Preamble: Developers are primarily responsible for the design and functionality of the AI, and when there is an error or harm, then the onus of responsibility often lies with them. Also, it is important to ensure that people affected can be compensated, in some way, when facing undesirable situations, in advance.

Issues to be addressed:

- Ask your team:
- Are there clear and concise allocation of responsibilities within the organisation?
- What is your (developer organization) responsibility if the system causes damage or otherwise has a negative impact?
- In the event of negative impact, can the ones affected seek redress?
- How do you inform users and other third parties about opportunities for redress?

Illustration: AI systems can inconvenience users in unforeseen, unpredictable ways. Depending on the situation, the company may or may not be legally responsible for the inconvenience. Nonetheless, by offering a digital platform for seeking redress, your company can seem more trustworthy while also offering additional value to your users.

Tool Suggestion:

TransparentAI

#14

Accountability

Responsibility

Preamble: Regulations may require to audit AI based systems in the course of time, consequently, it is advisable to have mechanisms for such task in advance.

Issues to be addressed:

- Ask your team:
- What are the issues involved with using poor data? What are the consequences?
- Is the system auditable?
- Can an audit be conducted independently?
- Is the system available for inspection?
- What mechanisms facilitate the system's auditability? How is traceability and logging of the systems's processess and outcomes ensured?
- How is adequate record-keeping, system for registration and documentation achieved?

Illustration: In facing negative outcomes from using AI, an accountability gap may impair the autonomy and violate the rights of the affected individuals. Accountability also requires that your project is auditable, that is, the project team keep records and make accessible information that enables monitoring of data provenance and analysis from the stages of collection, pre-processing, and modelling to training, testing, and deploying.

Tool Suggestion:

TransparentAI

#15

Acting with integrity

Responsibility

Preamble: AI organisations must ensure that their data meets quality and integrity standards at every stage of use, aligned with the objectives of the system, as the data used influences directly the operation of the system. Your project team should focus on the best practices of responsible data management, so that they are able to tend to data quality as an end-to-end priority.

Issues to be addressed:

- Ask your team:
- What are good or poor-quality data in the context of your system?
- How do you evaluate the quality and integrity of your own data? Are there alternative ways?
- How can you monitor data quality throughout the data lifecycle, before and after system's deployment?
- If you utilize data from external sources, how do you control their quality?
- Did you align your system with relevant standards (for example ISO, IEEE) or widely adopted protocols for daily data management and governance?
- How can you tell if your data sets have been hacked or otherwise compromised? Eg., data pollution, misdirected reinforcement learning behaviour
- Who handles data collection, storage, and use?
- Can your team report relevant issues (errors, security breaches, data leaks)? To whom?
- Is ethics training implemented in my organisation?

Illustration: In 2016, a beauty contest was judged by an AI, it selected nearly all white winners, with a lack of diversity. As they had used low quality data (lack of diversity) for data training, it favoured white participants.

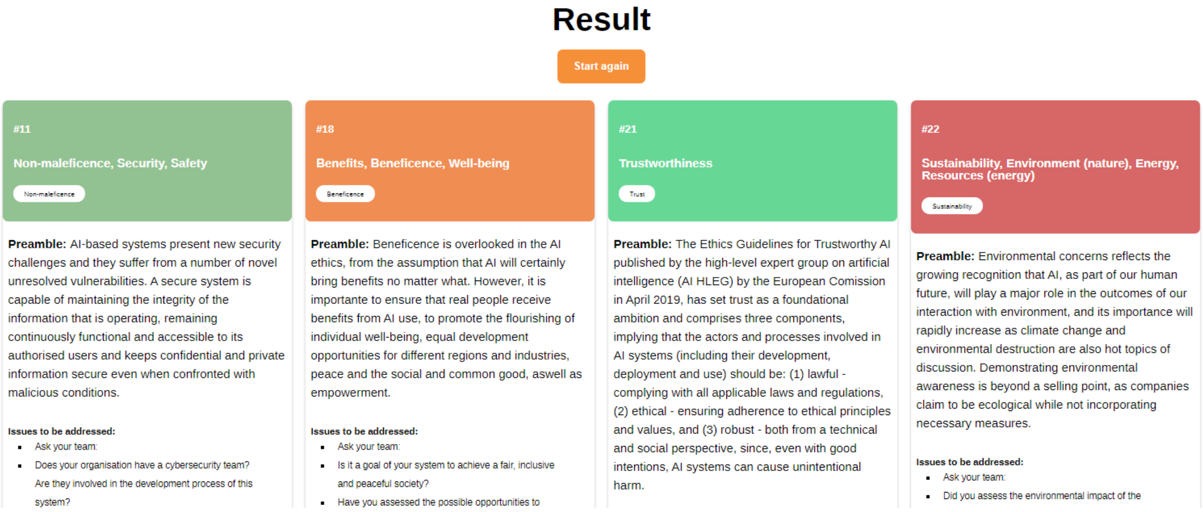
Tool Suggestion:

TransparentAI

Figure 4.7: Cards filtered through the Responsibility principle. Own source

sharing and adaptation of the guide preserving the attribution of the credit to the authors [144]. In addition, a free font is used throughout the system – UnB Office, allowing greater portability and maintainability, while users (i.e., developers, ethicists, public organizations, academics) of the system should not have to worry about patents or copyright licenses for fonts, or any other aspect of the system.

The source code of the guide is available at https://github.com/josesiqueira/RE4AIEthicalGuide and the system at https://josesiqueira.github.io/RE4AIEthicalGuide/. In the guide 24 cards are provided, distributed along the 11 principles adopted for the elaboration of the



Result

#11

Non-maleficence, Security, Safety

Preamble: AI-based systems present new security challenges and they suffer from a number of novel unresolved vulnerabilities. A secure system is capable of maintaining the integrity of the information that is operating, remaining continuously functional and accessible to its authorised users and keeps confidential and private information secure even when confronted with malicious conditions.

Issues to be addressed:

- Ask your team:
- Does your organisation have a cybersecurity team? Are they involved in the development process of this system?

#18

Benefits, Beneficence, Well-being

Preamble: Beneficence is overlooked in the AI ethics, from the assumption that AI will certainly bring benefits no matter what. However, it is importante to ensure that real people receive benefits from AI use, to promote the flourishing of individual well-being, equal development opportunities for different regions and industries, peace and the social and common good, aswell as empowerment.

Issues to be addressed:

- Ask your team:
- Is it a goal of your system to achieve a fair, inclusive and peaceful society?
- Have you assessed the possible opportunities to

#21

Trustworthiness

Preamble: The Ethics Guidelines for Trustworthy AI published by the high-level expert group on artificial intelligence (AI HLEG) by the European Comission in April 2019, has set trust as a foundational ambition and comprises three components, implying that the actors and processes involved in AI systems (including their development, deployment and use) should be: (1) lawful - complying with all applicable laws and regulations, (2) ethical - ensuring adherence to ethical principles and values, and (3) robust - both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm.

#22

Sustainability, Environment (nature), Energy, Resources (energy)

Preamble: Environmental concerns reflects the growing recognition that AI, as part of our human future, will play a major role in the outcomes of our interaction with environment, and its importance will rapidly increase as climate change and environmental destruction are also hot topics of discussion. Demonstrating environmental awareness is beyond a selling point, as companies claim to be ecological while not incorporating necessary measures.

Issues to be addressed:

- Ask your team:
- Did you assess the environmental impact of the

Figure 4.8: Different cards compared. Own source

guide, plus 2 additional cards: Stakeholders' assessment, and Overall ethical evaluation, both under the topic of Assessment. Thus, in total there are 26 cards. In the sprint backlog meeting, the actors must choose the cards that will be used in that sprint, read aloud the content of the card, then the development team will elicit the ethical requirements in the form of user stories, also writing down the reasoning that led them to those user stories. Validation should be done by development teams together with customers and multiple stakeholders, who may request changes.

The prototype designed – the Guide to Eliciting Ethical Requirements for AI –, and the ECCOLA method developed by Vakkuri et al. [6], differ in many aspects. While the latter is presented only as a deck of cards in Portable Document Format, our guide is developed as a web-based system (using HTML, CSS and JS), allowing interactivity in card selection through filters and comparisons between multiple cards, as well as extensive supporting material (how to use, principles, tools, trade-offs) and the addition of tool suggestion in the content of the cards. We also assigned free licenses and made the source code available, in order to allow the study of the tool and future adaptations to new contexts. Moreover, we contemplated in our guide all the 11 principles listed by Ryan and Stahl [10] and presented 26 cards, while in the ECCOLA method are contemplated only 7 principles with a total of 21 cards. In Table 4.6 we point out the main differences between ECCOLA and the Guide for Artificial Intelligence Ethical Requirements Elicitation.

Table 4.6: Main differences between ECCOLA and the Guide for Artificial Intelligence Ethical Requirements Elicitation

ECCOLA	RE4AI Ethical Guide
Static	Interactive
7 principles	11 principles

21 cards	26 cards
Does not suggest any tools	Suggested tools available
Copyrighted	Open license with source code available
No support material available	Support material available

4.4 Running example

This section will present an example of using the Guide as a short tutorial on how to use the Guide based on a hypothetical scenario. A medical institution wants to deploy a Facial Recognition Technology (FRT) as an AI-based system to monitor medical and behavioral conditions, as well as to diagnose medical and genetic conditions. Nevertheless, FRT in health care, like in other domains, raises ethical questions about “privacy and data protection, potential bias in the data or analysis, and potential negative implications for the therapeutic alliance in patient-clinician relationships” [145]. Given this scenario, the development team along with the Product Owner starts the use of the RE4AI Ethical Guide. As stated in the guide’s introduction, the user must start with card #0, then select any card, and can: use the Compare option, selecting specific cards to display them in order to compare them with others - from different principles or not; use the Filter option, displaying all cards from a specific principle. Moreover, reasoning done to elicit ethical requirements should be documented. Users are encouraged to answer the questions presented in cards linearly, however, it is not mandatory. In this sense, card #0 is the first to be addressed. In this example, ethical implications in FRT by Martinez-Martin [145] are explored.

Card #0: Stakeholders’ assessment

The different stakeholders identified are: the patient’s health insurance, the patient, the patient’s family, the doctors and nurses, the health care organization. These have contractual interests in providing the patient with the best treatment available. The system provides health staff with information regarding patient’s medical conditions.

Card #16: Privacy, Personal or private information. Ethical requirements elicited:

- The system should store data about patient’s complete facial image or as facial template.
- The system should collect personal data, as facial template is considered biometric data (personally identifiable information).

- Personal data should be used to detect genetic disorders, predict health characteristics such as longevity and aging, to predict behavior, pain, and emotions, also for identification and monitoring.
- Patients should be clearly informed about personal data collection and the organization's use of FRT, and must be able to consent and revoke access to their personal or private information at any given time.
- The system should report incidental findings to patients.
- The system should encrypt and anonymize personal data.
- It should be possible to detect anomalies in private data.

There is no minimum number of cards required in each Sprint, nor is there an order to be followed. Now, the team is free to choose any card from the Guide.

In the next sprint, the team can start by visiting the last card # 25: Overall ethical evaluation, in order to assess the system being designed, or, choose any card. It is suggested to reassess the system after each sprint iteration, so it is possible to mitigate possible ethical implications overlooked and improve others. In Section 5.8 another example of use of the Guide is provided.

4.5 Chapter Summary

In this chapter we have presented the conduction of stages 2 – Suggestion – and 3 – Development – of Design Science Research, in which we have provided a pilot project – the implementation of the ECCOLA method – and a prototype – the Guide for Artificial Intelligence Ethical Requirements Elicitation. In addition, we present the guide definition, the conceptual proposal, its criteria, the technologies used in the development, the reasons for the choices of principles and tools used in the guide, the definition of the content of the cards and the logo, and, finally, we make available the source code with open license for future modifications by anyone interested. Furthermore, a running example of the Guide is provided. In Chapter 5 the conduction of the evaluation of the Guide will be presented.

Chapter 5

Evaluation of the Guide for Artificial Intelligence Ethical Requirements Elicitation

In this Chapter will be presented the phase 4 – Evaluation – of the Design Science Research, the methodology adopted for the development of this work. After the creation of the Guide presented in Section 4.3, its evaluation is necessary, so that we can identify the perceptions of users when using it. It will be performed a mixed method evaluation, in a first stage through a survey, and in a second stage through a focus group. The methodologies employed for the evaluations of the RE4AI Ethical Guide, their planning, execution, and analysis will be presented. The evaluation of the proposed Guide is in accordance with our objectives pointed in Section 1.3.

5.1 Evaluation with a Survey

In a first stage, an evaluation was conducted, in order to obtain an initial perception, through a survey consisting of a questionnaire to be answered by the participants – undergraduate and postgraduate students. A survey is "a comprehensive system for collecting information to describe, compare or explain knowledge, attitudes and behaviour" [146]. For the planning of the survey, its execution and analysis of the responses, the guideline proposed by Pfleeger and Kitchenham [146] and the phases proposed by Molleri et al. [147] were used. In preparing the questionnaire, the guidelines presented by Kitchenham and Pfleeger [8] were used. The online questionnaire was developed using the free software LimeSurvey. Figure 5.1 presents the steps adapted for this research.

We started in the planning phase with the following steps: defining the objectives and questions of our evaluation – we created and defined the objectives for obtaining feedback

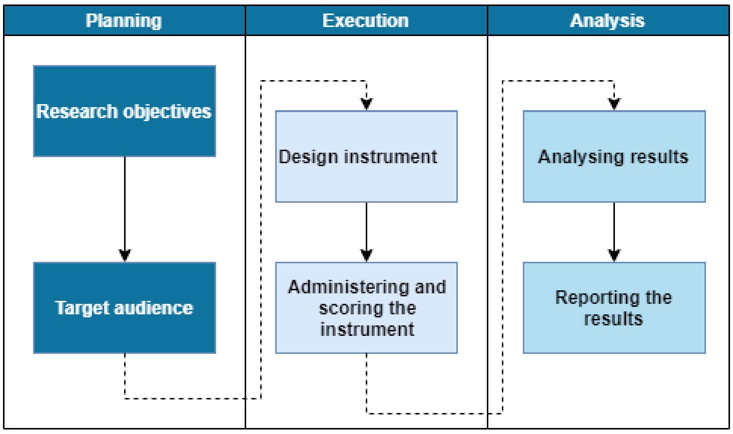


Figure 5.1: Steps for the evaluation of the Guide through the Survey. Adapted from [7] e [8].

from the use of our Guide, and the questions to be answered by the participants; choosing participants – undergraduate and graduate students. Then, in the execution phase: we designed a questionnaire – we created questions of different types with Likert-scale and open-ended questions; and administered the questionnaires. Finally, in the analysis phase: we analyzed and reported the results – presenting the qualitative analysis through the Krippendorff [148] content based analysis technique, and the quantitative analysis.

5.2 Planning the Survey

5.2.1 Survey Objective and Questions

The objective of the evaluation of the Guide through the survey is to verify the viability of the guide, as well as the perceptions of users about the content provided. To this end, undergraduate and postgraduate students were selected and invited to use the guide and answer a questionnaire, remotely. According to Morley et al. [132], there is little evidence that the use of tools that operationalise AI ethics impacts the governability of a system. Thus, the overall aim of the evaluation of the Guide is to provide evidence that its use may have an impact on the governability of AI-based systems. After using the guide, the participants answered a questionnaire with the purpose of knowing their opinions regarding the RE4AI Ethical Guide. In addition to questions related to the respondents' characteristics, they answered the following questions:

- Q1: Regarding the supporting content present in the guide, was the information sufficient for its understanding and use?
- Q2: Did you already know any of the tools suggested by the Guide?

- Q3: Regarding the suggested tools, do you believe they have utility in implementing AI ethics?
- Q4: Which Principles do you consider most easily implementable?
- Q5: Did you find the questions in the Guide cards easy to understand?
- Q6: In relation to the questions present in the Guide cards, can the questions answered by the use of the cards help to elicit ethical requirements?
- Q7: Has the Guide improved your ethical awareness and learning?
- Q8: At which stage of the software development process do you consider it most feasible to use the Guide?
- Q9: Would you use the RE4AI Ethical Guide in requirements elicitation?
- Q10: Do you have any suggestions for improving the Guide?

5.2.2 Survey Participants

The participants are undergraduate and graduate students of the disciplines of Data Science, Requirements Analysis, Artificial Intelligence and Cognitive Computing, and Artificial Intelligence of the semester of 2021.1. The disciplines were lectured at the University Center of Brasilia (UniCeub), in Brasilia-DF, Brazil. These disciplines are part of the curriculum of the courses of Computer Science, Computer Engineering and Systems Analysis. There was an average of 20 students per class.

5.3 Survey Execution

For the construction of the questionnaire the LimeSurvey tool, a free software for the creation of questionnaires and surveys, was used. The following considerations of Kitchenham and Pfleeger [8] were taken into account for the questionnaire design:

1. The way a question is worded – maintaining an appropriate technical level;
2. The number of questions in the questionnaire – adequate number of questions to neither discourage nor overwhelm the respondents;
3. The interval and type of response categories – use of Likert-scale type questions and open-ended questions;
4. Instructions to respondents – participants can understand and answer the questionnaires themselves through the instructions.

The questionnaire was divided into two groups of questions: a) Demographic questions on participants characteristics; b) Questions on the Guide evaluation. The language chosen for the elaboration of the questionnaire was Portuguese, due to the nationality of the participants. As an output of this stage we obtained a questionnaire, available on the link http://survey.josesiqueira.com/index.php?r=survey/index&sid=149126&lang=en, and presented in Appendix C. The topic was presented to the students in the virtual classroom environment of their disciplines, then the questionnaires were administered at the end of the classes, where the students answered the questionnaires voluntarily and on their own.

5.4 Survey Results

A total of 40 complete and 49 incomplete responses were received, amounting to 89 responses. Only complete questionnaires will be considered, therefore 40 complete questionnaires were analysed. In order to find out the characteristics of the questionnaire participants, the participants were asked to answer a first group of questions, related to their profile. Regarding Age Group, most of the respondents, 24 (60%) are between 18 to 24 years old, 13 (32.5%) between 25 to 35 years old, 2 (5.5%) between 36 to 50 years old, while only 1 (2.5%) is over 50 years old, as shown in Figure 5.2.

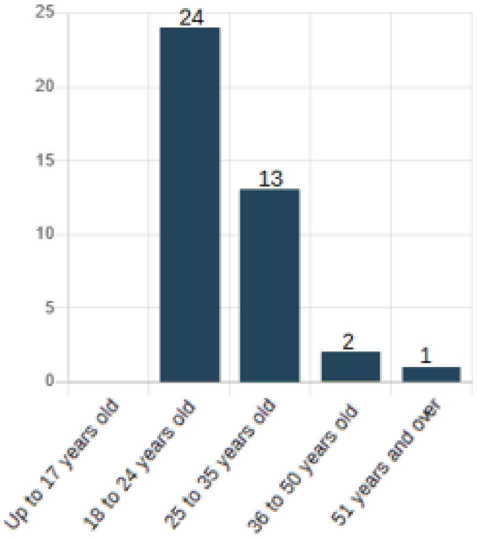


Figure 5.2: Age group of survey respondents

In relation to the education of the participants, the majority, 29 (72.5%) are attending undergraduate courses, while 6 (15%) have completed higher education, 2 (5%) are attending master's courses, and 2 (5%) informed the degree of education as, attending post-graduate studies, and incomplete higher education. Finally, only 1 (2.5%) is a doctoral student. Only 5 (12.5%) participants answered that they are currently participating in some software development project in the context of AI, while the vast majority, 35 (87.5%) are not working with the subject, as presented in Figure 5.3.

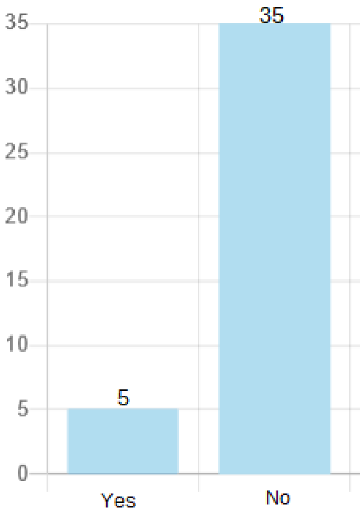


Figure 5.3: Current participation in AI projects

Regarding participation in previous software development projects in the context of AI, most, 26 (65%), responded that they have never participated in any such project, 8 (20%) responded that they have participated in 1 project, 3 (7.5%) responded that they have participated in 2 projects, while only 3 (7.5%) responded that they have had previous experience in 5 or more projects in the context of AI, as presented in Figure 5.4.

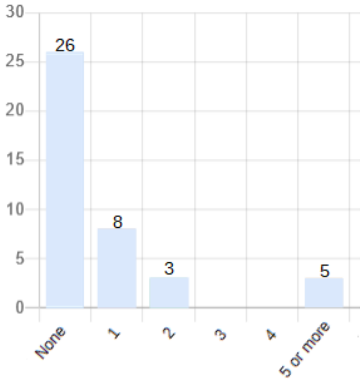


Figure 5.4: Previous participation in AI projects

The majority of respondents, 22 (55%), have no prior knowledge about ethical issues in AI, while 18 (45%) claimed to have prior knowledge about ethical issues in the context of AI, as presented in Figure 5.5.

Whilst the subjectivity of developers influences the ethical outcomes of AI-based systems, their training is crucial to address this issue. In this question, we found that only 5 (12.5%) have received some kind of training related to ethical guidelines for AI previously,

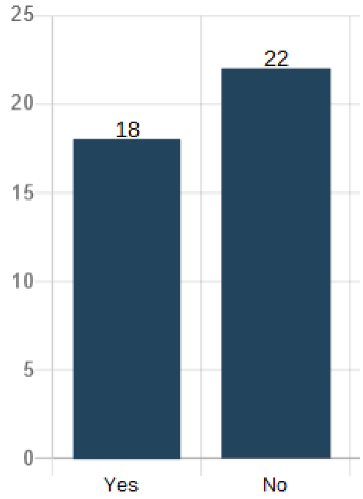


Figure 5.5: Prior knowledge about AI ethical issues

while the vast majority, 35 (87.5%) have never received any kind of training to address this issue, as presented in Figure 5.6.

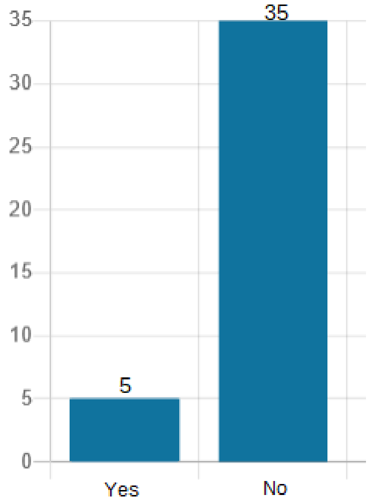


Figure 5.6: Previous training on ethical guidelines in the context of AI

In the second group of questions, only questions related to the Guide were presented, and participants were required, before answering the questions, to access the RE4AI Ethical Guide – and to navigate, on their own, through the features and functionalities available in the proposed guide.

Regarding the participants' opinion on the support material provided in the Guide, i.e., the introduction, principles, tools, and trade-offs sections, in relation to their understanding and usefulness, the majority of the participants, 24 (60%), agreed that the information presented in the Guide was sufficient for its understanding and use; 9 (22.5%) strongly agreed; and 7 (17.5%) neither agreed nor disagreed, as presented in Figure 5.7.

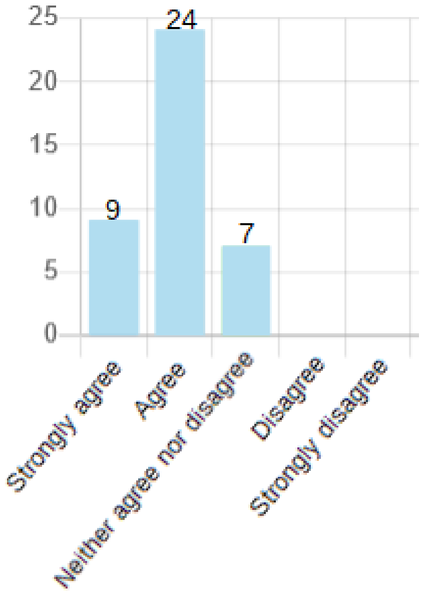


Figure 5.7: Feedback on the supporting content of the Guide

Regarding the previous knowledge of the tools suggested by the Guide, most of them are unknown to the participants, 36 (90%) said they did not know any of the tools presented in the guide. Regarding the applicability of the tools, 11 participants (27.5%) strongly agreed about the applicability of the tools suggested by the Guide in the implementation of ethics in AI, 19 (47.5%) agreed, 8 (20%) neither agreed nor disagreed. Only 2 (5.5%) disagreed with its applicability. With regard to the perception about the applicability of the ethical principles in AI provided by the guide, 30 (75%) participants chose the principle of Transparency. This choice reinforces the idea that this is the principle that enables the other principles [13]. In second place, participants chose the principle of Responsibilitity (with 23) and in third place the principle of Privacy (with 21), as presented in Figure 5.8.

For the success in the elicitation of ethical requirements, it is imperative to understand the questions that must be answered by the development team. Regarding the ease of understanding of the questions available in the cards, 12 (30%) participants strongly agreed, 19 (47.5%) agreed, while 9 (22.5%) were neutral, as presented in Figure 5.9.

9 participants (22.5%) strongly agreed regarding the usefulness of the answers obtained through the questions on the cards in creating user stories, 21 (52.5%) agreed, 9 (22.25%) were neutral , while only 1 (2.5%) disagreed, as presented in Figure 5.10.

The final objective of the proposed Guide is to assist in the creation of user stories. Some participants commented positively on this assistance: “The guide cards help a lot in the elaboration of clear user stories, placing emphasis mainly on what the software should do, for example making it explicit that there should be no discrimination of softwares

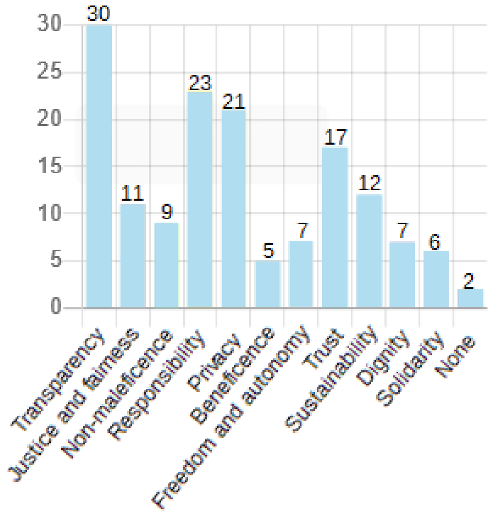


Figure 5.8: Applicability of ethical principles in AI

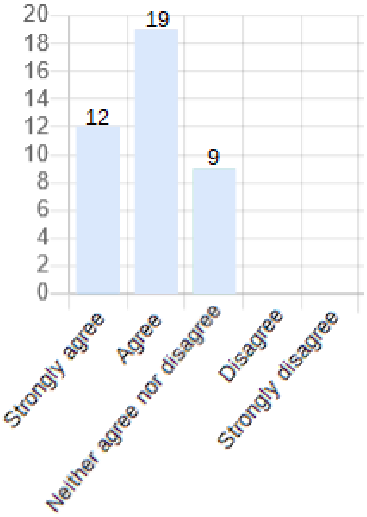


Figure 5.9: Comprehension of the questions available on the cardss

users."

In addition to producing ethical requirements, the Guide is intended to assist in increasing the ethical awareness of its users. 11 (27.5%) of the participants strongly agreed that the Guide can increase ethical awareness, 16 (40%) said they agreed, 11 (27.5%) remained neutral, 1 (2.5%) disagreed, and 1 (2.5%) strongly disagreed, as presented in Figure 5.11.

Several participants stated that the Guide was helpful for learning and noted a positive experience regarding learning about the topic of AI ethics: "In the same way that it helps the team to keep the project on track, the guide is also helpful for learning with its consistent presentation of information and context". Some participants who had some superficial or no contact with issues related to AI ethics stated: "The guide helped me

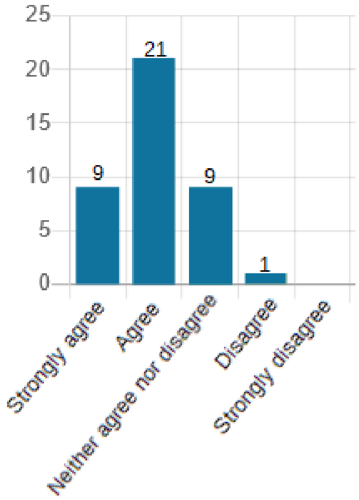


Figure 5.10: Feasibility of eliciting requirements through the answers obtained

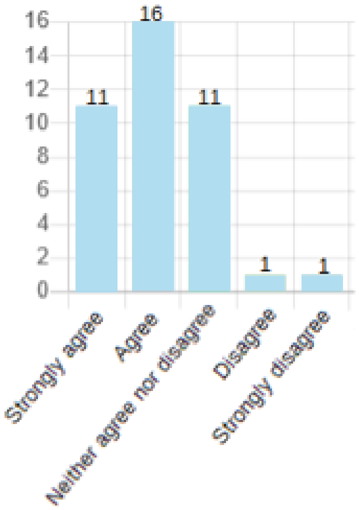


Figure 5.11: Ethical awareness acquired through the use of the Guide

understand the importance of ethics in AI software projects and I had no idea about the principles that need to be taken into consideration when building software in this context"; "The guide opened my mind about AI ethics. This subject needs to be increasingly thought and discussed by the software development community, due to the evolution of systems that use some AI component and its direct and indirect impacts on end users social well-being".

On the other hand, some participants stated that the proposed Guide is too extensive and commented on the practicality: "I learned a lot of new concepts from reading the guide. Although I miss something briefer to help in practice, like a checklist for the day-to-day, or a template for documentation related to ethical issues."

Regarding the applicability of the Guide in relation to the software development process phases, 33 (82.5%) of the participants consider it to be applicable in the Requirements Analysis phase, 22 (55%) in the Design phase, 5 (12.5%) in the Coding phase, 6 (15.5%) during the Testing phase, finally, 7 (17.5%) stated that it is applicable in the Implementation and Maintenance phase, as presented in Figure 5.12.

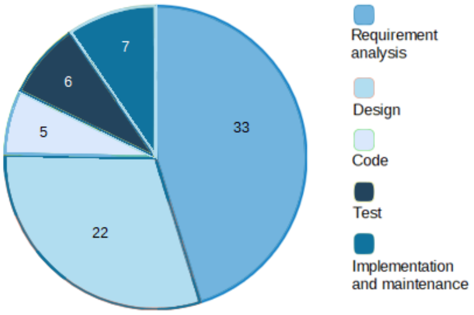


Figure 5.12: Applicability of the Guide in relation to the software development phases

About the future use of the Guide during the requirements elicitation phase of an AI-based system, 9 (22.5%) of the participants strongly agreed to use it, 21 (52.5%) agreed, 8 (20%) were neutral, 1 (2.5%) disagreed, and 1 (2.5%) strongly disagreed, as presented in Figure 5.13.

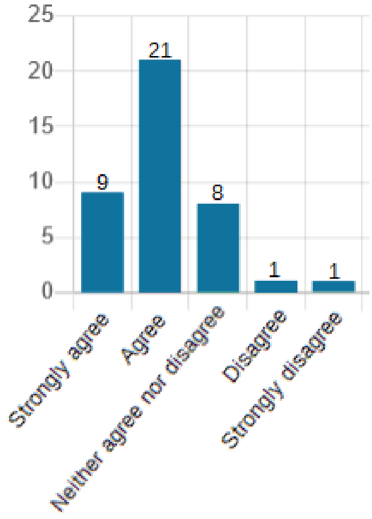


Figure 5.13: Future use of the RE4AI Ethical Guide by participants

Overall, the Guide was well accepted by the respondents, in relation to practicality. Some comments were: “I would use it because of the practicality of addressing specific requirements elicitation contexts that can be improved”; “I would use the guide because of the practicality and the usefulness of the cards. I believe that these resources would help a lot in requirements elicitation”. Regarding commonly overlooked aspects of requirements:

“The guide is very useful and is a good way, especially in the requirements phase, to not overlook commonly neglected aspects of building AI systems.”

Some participants made suggestions for improvement:

- Regarding the extension of the content, one finds the need for the presentation of the Guide’s content in a reduced form and adjustments in the interface: "I missed a slightly more summarized version of the guide. A version that could be used for smaller and simpler AI projects, like a project that does not necessarily have a team of developers, but for example 1 or 2 people just working on an AI model that is going to be used or consumed by some other process.”
- In relation to how to use the guide, impacting its use process: “It would be interesting to have a prioritization of the principles, something that would give an order of importance. What should be treated or resolved with more priority, for cases where the software development company does not have enough resources or time to evaluate all the ethical principles proposed by the guide.”
- We also observed the need to make the Guide available in other languages, in order to facilitate understanding by users who are not proficient in English, or are not native speakers: "It would be interesting to make the content of the guide also available in Portuguese to facilitate understanding.”; “I couldn’t find a translation on the site and the google translation often leaves something to be desired, it would be nice to implement in other languages.”

5.5 Focus Group Evaluation

In the second stage, through a focus group, the moderator presented the RE4AI Ethical Guide and its contextualization to the participants – experienced AI professionals. The moderator conducted the participants in the use of the Guide, allowed interaction between the participants when using the Guide and eliciting requirements, and after this phase addressed a script of questions. Focus group is a quick and low-cost method used to obtain information related to the experiences of professionals and users of some technology/product/service, providing qualitative information [7]. Focus groups are “carefully planned discussions designed to elicit the perceptions of group members about a defined area of interest” [7]. As data collection methods, we used the video recording of the focus group session, and the question script. We used the RNP Web Conference and the free software OBS Studio, for conducting and recording the remote session of the focus group, respectively.

We used adaptations of Kotio's guideline [7] to conduct the focus group. The steps for evaluating the Guide through the focus group are presented in Figure 5.14, adapted for our case, and arranged into three phases: planning, execution and analysis.

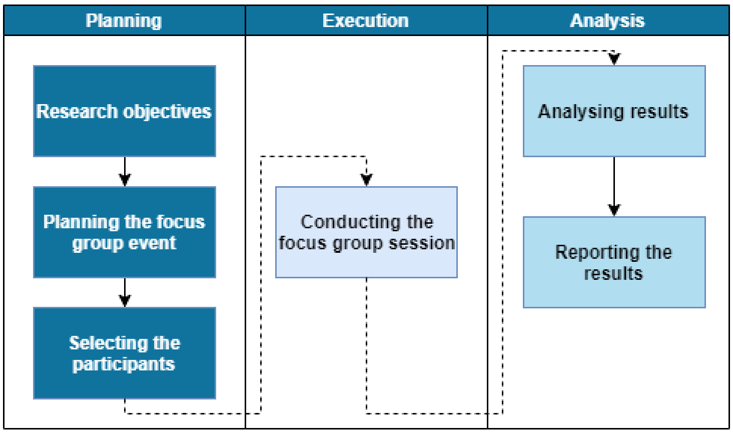

Planning
Execution
Analysis
Research objectives
Planning the focus group event
Selecting the participants
Conducting the focus group session
Analysing results
Reporting the results

Figure 5.14: Steps for the evaluation of the Guide through the Focus Group. Adapted from [7] e [8]

We began in the planning phase with the steps of: defining the objectives and research questions of our evaluation – creating and defining the objectives of the feedback obtained from the use of our Guide, and the script of questions to be answered by the participants; planning the focus group event – setting a pre-determined structure and sufficient time for participants to understand the issues and meaningful discussions to occur [7]; selecting participants – choosing representative, experienced, and motivated participants. Then, in the execution phase: conducting the focus group session – presenting an introduction of the objectives and fostering discussion and interaction, and video recording of the session was made. In the analysis phase: analysing and reporting the results, and finally, the final discussion.

5.6 Focus Group Planning

5.6.1 Objective and question script

The objective of the evaluation of the Guide through the Focus Group is the same as the survey presented in Section 5.2.1. For this purpose, participants working in software development teams that implement AI-based systems were selected. They were invited to use the Guide and to answer a question script composed of the same questions present in the survey, with the addition of one question: Do you think the Guide helped the software development team to identify and elicit the ethical requirements?

5.6.2 Planning the Focus Group Event

It is crucial to have well-planned focus group session agendas, especially in situations where the meetings are conducted electronically [7], as is our case. The focus group was designed so that it could be conducted remotely. There was the presence of a moderator who described the problem and objectives, as well as presenting the RE4AI Ethical Guide. The session lasted one hour and thirty minutes. The moderator was careful to create an informal atmosphere where participants felt unhindered, as well as [149]:

1. Not judge or criticise ideas during the session;
2. Encourage thoughtless or seemingly irrelevant ideas;
3. Build on the ideas of others;
4. Strive for quantity.

The session was divided into three parts: a) introductory presentation – contextualization, description of the problem, the objectives; b) presentation and use of the guide – orienting the participants in the use of the guide; c) presentation of the questions from the question script – the moderator read aloud the questions and the participants were encouraged to answer them aloud.

5.6.3 Focus Group Participants

The focus group was composed of 5 professionals who develop AI-based systems. The working area of these professionals are: major Brazilian banks (private and public), private companies that provide services to public organizations, and major multinational IT companies.

5.7 Conduction of the Focus Group Session

The session began with an overview of the study objectives and a discussion of how participants should act during the session [7]. Participants were encouraged to collaborate with each other, and the organiser ensured the confidentiality and anonymity of the discussion.

We chose the Product and System User Testing technique [149] for the focus group session. This is because of the technique's main advantages: feedback is based on participants' experience of performing real tasks; providing stimulus for discussion; useful when evaluating systems or prototypes under development. Observations (comments made throughout the session) and a question script were used to provide additional information.

The session was held through the platform Web Conference of the Brazilian National Education and Research Network (RNP). The moderator made available the link of the Guide, and shared the screen, in which the participants experienced the assisted use of the guide, exploring the ethical requirements elicitation, and collaborated with each other through this task. Next, the questions in the questionnaire were read aloud, and participants were able to answer the questions, allowing for a discussion. Finally, participants were asked to present their final considerations.

5.8 Focus Group Analysis

Regarding the use of the Guide, the participants created a hypothetical scenario of an AI-based essay correction system, to be employed by a selection board for admission to an institution. The system aims to select candidates by assigning a score to their essays. Throughout the session, they addressed three cards, and answered questions in the Issues to be addressed section. As detailed in the Guide information, they should start with card 0, on stakeholder assessment. After that, they chose among themselves the next cards to be addressed – cards 7 and 6 – where various ethical requirements were elaborated by the participants.

Card #0: Stakeholder's Assessment. Analysis made by the group:

The different stakeholders identified are: the selection board, the candidates, and the institution. These have contractual interests in selecting the best candidate, given the available criteria. "The system is the one that decides the selection, it can select the candidates automatically, but if it has any bias, it does so in a wrong way, negatively impacting the stakeholders." Collaboration among multiple stakeholders was not identified.

Card #7: Interpretability, Showing - Transparency. Ethical requirements elicited:

1. The algorithm should be known to the candidates.
2. The dataset should be disclosed after the application of the test. This can help students to prepare for the next exams, since the theme of the essay is modified in each selection process.
3. It should be communicated widely that the correction is being done automatically by a system, in the public notice, in the form of a contract.

4. Regarding compliance with the General Law on Data Protection (LGPD): There should be no personal data within the dataset, and users' data should be anonymised.

Card #6: Explainability, Explicability, Disclosure - Transparency. Ethical requirements elicited:

1. The system should not retain personal data.
2. The system should ensure that a candidate receives their own score, not someone else's.
3. Data used for training must be anonymous.
4. Reproducibility of the system should be allowed by the Ministry of Education.
5. Documented system code should be disclosed to external auditing bodies.
6. The system should be periodically monitored, with a portion of the training data, and made publicly available – create these monitors and publicly disclose the results.

Regarding testing. Ethical requirements elicited:

1. Professors on the board should be able to perform curation by evaluating an essay and observing the results of the AI-based system, thus enhancing the system – Enable a curation process.
2. Record how and who performed the curation process, make reports of that process (making logs of the curation).
3. It should be possible to explain and document the code and metrics involved.
4. Make available non-technical documentation of the system part – make metrics public.
5. Tests that fail (e.g. essays that receive unduly low scores) should be exposed to the public.
6. Appropriate metrics should be used in order to publicly demonstrate the percentage of successes, as well as which essays were wrongly graded or scored.
7. It should be checked in which cases the metric was not satisfied.
8. The cases where the AI-based system is not able to repair should be identified, and then a curation process should be undertaken – a follow up by a certified essay proofreader.

Regarding the tools presented, the participants explored the tool **InterpretML** on GitHub. It was identified that the tool helps to explain how the model works, as well as why the system acts in a certain way. The advantages found are: “finding out if the model being developed is doing what we want, if the dataset is sufficient, or if counter example data is needed”, “Helps to find out how the model works and how to improve it, as well as finding out the reasons, of the model’s classification features, even when using a black box model”.

Regarding the question script addressed, they assessed that the principles present are not so basic: “propose questions that you would never ask yourself nor raise the possibility of the problem, and each card brings discussions of hours, it’s very complex, and it all becomes easy to visualize, in items”. They have never had contact with the suggested tools, but believe that the tools have application to ethics in AI. 3 participants considered that the principle of Transparency is easier to apply (in particular, card #5). While Dignity is the most difficult, because “not all systems reach that point”. Only 2 participants mentioned that no principle is easily implementable, because "all of them are very complex, one should be careful not to see them as simple".

Concerning the questions presented in the cards, they mentioned in relation to their practicality: “very practical, direct and objective”. However, they also mentioned that they can be very broad: “the questions are clear, but the answers are not so clear.” They pointed out the help of the questions for the elicitation of ethical requirements: “When provoked, we remember curating process, public notice, among others.”

The participants, acting as a software development team, reported that the Guide helped to identify and elicit ethical requirements: “The guide helped us a lot to elicit ethical requirements, because the Guide is well structured, well divided and in a simple way.” Also, a Product Owner present would help to have more questions to be raised. Overall, the Guide has improved the ethical awareness and learning of the participants: “By reading all the cards, we see principles that we were not aware of.”

4 participants considered the requirements analysis phase as the most suitable phase of the software development process for using the Guide, while only 1 considered the implementation and maintenance phase. All participants stated that they would use the RE4AI Ethical Guide in requirements elicitation in their projects.

Several suggestions for improvement were offered. The Guide was considered to be too lengthy, and they suggested reducing the scope of the Guide and providing a "reduced version". Due to the high coverage and complexity of the various questions, they suggested more guidance on the use of the Guide: “guidance on which cards and tools to use for specific problems”. This is also due to the lack of availability of a Product Owner, who usually works in areas outside IT, and: ”would not have enough time to use this guide”.

In addition, they suggested a separation of the Guide into categories: "possible phase divisions, such as documentation, testing, coding and maintenance".

Finally, they considered the scope of the Guide of great interest, where "just the content already provides considerable knowledge" and "Fundamental to bring the debate of this problem, which is already a current problem that impacts everyone". They pointed out that there are other ways of benefiting from the content: "not just as Planning Poker, but the reading itself already raises the debate". They mentioned the possible application of the Guide acting as a checklist: "to assess and grade an already implemented system."

5.9 Main evaluation discussions

In our two studies, we identified 6 perceived positive points, such as: a) the support information presented is adequate for understanding and use; b) the questions contained in the cards are easy to understand – objective and clear; c) the use of the Guide helps the creation of user stories through the questions in the cards; d) there is an increase in ethical awareness through the use of the Guide; e) applicability of the Guide in the requirements elicitation phase; f) there is an interest from the participants in using the guide in the requirements elicitation phase in their future projects. In addition, from the focus group we noted the usefulness and practicality of the Guide in eliciting requirements in development teams, since the participants were able to elicit 18 requirements for a hypothetical scenario, during the session.

Our findings suggest that the RE4AI Ethical Guide is perceived to be of great interest by participants, receiving an overall positive evaluation. The Guide, by operationalising ethical principles, can help mitigate challenges present in the literature, such as: lack of tools to implement AI ethics at the project level [13], [14]; lack of tools that assist software development teams as a whole [6]; with practicality and usability offering help to be used in practice [13]; as well as the lack of tools that do not focus mostly on explicability [13].

We observed 5 negative points and suggestions for improvement offered by the participants, such as: a) the suggested tools are not known by the participants; b) very extensive and broad guide, suggesting a reduced version (reduction of the scope) with cards and tools oriented to a particular context/problem; c) to divide the Guide in categories for the phases of documentation, tests, codification and maintenance; d) make the Guide available in other languages; e) offer an order of importance of the principles (prioritization of the principles). These problems will be mitigated in future works and in further versions of the Guide.

5.10 Limitations and threats to validity

There is an impossibility to generalize the result of the survey, in part because it was conducted only in Brazil, with undergraduate and graduate students, who may have their perceptions impacted, both by their previous experiences and by the quality of the educational institution. Also, there was a difficulty in obtaining a significant number of participants. However, these limitations indicate opportunities to replicate this study in different countries and contexts. Furthermore, it was detected that more than half of the students, 26 (65%), had never participated in a project of this type. In order to mitigate this problem, a focus group was conducted with AI professionals.

Concerning the generalization of the focus group results, like other qualitative studies, they have usual limitations on this topic [149]. We had available only 5 professionals who develop AI-based systems in different contexts, which influenced the generalization of the final results, increased by the fact that their answers are affected by the organization's goals and previous experiences on the subject. However, despite the fact that generalization is not possible, these data are valid and complementary with other studies.

The evaluation did not consider implemented AI-based systems where it would be possible to test the functionalities and trace them to their requirements, i.e. a validation step of the requirement elicited through our Guide was not contemplated. This proved to be unfeasible in the context of the development of this study, since the developers – the participants – would not have the time to develop complete systems. Therefore, one of the limitations of the Guide is the impossibility to ensure that the requirements elicited from its use comply with AI ethics guidelines or regulations.

Furthermore, another limitation to our study is the application context of the AI-based system, i.e., a possible bias about the deployment area of the system – where the AI-based system will be deployed (e.g., banking, medicine, surveillance, business, transportation), working in one but not in others. This limitation stems from the finding of Morley et al. [132]: "The ethical implications of deploying an AI-based system in a healthcare context are unlikely to be the same as the ethical implications of deploying an AI-based system in an educational context." In order to mitigate this evaluation problem, the evaluation was conducted with participants embedded in different contexts, such as students, and professionals working in banking, public sector and representative private sector companies providing IT services. Similarly, it is possible that our Guide has application limitations regarding its usefulness for the development of an ethical AI-based system using a specific AI technique (e.g., Machine Learning, Decision Trees, or Deep learning).

There is a need for further evaluation of the proposed Guide in order to compare it with other Software Engineering methods for ethical requirements elicitation in the context of AI-based systems.

5.11 Chapter Summary

In this chapter we conducted the evaluation of the Guide for Artificial Intelligence Ethical Requirements Elicitation, comprising phase 4 – Evaluation – of the adopted Design Science Research methodology. To characterize the viability of the RE4AI Ethical Guide, a mixed study was conducted through a survey with 40 undergraduate and graduate students who evaluated the Guide through an online questionnaire, as well as a focus group with 5 experts in the field from different contexts. In Chapter 6, we present our final considerations, contributions, and indicate future work.

Chapter 6

Final Remarks

In this work a Guide for Artificial Intelligence Ethical Requirements Elicitation was developed, also referred to as RE4AI Ethical Guide, to assist the operationalisation of ethics in AI by software development teams. For its creation, the Design Science Research methodology was adopted in order to understand the problem, suggest a pilot project, develop a prototype and evaluate it.

For the first stage, a Systematic Literature Review was performed with the selection of 33 primary studies, in which few of them explicitly address the use or present new proposals for practical means to implement AI ethics, demonstrating how ethics in practical AI is still in its early stages, especially regarding practical guidelines, ethical requirements, and tools. After analyzing the techniques, tools, methods, frameworks and processes found in the literature, we identified the ECCOLA method [6] as the most suitable for our context, consisting of a deck of cards, based on Planning Poker, for the elicitation of ethical requirements in AI, made available in a static way. We also found the need for the inclusion of traditional software engineering practices, such as requirements elicitation, for the context of Artificial Intelligence, in addition to the characteristics of a Guide to implement ethics in AI [66]: broad, operationalizable, flexible, iterative, guided and participatory.

In the second stage, a pilot project was created, serving as an implementation of the ECCOLA method, with the original 21 cards, covering 7 principles, proposed by the authors of ECCOLA [6]. In the third stage, the pilot project was further developed and a prototype – RE4AI Ethical Guide – was conceived, composed of 26 cards, comprising the 11 principles found in the SLR. The Guide was developed as a web-based system allowing interactivity in the selection of the cards through filters and comparisons among multiple cards, as well as a support material. Finally, in the fourth stage, evaluation, a mixed study was conducted through a survey with 40 undergraduate and graduate students who evaluated the Guide through an online questionnaire. In addition, we also conducted a focus group with 5 experts in the field of AI software development. The RE4AI Ethical Guide received a positive evaluation in the assessments conducted.

The implementation of ethics in AI is an ongoing challenge, which should not be seen as a final goal that can be objectively achieved, as a checklist to be completed, but as a development process, i.e., a set of repeatable procedures, and re-evaluated on a recurring basis. Among other features, we have tried to develop the Guide so that it is perceived by users as a reflective development process, which helps AI practitioners to understand their own subjectivity and biases within a given set of circumstances. Therefore, a more diverse and interdisciplinary development team would be fruitful to AI ethics in practice.

We hope to contribute in the development of future research in the context of AI ethics, both in academia and industry, and in choosing tools and processes that support the implementation of ethics in AI-based systems, as well as raising awareness about the various ethical issues involved in the use of AI-based systems and their challenges in the development process. The main contribution of this work is the presentation of the Guide for Artificial Intelligence Ethical Requirements Elicitation – RE4AI Ethical Guide.

6.1 Future work

Throughout this study, several research possibilities were identified that address ethics in AI sharing an interest in assisting to operationalise, i.e., put into practice in some way this topic. We identified the need for the implementation of a client side in the developed web-based system, in which software development teams can create and access an account, modify and insert cards through a graphical interface, and that users can store the elicited requirements related to a particular card, i.e., to the ethical principle. These requirements can serve as examples to new users of the guide. In addition, other ethical requirements may arise. In this way, we suggest the creation of a dataset of AI ethical requirements and the use of a Natural Language Processing technique, where it would be possible to train these data and generate a Machine Learning model so that new requirements can be automatically validated.

Some possible future work are:

- Further examples from the use of RE4AI Ethical Guide are needed, as well as from the available tools, processes, frameworks, and methods, and in different contexts;
- To conduct further evaluations of the guide to identify the perceptions of a diverse set of AI practitioners in the use of the guide's tools, processes, frameworks and methods and propose improvements;

- Further improve the Guide based on evaluations performed, presenting enhanced versions, i.e., more iterations on the Design Science Research cycle are needed;
- External ethical auditing is a key component of any ethical AI-based system [132]. We therefore propose to issue an ethical AI certificate through official public and/or authorised auditing bodies;
- The traceability of ethical requirements in the implemented code is also an attractive field of research where attention is needed, as it requires the evaluation and understanding of what has been accomplished by the developers. It is interesting to provide ways to perform this task, as well as examples of these mappings – between ethical requirements and code;
- Presentation of a catalogue or database of ethical requirements in AI;
- To evaluate the application of the proposed Guide when it is desired to evaluate AI-based systems already developed, deployed and in use by users;
- The operationalisation of ethical principles and guidelines in AI is subject to the subjectivity of those involved in the elicitation process, and more importantly, in the developers. There is a need for more work that focuses on teaching AI ethics in the training of future professionals as part of the curriculum adopted in courses related to the development of AI-based systems, in order to increase ethical awareness among students in computing courses, as well as the training of IT professionals by organizations. Furthermore, there is a need for inclusion of interdisciplinary and diverse workers on teams that develop AI-based systems, besides the inclusion of civil society and other relevant stakeholders.

Overall, our work is an important cornerstone for enabling and steering such future research through the presentation and use of the proposed Guide, among other aspects.

References

[1] Vaishnavi, Vijay K and William Kuechler: *Design science research methods and patterns: innovating information and communication technology.* Crc Press, 2015. xv, 5, 6, 7

[2] Dignum, Virginia: *Responsible Artificial Intelligence - How to Develop and Use AI in a Responsible Way.* Artificial Intelligence: Foundations, Theory, and Algorithms. Springer, https://doi.org/10.1007/978-3-030-30371-6, 2019. xv, 10, 11, 16

[3] Jameel, Tanzeela, Rukhsana Ali, and Iqra Toheed: *Ethics of artificial intelligence: Research challenges and potential solutions.* 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), pages 1–6, 2020. xv, 10, 12

[4] Gunning, David and David W Aha: *Darpa's explainable artificial intelligence program.* AI Magazine, 40(2):44–58, 2019. xv, 11, 13

[5] Kitchenham, Barbara A., Pearl Brereton, David Budgen, Mark Turner, John Bailey, and Stephen G. Linkman: *Systematic literature reviews in software engineering - A systematic literature review.* Inf. Softw. Technol., 51(1):7–15, 2009. xv, 44

[6] Vakkuri, Ville, Kai-Kristian Kemell, and Pekka Abrahamsson: *ECCOLA - a method for implementing ethically aligned AI systems.* CoRR, abs/2004.08377, 2020. `https://arxiv.org/abs/2004.08377`. xv, xvii, 2, 35, 41, 50, 55, 58, 66, 68, 69, 70, 72, 74, 75, 76, 77, 78, 79, 85, 86, 90, 109, 112

[7] Kontio, Jyrki, Laura Lehtola, and Johanna Bragge: *Using the focus group method in software engineering: Obtaining practitioner and user experiences.* In *2004 International Symposium on Empirical Software Engineering (ISESE 2004), 19-20 August 2004, Redondo Beach, CA, USA*, pages 271–280. IEEE Computer Society, 2004. `http://doi.ieeecomputersociety.org/10.1109/ISESE.2004.35`. xvi, 94, 103, 104, 105

[8] Kitchenham, Barbara A. and Shari Lawrence Pfleeger: *Principles of survey research part 4: questionnaire evaluation.* ACM SIGSOFT Softw. Eng. Notes, 27(3):20–23, 2002. `https://doi.org/10.1145/638574.638580`. xvi, 93, 94, 95, 104

[9] Eleanor Shearer, Richard Stirling and Walter Pasquarelli: *Government AI Readiness Index 2020.* Oxford Insights, 2020. `https://www.oxfordinsights.com/government-ai-readiness-index-2020`, visited on 2020-10-16. xvii, 17, 18

[10] Ryan, Mark and Bernd Carsten Stahl: *Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications.* Journal of Information, Communication and Ethics in Society, 2020. xvii, 10, 13, 20, 21, 26, 52, 63, 66, 69, 70, 76, 77, 78, 79, 80, 90

[11] Hagendorff, Thilo: *The ethics of AI ethics: An evaluation of guidelines.* Minds Mach., 30:99–120, 2020. xvii, 1, 2, 3, 14, 19, 20, 28, 30, 34, 51, 64, 76, 79, 80, 84

[12] Vogelsang, Andreas and Markus Borg: *Requirements engineering for machine learning: Perspectives from data scientists.* In *2019 IEEE 27th International Requirements Engineering Conference Workshops (REW)*, pages 245–251. IEEE, 2019. 1, 2, 35, 51, 60

[13] Morley, Jessica, Luciano Floridi, Libby Kinsey, and Anat Elhalal: *From what to how: An initial review of publicly available ai ethics tools, methods and research to translate principles into practices.* Science and Engineering Ethics, pages 1–28, 2019. 1, 2, 3, 4, 14, 21, 27, 28, 35, 40, 51, 56, 72, 99, 109

[14] Vakkuri, Ville, Kai Kristian Kemell, Joni Kultanen, and Pekka Abrahamsson: *The current state of industrial practice in artificial intelligence ethics.* IEEE Software, 2020. 1, 2, 4, 13, 14, 26, 34, 35, 51, 57, 58, 109

[15] Hagerty, Alexa and Igor Rubinov: *Global AI ethics: A review of the social impacts and ethical implications of artificial intelligence.* CoRR, abs/1907.07892(1):1–27, 2019. 1, 51, 60

[16] Rahman, Md. Abdur, M. Hossain, N. Alrajeh, and Nadra Guizani: *B5g and explainable deep learning assisted healthcare vertical at the edge: Covid-19 perspective.* IEEE Network, 34:98–105, 2020. 1

[17] Times, The New York: *Cambridge Analytica and Facebook: The Scandal and the Fallout So far*, April 2018. `https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html`. 1

[18] Reuters: *Amazon scraps secret AI recruiting tool that showed bias agaisnt women*, October 2018. `https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G`. 1

[19] Review, MIT Technology: *Predictive policing algorithms are racist. They need to be dismantled.*, July 2020. `https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/`. 1

[20] Siau, Keng and Weiyu Wang: *Artificial intelligence (ai) ethics: Ethics of ai and ethical ai.* Journal of Database Management (JDM), 31(2):74–87, 2020. 1, 2, 50, 58

[21] Tolosana, Rubén, Rubén Vera-Rodríguez, Julian Fiérrez, Aythami Morales, and Javier Ortega-Garcia: *Deepfakes and beyond: A survey of face manipulation and fake detection.* CoRR, abs/2001.00179, 2020. `http://arxiv.org/abs/2001.00179`. 1

[22] Malolan, Badhrinarayan, Ankit Parekh, and Faruk Kazi: *Explainable deep-fake detection using visual interpretability methods.* In *3rd International Conference on Information and Computer Technologies, ICICT 2020, San Jose, CA, USA, March 9-12, 2020*, pages 289–293. IEEE, 2020. https://doi.org/10.1109/ICICT50521.2020.00051. 1

[23] Guizzardi, Renata S. S., Glenda Carla Moura Amaral, Giancarlo Guizzardi, and John Mylopoulos: *Ethical requirements for AI systems.* In *Canadian Conference on AI*, volume 12109 of *Lecture Notes in Computer Science*, pages 251–256, https://doi.org/10.1007/978-3-030-47358-7_24, 2020. Springer. 1, 13, 38, 39, 41, 50, 55, 86

[24] Aberkane, A: *Exploring ethics in requirements engineering.* Master's thesis, UTRECHT UNIVERSITY, 2018. 1, 2

[25] McNamara, Andrew, Justin Smith, and Emerson Murphy-Hill: *Does acm's code of ethics change ethical decision making in software development?* In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 729–733, 2018. 2, 50, 58, 59

[26] Aydemir, Fatma Basak and Fabiano Dalpiaz: *A roadmap for ethics-aware software engineering.* In Brun, Yuriy, Brittany Johnson, and Alexandra Meliou (editors): *Proceedings of the International Workshop on Software Fairness, FairWare@ICSE 2018, Gothenburg, Sweden, May 29, 2018*, pages 15–21. ACM, 2018. https://doi.org/10.1145/3194770.3194778. 2, 34, 50, 53, 57, 70

[27] Jobin, Anna, Marcello Ienca, and Effy Vayena: *The global landscape of ai ethics guidelines.* Nature Machine Intelligence, 1(9):389–399, 2019. 3, 13, 14, 15, 16, 19, 20, 26, 47, 51, 62, 66, 80

[28] Rothenberger, Lea, Benjamin Fabian, and Elmar Arunov: *Relevance of ethical guidelines for artificial intelligence - a survey and evaluation.* In Brocke, Jan vom, Shirley Gregor, and Oliver Müller (editors): *27th European Conference on Information Systems - Information Systems for a Sharing Society, ECIS 2019, Stockholm and Uppsala, Sweden, June 8-14, 2019*, volume 27, pages 1–12, 2019. https://aisel.aisnet.org/ecis2019_rip/26. 3, 20, 51, 63, 80

[29] Floridi, Luciano, 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: *Ai4people - an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations.* Minds Mach., 28(4):689–707, 2018. https://doi.org/10.1007/s11023-018-9482-5. 3, 13, 50, 53, 64

[30] Mittelstadt, Brent: *Principles alone cannot guarantee ethical AI.* Nature Machine Intelligence, 1(11):501–507, November 2019. https://doi.org/10.1038/s42256-019-0114-4. 3, 15, 19, 20, 26, 27, 31, 32, 68, 76

[31] Cath, Corinne: *Governing artificial intelligence: ethical, legal and technical opportunities and challenges.* Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133):20180080, October 2018. https://doi.org/10.1098/rsta.2018.0080. 3

[32] Vakkuri, Ville, Kai-Kristian Kemell, Marianna Jantunen, and Pekka Abrahamsson: *"this is just a prototype": How ethics are ignored in software startup-like environments.* In Stray, Viktoria, Rashina Hoda, Maria Paasivaara, and Philippe Kruchten (editors): *Agile Processes in Software Engineering and Extreme Programming - 21st International Conference on Agile Software Development, XP 2020, Copenhagen, Denmark, June 8-12, 2020, Proceedings*, volume 383 of *Lecture Notes in Business Information Processing*, pages 195–210. Springer, 2020. https://doi.org/10.1007/978-3-030-49392-9_13. 3, 4, 52, 59

[33] Iivari, Juhani and John Venable: *Action research and design science research - seemingly similar but decisively dissimilar.* In Newell, Susan, Edgar A. Whitley, Nancy Pouloudi, Jonathan Wareham, and Lars Mathiassen (editors): *17th European Conference on Information Systems, ECIS 2009, Verona, Italy, 2009*, pages 1642–1653, 2009. http://aisel.aisnet.org/ecis2009/73. 5

[34] Vernadat, F: *Enterprise modeling and integration : principles and applications.* Chapman & Hall, 1996. 6

[35] Russell, Stuart J and Peter Norvig: *Artificial intelligence-a modern approach, third international edition.*, 2010. 9, 10, 11, 13

[36] Benjamins, Richard, Alberto Barbado, and Daniel Sierra: *Responsible ai by design in practice.* arXiv, 1:1–10, 2019. 9, 13, 29, 52, 55, 64

[37] Commission, European: *A definition of AI: Main capabilities and scientific disciplines High-Level Expert Group on artificial intelligence*, April 2019. https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines. 10, 11

[38] Dilek, Selma, Hüseyin Çakir, and M. Aydin: *Applications of artificial intelligence techniques to combating cyber crimes: A review.* ArXiv, abs/1502.03552, 2015. 11

[39] Krafft, Tobias, Marc Hauer, Lajla Fetic, Andreas Kaminski, Michael Puntschuh, Philipp Otto, Christoph Hubig, Torsten Fleischer, Paul Grünke, Rafaela Hillerbrand, Carla Hustedt, and Sebastian Hallensleben: *From principles to practice - an interdisciplinary framework to operationalise ai ethics.* April 2020. 11, 13, 14, 32, 34, 40, 41, 52, 54

[40] Dosilovic, Filip Karlo, Mario Brcic, and Nikica Hlupic: *Explainable artificial intelligence: A survey.* In Skala, Karolj, Marko Koricic, Tihana Galinac Grbac, Marina Cicin-Sain, Vlado Sruk, Slobodan Ribaric, Stjepan Gros, Boris Vrdoljak, Mladen Mauher, Edvard Tijan, Predrag Pale, and Matej Janjic (editors): *41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018, Opatija, Croatia, May 21-25, 2018*, pages 210–215. IEEE, 2018. https://doi.org/10.23919/MIPRO.2018.8400040. 12

[41] Zellers, Rowan, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi: *Defending against neural fake news.* CoRR, abs/1905.12616, 2019. `http://arxiv.org/abs/1905.12616`. 12

[42] European Commission: *Ethics Guidelines for Trustworthy AI High-Level Expert Group on artificial intelligence*, April 2019. `https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai`, visited on 2020-10-08. 13, 19, 21, 28, 47, 69, 77, 81, 82, 84, 133

[43] Peters, Dorian, Karina Vold, Diana Robinson, and Rafael A Calvo: *Responsible ai—two frameworks for ethical design practice.* IEEE Transactions on Technology and Society, 1(1):34–47, 2020. 13, 27, 31, 32, 33, 34, 51, 54, 71

[44] Bird, Eleanor, Jasmin Fox-Skelly, Nicola Jenner, Ruth Larbey, Emma Weitkamp, and Alan Winfield: *The ethics of artificial intelligence issues and initiatives : study Panel for the Future of Science and Technology.* European Union, Brussels, 2020, ISBN 978-92-846-5799-5. 13, 14, 18, 20

[45] Smit, Koen, Martijn Zoet, and John van Meerten: *A review of AI principles in practice.* In Vogel, Doug, Kathy Ning Shen, Pan Shan Ling, Carol Hsu, James Y. L. Thong, Marco De Marco, Moez Limayem, and Sean Xin Xu (editors): *24th Pacific Asia Conference on Information Systems, PACIS 2020, Dubai, UAE, June 22-24, 2020*, page 198, 2020. `https://aisel.aisnet.org/pacis2020/198`. 14, 20, 21, 27, 51, 62, 80, 134

[46] Belani, Hrvoje, Marin Vukovic, and Zeljka Car: *Requirements engineering challenges in building ai-based complex systems.* In *27th IEEE International Requirements Engineering Conference Workshops, RE 2019 Workshops, Jeju Island, Korea (South), September 23-27, 2019*, pages 252–255, 10.1109/REW.2019.00051, 2019. IEEE. `https://doi.org/10.1109/REW.2019.00051`. 14, 37, 39, 51, 59

[47] Benjamins, Richard: *Towards organizational guidelines for the responsible use of AI.* CoRR, abs/2001.09758:1–2, 2020. `https://arxiv.org/abs/2001.09758`. 15

[48] Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar: *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI.* SSRN Electronic Journal, 1(1):1–39, 2020. `https://doi.org/10.2139/ssrn.3518482`. 15, 16, 20, 26, 28, 80

[49] Zeng, Y., Enmeng Lu, and Cunqing Huangfu: *Linking artificial intelligence principles.* ArXiv, abs/1812.04814, 2019. 16, 20, 80

[50] Whittlestone, Jess, Rune Nyrup, Anna Alexandrova, and Stephen Cave: *The role and limits of principles in AI ethics: Towards a focus on tensions.* In Conitzer, Vincent, Gillian K. Hadfield, and Shannon Vallor (editors): *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, January 27-28, 2019*, pages 195–200. ACM, 2019. `https://doi.org/10.1145/3306618.3314289`. 16, 20

[51] Stix, Charlotte: *A survey of the european union's artificial intelligence ecosystem.* Leverhulme Centre for the Future of Intelligence, University of Cambridge, 2019. 16, 19, 20

[52] Carrillo, Margarita Robles: *La gobernanza de la inteligencia artificial: contexto y parámetros generales.* Revista Electrónica de Estudios Internacionales, 2020(39), June 2020. `https://doi.org/10.17103/reei.39.07`. 19

[53] Greene, D., A. Hoffmann, and Luke Stark: *Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning.* In *Proceedings of the 52nd Hawaii International Conference on System Sciences–HICSS*, pages 1–10, https://hdl.handle.net/10125/59651, 2019. HICSS. 19, 31

[54] Whittaker, Meredith, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Mysers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz: *AI now report 2018.* AI Now Institute at New York University New York, 2018. 19

[55] Service, European Parliamentary Research: *EU guidelines on ethics in artificial intelligence: Context and implementation*, September 2019. `https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_BRI(2019)640163`. 19

[56] Nemitz, Paul: *Constitutional democracy and technology in the age of artificial intelligence.* Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133):20180089, October 2018. `https://doi.org/10.1098/rsta.2018.0089`. 19

[57] Smuha, Nathalie A.: *The eu approach to ethics guidelines for trustworthy artificial intelligence.* Computer Law Review International, 20:106 – 97, 2019. 19, 133

[58] IEEE: *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (EAD FirstEdition)*, 2019. `https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf`, visited on 2020-10-08. 20, 21, 47, 69, 77, 137

[59] Conselho da União Europeia: *General Data Protection Regulation (GDPR) (UE) 2016/679*, 2016. `https://data.consilium.europa.eu/doc/document/ST-5419-2016-INIT/en/pdf`, visited on 2020-10-08. 20, 24, 47

[60] Brasil: *Lei nº 13.709, de 14 de agosto de 2018. lei geral de proteção de dados pessoais (lgpd).* Diário Oficial da República Federativa do Brasil, 1:1–5, 2018. `http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm`, visited on 2020-10-08. 20, 24, 47

[61] Verdélio, Andreia: *Lei Geral de Proteção de Dados entra em vigor.* Agência Brasil, 2020. `https://agenciabrasil.ebc.com.br/geral/noticia/2020-09/lei-geral-de-protecao-de-dados-entra-em-vigor`, visited on 2020-10-08. 20

[62] Floridi, Luciano and Josh Cowls: *A unified framework of five principles for AI in society.* Issue 1, 1, June 2019. `https://doi.org/10.1162/99608f92.8cd550d1`. 20, 21, 33, 40, 41, 80, 136

[63] Ágreda, Ángel Gómez de: *Ethics of autonomous weapons systems and its applicability to any ai systems.* Telecommunications Policy, 44:101953, 2020. 20, 80

[64] Vakkuri, Ville and Kai Kristian Kemell: *Implementing artificial intelligence ethics: A tutorial.* In *Lecture Notes in Business Information Processing*, pages 439–442. Springer International Publishing, 2019. `https://doi.org/10.1007/978-3-030-33742-1_38`. 26, 70

[65] Newman, Jessica: *Decision Points in AI Governance: Three case studies explore efforts to operationalize AI principles.* Center for Long-Term Cybersecurity - UC Berkeley, 2020. `https://cltc.berkeley.edu/wp-content/uploads/2020/05/Decision_Points_AI_Governance.pdf`, visited on 2020-10-15. 27

[66] Schiff, Daniel, Bogdana Rakova, Aladdin Ayesh, Anat Fanti, and Michael Lennon: *Principles to practices for responsible ai: Closing the gap.* arXiv preprint arXiv:2006.04707, 2020. 27, 28, 29, 31, 33, 34, 41, 52, 54, 71, 112

[67] Arrieta, Alejandro Barredo, Natalia Díaz Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera: *Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI.* Inf. Fusion, 58:82–115, 2020. `https://doi.org/10.1016/j.inffus.2019.12.012`. 28, 31, 51, 57

[68] Nori, H., S. Jenkins, P. Koch, and R. Caruana: *Interpretml: A unified framework for machine learning interpretability.* ArXiv, abs/1909.09223, 2019. 28

[69] Ribeiro, Marco Túlio, Sameer Singh, and Carlos Guestrin: *"why should I trust you?": Explaining the predictions of any classifier.* In Krishnapuram, Balaji, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi (editors): *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM, 2016. `https://doi.org/10.1145/2939672.2939778`. 28

[70] Lundberg, Scott M. and Su-In Lee: *A unified approach to interpreting model predictions.* In Guyon, Isabelle, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (editors): *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 4765–4774, 2017. `http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions`. 28

[71] Madaio, Michael A., Luke Stark, Jennifer Wortman Vaughan, and H. Wallach: *Co-designing checklists to understand organizational challenges and opportunities around fairness in ai.* Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020. 28, 84

[72] DrivenData: *Deon: An ethics checklist for data scientists*, 2019. `https://deon.drivendata.org/`, visited on 2020-10-20. 28

[73] Johns Hopkins Center for Government Excellence: *Ethics & Algorithms Toolkit*, 2019. `http://ethicstoolkit.ai/`, visited on 2020-10-20. 28

[74] Machine Intelligence Garage: *Digital Catapult's Ethics Committee AI Ethics Framework*, 2019. `https://www.migarage.ai/ethics/ethics-framework/`, visited on 2020-10-20. 28

[75] Havrda, Marek and Bogdana Rakova: *Enhanced well-being assessment as basis for the practical implementation of ethical and rights-based normative principles for AI.* CoRR, abs/2007.14826, 2020. `https://arxiv.org/abs/2007.14826`. 28, 29, 50, 55

[76] Kaminski, Margot E. and G. Malgieri: *Multi-layered explanations from algorithmic impact assessments in the gdpr.* Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020. 28

[77] Schiff, D., A. Ayesh, Laura Musikanski, and John C. Havens: *Ieee 7010: A new standard for assessing the well-being implications of artificial intelligence.* ArXiv, abs/2005.06620, 2020. 28, 29, 32

[78] Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker: *Algorithmic impact assessments. ai now institute*, 2018. 28

[79] Raji, Inioluwa Deborah, Andrew Smart, Rebecca White, M. Mitchell, Timnit Gebru, B. Hutchinson, Jamila Smith-Loud, Daniel Theron, and P. Barnes: *Closing the ai accountability gap: defining an end-to-end framework for internal algorithmic auditing.* Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020. 29, 32, 39, 40

[80] Bellamy, Rachel K. E., Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John T. Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang: *AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias.* CoRR, abs/1810.01943:1–20, 2018. `http://arxiv.org/abs/1810.01943`. 30

[81] Chakraborty, Joymallya, Suvodeep Majumder, Zhe Yu, and Tim Menzies: *Fairway: a way to build fair ML software.* In *ESEC/SIGSOFT FSE*, pages 654–665, https://doi.org/10.1145/3368089.3409697, 2020. ACM. 30

[82] Oneto, Luca and Silvia Chiappa: *Fairness in machine learning.* In *Recent Trends in Learning From Data*, pages 155–196. Springer, 2020. 30

[83] Chakraborty, Joymallya, Kewen Peng, and Tim Menzies: *Making fair ML software using trustworthy explanation.* CoRR, abs/2007.02893:1–5, 2020. 30

[84] Adebayo, Julius: *Fairml : Toolbox for diagnosing bias in predictive modeling*, 2016. 30, 31

[85] Saleiro, Pedro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, J. London, and R. Ghani: *Aequitas: A bias and fairness audit toolkit.* ArXiv, abs/1811.05577, 2018. 30

[86] Sharma, S., J. Henderson, and Joydeep Ghosh: *Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models.* ArXiv, abs/1905.07857, 2019. 30, 31

[87] Umbrello, Steven: *Beneficial artificial intelligence coordination by means of a value sensitive design approach.* Big Data Cogn. Comput., 3(1):5, 2019. `https://doi.org/10.3390/bdcc3010005`. 31

[88] Cawthorne, Dylan and Aimee Robbins-van Wynsberghe: *An ethical framework for the design, development, implementation, and assessment of drones used in public healthcare.* Science and Engineering Ethics, 26:1–25, 2020. 31, 50, 58, 65, 66

[89] Umbrello, Steven and Angelo F. De Bellis: *A value-sensitive design approach to intelligent agents.* 2018. 31

[90] Leikas, Jaana, Raija Koivisto, and Nadezhda Gotcheva: *Ethical framework for designing autonomous intelligent systems.* Journal of Open Innovation: Technology, Market, and Complexity, 5(1):18, 2019. 31, 50, 53, 62, 63

[91] Detweiler, Christian and Maaike Harbers: *Value stories: Putting human values into requirements engineering.* In *REFSQ Workshops*, volume 1138 of *CEUR Workshop Proceedings*, pages 2–11, http://ceur-ws.org/Vol-1138/creare.pdf, 2014. CEUR-WS.org. 31

[92] a3i: *The Trust-in-AI Framework*, 2019. `http://a3i.ai/trust-in-ai`, visited on 2020-10-20. 32

[93] Governo do Reino Unido: *Data Ethics Framework*, 2019. `https://www.gov.uk/government/publications/data-ethics-framework`, visited on 2020-10-20. 32

[94] Leslie, David: *Understanding artificial intelligence ethics and safety.* arXiv preprint arXiv:1906.05684, 2019. 32, 52, 54

[95] Kostova, Blagovesta, Seda Gurses, and Alain Wegmann: *On the interplay between requirements, engineering, and artificial intelligence.* In *Proceedings of REFSQ-2020 Workshops, Pisa, Italy, 2020*, pages 1–5, 2020. `http://ceur-ws.org`. 35, 39, 51, 61, 66, 68

[96] Guizzardi, Renata S. S., Jennifer Horkoff, Anna Perini, and Angelo Susi: *Preface: The international workshop on requirements engineering for artificial intelligence (RE4AI 2020).* In *REFSQ Workshops*, volume 2584 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020. 35

[97] Abualhaija, Sallam, Chetan Arora, Mehrdad Sabetzadeh, Lionel C. Briand, and Eduardo Vaz: *A machine learning-based approach for demarcating requirements in*

textual specifications. In Damian, Daniela E., Anna Perini, and Seok-Won Lee (editors): *27th IEEE International Requirements Engineering Conference, RE 2019, Jeju Island, Korea (South), September 23-27, 2019*, pages 51–62. IEEE, 2019. https://doi.org/10.1109/RE.2019.00017. 35

[98] Navarro-Almanza, Raúl, Reyes Juárez-Ramírez, and G. Licea: *Towards supporting software engineering using deep learning: A case of software requirements classification.* 2017 5th International Conference in Software Engineering Research and Innovation (CONISOFT), pages 116–120, 2017. 35

[99] Binkhonain, Manal and Liping Zhao: *A review of machine learning algorithms for identification and classification of non-functional requirements.* Expert Syst. Appl. X, 1:100001, 2019. https://doi.org/10.1016/j.eswax.2019.100001. 35

[100] Dalpiaz, Fabiano, Davide Dell'Anna, Fatma Basak Aydemir, and Sercan Çevikol: *Requirements classification with interpretable machine learning and dependency parsing.* In Damian, Daniela E., Anna Perini, and Seok-Won Lee (editors): *27th IEEE International Requirements Engineering Conference, RE 2019, Jeju Island, Korea (South), September 23-27, 2019*, pages 142–152. IEEE, 2019. https://doi.org/10.1109/RE.2019.00025. 35

[101] Liu, Su, Reng Zeng, and Xudong He: *An empirical study on classification of non-functional requirements.* In *Proceedings of the 23rd International Conference on Software Engineering & Knowledge Engineering (SEKE'2011), Eden Roc Renaissance, Miami Beach, USA, July 7-9, 2011*, pages 444–449. Knowledge Systems Institute Graduate School, 2011. 35

[102] Mahmoud, Anas and G. Williams: *Detecting, classifying, and tracing non-functional software requirements.* Requirements Engineering, 21:357–381, 2016. 35

[103] Slankas, John and L. Williams: *Automated extraction of non-functional requirements in available documentation.* 2013 1st International Workshop on Natural Language Analysis in Software Engineering (NaturaLiSE), pages 9–16, 2013. 35

[104] Winkler, Jonas and Andreas Vogelsang: *Automatic classification of requirements based on convolutional neural networks.* In *24th IEEE International Requirements Engineering Conference, RE 2016, Beijing, China, September 12-16, 2016*, pages 39–45. IEEE Computer Society, 2016. https://doi.org/10.1109/REW.2016.021. 35

[105] Abad, Zahra Shakeri Hossein, Oliver Karras, Parisa Ghazi, Martin Glinz, Guenther Ruhe, and Kurt Schneider: *What works better? A study of classifying requirements.* In Moreira, Ana, João Araújo, Jane Hayes, and Barbara Paech (editors): *25th IEEE International Requirements Engineering Conference, RE 2017, Lisbon, Portugal, September 4-8, 2017*, pages 496–501. IEEE Computer Society, 2017. https://doi.org/10.1109/RE.2017.36. 35

[106] Sommerville, Ian: *Software engineering. 10th.* In *Book Software Engineering. 10th, Series Software Engineering.* Addison-Wesley, 2015. 36, 37

[107] Ghazi, Parisa and Martin Glinz: *Challenges of working with artifacts in requirements engineering and software engineering.* Requir. Eng., 22(3):359–385, 2017. 36

[108] Franco, Áldrin Jaramillo and Saïd Assar: *Leveraging creativity techniques in requirements elicitation : a literature review.* Applied Zeger to Parisien Life, 2016(02):., June 2016. `https://hal.archives-ouvertes.fr/hal-02375817`. 37

[109] Canedo, Edna Dias, Angélica Toffano Seidel Calazans, Eloisa Toffano Seidel Masson, Pedro Henrique Teixeira Costa, and Fernanda Lima: *Perceptions of ICT practitioners regarding software privacy.* Entropy, 22(4):429, 2020. 37

[110] Alsaqaf, Wasim, Maya Daneva, and Roel J. Wieringa: *Agile quality requirements engineering challenges: First results from a case study.* In *ESEM*, pages 454–459. IEEE Computer Society, 2017. 37

[111] Martins, Hugo Ferreira, Antônio Carvalho de Oliveira Junior, Edna Dias Canedo, Ricardo Ajax Dias Kosloski, Roberto Ávila Paldês, and Edgard Costa Oliveira: *Design thinking: Challenges for software requirements elicitation.* Inf., 10(12):371, 2019. 37

[112] Chazette, Larissa: *Mitigating challenges in the elicitation and analysis of transparency requirements.* In *RE*, pages 470–475. IEEE, 2019. 37

[113] Kahan, Ezequiel, Marcela Genero, and Alejandro Oliveros: *Challenges in requirement engineering: Could design thinking help?* In *QUATIC*, volume 1010 of *Communications in Computer and Information Science*, pages 79–86. Springer, 2019. 37

[114] Schleier-Smith, Johann: *An architecture for agile machine learning in real-time applications.* In Cao, Longbing, Chengqi Zhang, Thorsten Joachims, Geoffrey I. Webb, Dragos D. Margineantu, and Graham Williams (editors): *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pages 2059–2068. ACM, 2015. `https://doi.org/10.1145/2783258.2788628`. 37, 39

[115] Amershi, Saleema, Andrew Begel, Christian Bird, Robert DeLine, Harald C. Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann: *Software engineering for machine learning: a case study.* In Sharp, Helen and Mike Whalen (editors): *Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Practice, ICSE (SEIP) 2019, Montreal, QC, Canada, May 25-31, 2019*, pages 291–300. IEEE / ACM, 2019. `https://doi.org/10.1109/ICSE-SEIP.2019.00042`. 37

[116] Heikkilä, Ville T., Daniela E. Damian, Casper Lassenius, and Maria Paasivaara: *A mapping study on requirements engineering in agile software development.* In *41st Euromicro Conference on Software Engineering and Advanced Applications, EUROMICRO-SEAA 2015, Madeira, Portugal, August 26-28, 2015*, pages 199–207. IEEE Computer Society, 2015. `https://doi.org/10.1109/SEAA.2015.70`. 37, 38, 72

[117] Schwaber, Ken and Mike Beedle: *Agile software development with Scrum*, volume 1. Prentice Hall Upper Saddle River, 2002. 37

[118] Dybå, T. and Torgeir Dingsøyr: *Empirical studies of agile software development: A systematic review.* Inf. Softw. Technol., 50:833–859, 2008. 37

[119] Curcio, Karina, T. Navarro, A. Malucelli, and S. Reinehr: *Requirements engineering: A systematic mapping study in agile software development.* J. Syst. Softw., 139:32–50, 2018. 37, 38

[120] Lucia, A. and Abdallah Qusef: *Requirements engineering in agile software development.* Journal of Emerging Technologies in Web Intelligence, 2:212–220, 2010. 38

[121] Otto, Paul N. and Annie I. Antón: *Addressing legal requirements in requirements engineering.* In *15th IEEE International Requirements Engineering Conference, RE 2007, October 15-19th, 2007, New Delhi, India*, pages 5–14. IEEE Computer Society, 2007. `https://doi.org/10.1109/RE.2007.65`. 38

[122] Nguyen-Duc, Anh, Ingrid Sundbø, Elizamary Nascimento, Tayana Conte, Iftekhar Ahmed, and Pekka Abrahamsson: *A multiple case study of artificial intelligent system development in industry.* In Li, Jingyue, Letizia Jaccheri, Torgeir Dingsøyr, and Ruzanna Chitchyan (editors): *EASE '20: Evaluation and Assessment in Software Engineering, Trondheim, Norway, April 15-17, 2020*, pages 1–10. ACM, 2020. `https://doi.org/10.1145/3383219.3383220`. 39

[123] Lwakatare, Lucy Ellen, Aiswarya Raj, Jan Bosch, Helena Holmström Olsson, and Ivica Crnkovic: *A taxonomy of software engineering challenges for machine learning systems: An empirical investigation.* In Kruchten, Philippe, Steven Fraser, and François Coallier (editors): *Agile Processes in Software Engineering and Extreme Programming - 20th International Conference, XP 2019, Montréal, QC, Canada, May 21-25, 2019, Proceedings*, volume 355 of *Lecture Notes in Business Information Processing*, pages 227–243. Springer, 2019. `https://doi.org/10.1007/978-3-030-19034-7_14`. 39

[124] Cysneiros, Luiz Marcio and Julio Cesar Sampaio do Prado Leite: *Non-functional requirements orienting the development of socially responsible software.* In Nurcan, Selmin, Iris Reinhartz-Berger, Pnina Soffer, and Jelena Zdravkovic (editors): *Enterprise, Business-Process and Information Systems Modeling - 21st International Conference, BPMDS 2020, 25th International Conference, EMMSAD 2020, Held at CAiSE 2020, Grenoble, France, June 8-9, 2020, Proceedings*, volume 387 of *Lecture Notes in Business Information Processing*, pages 335–342. Springer, 2020. `https://doi.org/10.1007/978-3-030-49418-6_23`. 40

[125] Mahnic, V. and T. Hovelja: *On using planning poker for estimating user stories.* J. Syst. Softw., 85:2086–2095, 2012. 41, 70

[126] Felizardo, Kátia Romero, Elisa Yumi Nakagawa, Sandra Camargo Pinto Ferraz Fabbri, and Fabiano Cutigi Ferrari: *Revisão Sistemática da Literatura em Engenharia de Software: Teoria e Prática.* Elsevier Brasil, 2017. 44, 45

[127] Brereton, Pearl, Barbara A Kitchenham, David Budgen, Mark Turner, and Mohamed Khalil: *Lessons from applying the systematic literature review process within the software engineering domain.* Journal of systems and software, 80(4):571–583, 2007. 45

[128] Scantamburlo, Teresa, Atia Cortés, and Marie Schacht: *Progressing towards responsible ai.* arXiv preprint arXiv:2008.07326, 2020. 51, 57, 58

[129] Antonov, Alexander and Tanel Kerikmäe: *Trustworthy ai as a future driver for competitiveness and social change in the eu.* In *The EU in the 21st Century*, pages 135–154. Springer, 2020. 52, 57, 63

[130] Sartor, Giovanni and Francesca Lagioia: *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence : study.* European Parliament, Brussels, 2020, ISBN 978-92-846-6771-0. 52, 64

[131] Vakkuri, Ville, Marianna Jantunen, Erika Halme, Kai-Kristian Kemell, Anh Nguyen-Duc, Tommi Mikkonen, and Pekka Abrahamsson: *Time for AI (ethics) maturity model is now.* CoRR, abs/2101.12701, 2021. `https://arxiv.org/abs/2101.12701`. 52, 54

[132] Morley, Jessica, Anat Elhalal, Francesca Garcia, Libby Kinsey, Jakob Mokander, and Luciano Floridi: *Ethics as a service: a pragmatic operationalisation of AI ethics.* CoRR, abs/2102.09364, 2021. `https://arxiv.org/abs/2102.09364`. 52, 58, 59, 68, 84, 86, 94, 110, 114

[133] Halme, Erika, Ville Vakkuri, Joni Kultanen, Marianna Jantunen, Kai-Kristian Kemell, Rebekah Rousi, and Pekka Abrahamsson: *How to write ethical user stories? impacts of the ECCOLA method.* In Gregory, Peggy, Casper Lassenius, Xiaofeng Wang, and Philippe Kruchten (editors): *Agile Processes in Software Engineering and Extreme Programming - 22nd International Conference on Agile Software Development, XP 2021, Virtual Event, June 14-18, 2021, Proceedings*, volume 419 of *Lecture Notes in Business Information Processing*, pages 36–52. Springer, 2021. `https://doi.org/10.1007/978-3-030-78098-2_3`. 53, 56

[134] Siqueira De Cerqueira, José Antonio, Lucas Dos Santos Althoff, Paulo Santos De Almeida, and Edna Dias Canedo: *Ethical perspectives in ai: A two-folded exploratory study from literature and active development projects.* In *Proceedings of the 54th Hawaii International Conference on System Sciences*, page 5240, 2021. 69, 70, 79, 80

[135] Nori, H., S. Jenkins, P. Koch, and R. Caruana: *InterpretML - Alpha Release*, October 2019. `https://github.com/interpretml/interpret`. 72

[136] Nathanlauga: *InterpretML - Alpha Release*, October 2019. `https://github.com/Nathanlauga/transparentai`. 72

[137] Consortium, World Wide Web: *HTML - Living Standard*, April 2021. `https://html.spec.whatwg.org/multipage/`. 73

[138] Bert Bos, World Wide Web Consortium: *CSS*, April 2021. `https://www.w3.org/Style/CSS/`. 73

[139] Docs, Mozilla MDN Web: *JavaScript*, January 2021. `https://developer.mozilla.org/en-US/docs/Web/JavaScript`. 73

[140] Slack, Dylan, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju: *Fooling lime and shap: Adversarial attacks on post hoc explanation methods.* In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2020. 82

[141] Kim, Michael P, Amirata Ghorbani, and James Zou: *Multiaccuracy: Black-box post-processing for fairness in classification.* In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254. ACM, 2019. 83

[142] Raff, Edward, Jared Sylvester, and Steven Mills: *Fair forests: Regularized tree induction to minimize model bias.* In Furman, Jason, Gary E. Marchant, Huw Price, and Francesca Rossi (editors): *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, pages 243–250. ACM, 2018. `https://doi.org/10.1145/3278721.3278742`. 83

[143] Theodorou, Andreas, Robert H. Wortham, and Joanna J. Bryson: *Designing and implementing transparency for real time inspection of autonomous robots.* Connect. Sci., 29(3):230–241, 2017. `https://doi.org/10.1080/09540091.2017.1310182`. 83

[144] Méndez, Daniel, Daniel Graziotin, Stefan Wagner, and Heidi Seibold: *Open science in software engineering.* In Felderer, Michael and Guilherme Horta Travassos (editors): *Contemporary Empirical Methods in Software Engineering*, pages 477–501. Springer, 2020. `https://doi.org/10.1007/978-3-030-32489-6_17`. 89

[145] Martinez-Martin, Nicole: *What are important ethical implications of using facial recognition technology in health care?* AMA journal of ethics, 21 2:E180–187, 2019. 91

[146] Pfleeger, Shari Lawrence and Barbara A. Kitchenham: *Principles of survey research: part 1: turning lemons into lemonade.* ACM SIGSOFT Softw. Eng. Notes, 26(6):16–18, 2001. `https://doi.org/10.1145/505532.505535`. 93

[147] Molléri, Jefferson Seide, Kai Petersen, and Emilia Mendes: *Survey guidelines in software engineering: An annotated review.* In *Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM 2016, Ciudad Real, Spain, September 8-9, 2016*, pages 58:1–58:6. ACM, 2016. `https://doi.org/10.1145/2961111.2962619`. 93

[148] Krippendorff, Klaus: *Content analysis: An introduction to its methodology.* Sage publications, 2018. 94

[149] Langford, J. and D. McDonagh: *Focus groups: Supporting effective product development.* 2002. 105, 110

[150] Beauchamp, Tom and James Childress: *Principles of biomedical ethics.* Oxford University Press, New York, 2013, ISBN 978-0-19-992458-5. 136

Appendix A

Publicly Available AI Ethics Guidelines

Group 1: Members of society

(a) Professional associations

Table A.1: Guidelines published by professional associations

Title	Origin	Place	Month	Year
Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (EAD First Edition)	IEEE	United States	March	2019
Principles for Algorithmic Transparency and Accountability	ACM US Public Policy Council	United States	January	2017

(b) Civil society/lawyers groups:

Table A.2: Guidelines published by civil society/lawyers groups

Title	Origin	Place	Month	Year
Top 10 Principles for Ethical AI	UNI Global Union	Switzerland	December	2017
Toronto Declaration	Amnesty International & Access Now	Canada	May	2018
Universal Guidelines for AI	The Public Voice Coalition	Belgium	October	2018
Human Rights in the Age of AI	Access Now	United States	November	2018
Future of Work and Education for the Digital Age	T20: Think20	Argentina	July	2018

(c) Nonprofit organisations

Table A.3: Guidelines published by nonprofit organisations

Title	Origin	Place	Month	Year
Asilomar AI Principles	Future of Life Institute	United States	January	2017
Three Rules for Artificial Intelligence Systems by the CEO of Allen Institute for Artificial Intelligence	Allen Institute for Artificial Intelligence (AI2)	United States	September	2017
Principles for the Governance of AI	The Future Society	United States	July	2017
Tenets of Partnership on AI	Partneship on AI	United States	Not found	2016

(d) Academia

Table A.4: Guidelines published by academia

Title	Origin	Place	Month	Year
The Japanese Society for Artificial Intelligence Ethical Guidelines	JSAI	Japan	February	2017
The Montreal Declaration for a Responsible Development of Artificial Intelligence	University of Montreal	Canada	November	2017
Three ideas from the Stanford Human-Centered AI Initiative	Stanford University	United States	Not found	2018
Harmonious Artificial Intelligence Principles	Chinese Academy of Science	China	Not found	2018

Group 2: National and international organisations

Table A.5: Guidelines published by national and international organisations

Title	Origin	Place	Month	Year
Ethics Guidelines for Trustworthy AI	High-Level Expert Group on Artificial Intelligence	Euopean Union	April	2019
Principles on AI	Organisation for Economic Co-operation and Development (OECD)	France	June	2019
UNESCO AHEG Draft text on Recommendation on the Ethics of Artificial Intelligence	UNESCO	France	April	2020

Principles to Promote FEAT AI in the Financial Sector	Monetary Authority of Singapore	Singapure	February	2019
AI in the UK	Select Committee on Artificial Intelligence of the UK House of Lords	United Kingdom	April	2018
Beijing AI Principles	Beijing Academy of Artificial Intelligence (BAAI)	China	May	2019
Governance Principles for a New Generation of AI	Chinese National Governance Committee for AI	China	June	2019
Social Principles of Human-Centric AI	Government of Japan; Cabinet Office; Council for Science, Technology and Innovation	Japan	March	2019
AI Principles and Ethics	Smart Dubai	United Arab Emirates	January	2019
AI Strategy	German Federal Ministries of Education, Economic Affairs, and Labour and Social Affairs	Germany	November	2018
AI in Mexico	British Embassy in Mexico City	Mexico	June	2018
National Strategy for AI	Niti Aayog	India	June	2018
AI for Humanity	French Strategy for Artificial Intelligence	France	March	2018

Group 3: Private sector and industry

Table A.6: Guidelines published by private sector and industry

Title	Origin	Place	Month	Year
Microsoft AI Principles	Microsoft	United States	February	2018
AI at Google: Our Principles	Google	United States	June	2018
DeepMind Ethics & Society Principles	Google DeepMind	United Kingdom	Not found	2017
IBM Everyday Ethics for AI	IBM	United States	October	2019
AI Principles at Telefónica	Telefónica	Spain	October	2018
Six Principles of AI	Tencent Institute	China	April	2017
Sony Group AI Ethics Guidelines	Sony Group	Japan	September	2018
SAP's Guiding Principles for Artificial Intelligence	SAP	Germany	September	2018

The Ethics of Code: Developing AI for Business with Five Core Principles	Sage	United Kingdom	June	2017
AI Policy Principles	ITI	United States	October	2017
OpenAI Charter	OpenAI	United States	April	2018

Appendix B

Principles of AI Ethics

B.1 Ethics Guidelines for Trustworthy AI (AI HLEG)

The AI ethics guidelines from the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG) [42], titled Ethics Guidelines for Trustworthy AI, prescribes four principles that should be considered as ethical imperatives in the context of AI:

1. Respect for human autonomy,
2. Prevention of harm,
3. Fairness,
4. Explicability.

From these principles, seven requirements that AI-based systems should take into account were devised, considering technical and non-technical methods to ensure the implementation of these requirements [57]:

1. Human agency and oversight (including fundamental rights, human agency and human oversight),
2. Technical robustness and safety (including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility),
3. Privacy and data governance (including respect for privacy, quality and integrity of data, and access to data),
4. Transparency (including traceability, explainability and communication),

5. Diversity, non-discrimination and fairness (including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation),
6. Societal and environmental wellbeing (including sustainability and environmental friendliness, social impact, society and democracy),
7. Accountability (including auditability, minimisation and reporting of negative impact, trade-offs and redress)

B.2 Smit et al.: A Review of AI Principles in Practice

They investigated several AI ethics guidelines by exploring which design-level AI principles are available in these guidelines. 22 categories of principles were found, followed by their synonyms (in parentheses when they exist) and the definition of the design principle [45]:

1. Human augmentation. An AI must be designed and used to enhance human productivity and/or capability.
2. Do Good (Sustainability). An AI must be designed and used to enhance financial, manufactured, intellectual, human, social & relation and/or natural capital.
3. Trustworthy (Honest). An AI must be designed and used so that it's deserving of trust, or able to be trusted.
4. Human-Centric. AI-based systems must be: A) designed and used so that humans are involved in the development, B) developed with the user in mind.
5. Autonomy. AI-based systems must be designed and executed such that a specified extent of human control is possible.
6. Equality, design level (Non-biased, fairness, preventing discrimination). A designed AI-based system must treat all people equal.
7. Equality, execution level (Non-biased, fairness, preventing discrimination). A deployed AI-based system must: be accessible for each human such that equal usage opportunity exists, treat all people equal.
8. Traceability (Reproducibility). AI-based system must be designed such that it can provide the applied business logic to reach its conclusion or perform an action.

9. Human dignity (Freedom of groups). AI-based system must be designed and executed such that it will respect human (inherent) dignity on a group and individual level.

10. Human rights. AI-based systems must be designed and executed such that it will respect human rights (status dignity).

11. Transparency, design level. AI-based systems must be designed and executed so that it is possible for humans to get insight into business logic applied.

12. Democrability. Democratic debate and public engagement should drive the design and execution of AI.

13. Privacy. AI-based systems must be designed and executed such that: A) it runs on anonymized data or anonymize data, B) users preserve power over access and use of their data.

14. Security. AI-based systems must be designed and executed to provide maximum security against internal and external malicious or accidental threats.

15. Safety, design level. AI-based systems must be: A) designed by people that have a technical background and understand security risks, B) designed and executed such that it takes into account human safety and reduces possible harm.

16. Safety, execution level. A deployed AI-based system must be reliable and save but also protect the privacy and security of individuals or groups.

17. Collaboration. AI-based systems must be designed and executed to promote human-AI collaboration.

18. Accountability (Liability, Oversight, Responsibility). A person or organization is responsible for the design and execution of an AI-based system.

19. Understandability (Interpretability, Explainability). AI-based systems must be designed so that humans are able to understand (language, presentation) the manner of working of the AI.

20. Responsible use of data (Narrowness). AI-based systems must be designed and used so that it only utilizes relevant and representative data.

21. Accuracy. AI-based systems must be designed to function as exact as possible.

22. Education & Promotion. Democratic debate and public engagement should drive the design and execution of AI.

The five most frequent principles in the documents explored by the authors are: Do good, Accountability, Equality, Privacy, and Education.

B.3 Floridi and Cowls: A unified framework of five principles for ethical AI

Exploram seis conjuntos de diretrizes, que juntos contêm 47 princípios, e, ao encontrar um certo grau de convergência e sobreposição entre os princípios, os comparam em relação aos 4 princípios de ética em biomédica: beneficência, não-maleficência, autonomia e justiça [150]; argumentando que um novo princípio deve ser adicionado, explicabilidade. Apresentamos os cinco princípios acompanhados de suas respectivas definições [62]:

They explored six guidelines, which together contain 47 principles, and, after finding a certain degree of convergence and overlap between the principles, they compared them in relation to the 4 principles of ethics in biomedical: beneficence, non-maleficence, autonomy and justice [150]; arguing that a new principle should be added, explicability. We present the five principles accompanied by their respective definitions [62]:

1. Beneficence: promoting well-being, preserving dignity and sustaining the planet. Promote the well-being of people and planet with AI.

2. Non-maleficence: privacy, security and capability caution. Precautions against the various negative consequences of AI misuse.

3. Autonomy: the power to decide. Promoting human autonomy while restricting the autonomy of AI-based systems and valuing the reversibility of their decision-making.

4. Justice: promoting prosperity, preserving solidarity, avoiding unfairness. Use AI-based systems to eliminate unfair discrimination, promote diversity, prevent new threats to justice from emerging.

5. Explicability: enabling the other principles through intelligibility and accountability. This principle concerns the transparency, accountability, comprehensibility, understanding and interpretation, of an AI-based system, complementing and enabling the other four principles.

B.4 IEEE Ethically Aligned Design (EAD) - First Edition

The first version of the IEEE Ethically Aligned Design, A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (A/IS) is a document created by the IEEE Global Initiative, and has eight general principles that should guide the ethical design, development and implementation of technologies [58]:

1. Human rights: AI-based systems should be created and employed to respect, promote, and protect internationally recognised human rights.

2. Prioritizing Well-being: Developers of AI-based systems should adopt increasing human well-being as the primary criterion for development.

3. Data Agency: The designers of AI-based systems must empower individuals with the ability to access and share their data securely, to maintain people's ability to have control over their identity.

4. Effectiveness: Developers and those operating AI-based systems should provide evidence of the effectiveness and capability for the purpose of the systems.

5. Transparency: The basis of a decision of an AI-based system should always be discoverable.

6. Accountability: AI-based systems should be developed and employed in a way that provides unambiguous rationale for all decisions taken.

7. Awareness of Misuse: Developers of AI-based systems must protect against all potential misuses and risks of these systems in use.

8. Competence: AI developers must specify while operators must adhere to the knowledge and skills required for safe and effective operation.

Appendix C

Evaluation Questionnaire for the RE4AI Ethical Guide

In order to evaluate the Guide created, a questionnaire was devised to investigate participants – AI practitioners and students. The questionnaire was divided into two groups of questions: a) Demographic questions about the characteristics of the participants; b) Questions regarding the evaluation of the Guide.

0%

Demográfico

*Qual a sua faixa etária?

Choose one of the following answers

○ Até 17 anos

○ De 18 a 24 anos

○ De 25 a 35 anos

○ De 36 a 50 anos

○ A partir de 51 anos

*Qual seu grau de formação?

Choose one of the following answers

○ Superior completo

○ Cursando graduação

○ Mestrado

○ Cursando mestrado

○ Doutorado

○ Cursando doutorado

○ Outro. Qual?

*Você está participando de algum projeto de desenvolvimento de software no contexto de IA?

Choose one of the following answers

○ Sim.

○ Não.

*Quantos projetos de desenvolvimento de software no contexto de IA você já participou?

Choose one of the following answers

○ Nenhum.

○ 1.

○ 2.

○ 3.

○ 4.

○ 5 ou mais.

*Possui conhecimento prévio sobre questões éticas no contexto de IA e suas implicações?

Choose one of the following answers

○ Sim.

○ Não.

*Você já recebeu algum treinamento relacionado às diretrizes éticas no contexto de IA anteriormente?

Choose one of the following answers

○ Sim.

○ Não.

50%

Avaliação do RE4AI Ethical Guide

*Em relação ao conteúdo de apoio presente no guia, as informações foram suficientes para o seu entendimento e utilização? (Introdução, Princípios, ferramentas, trade-offs; foi necessária a consulta de outras fontes?)

Comente (sua opinião é importante para a evolução do RE4AI Ethical Guide)

Choose one of the following answers

○ Concordo completamente.

○ Concordo.

○ Neutro.

○ Discordo.

○ Discordo completamente.

Please enter your comment here:

*Você já conhecia alguma das ferramentas sugeridas pelo Guia?

Choose one of the following answers

○ Não.

○ Sim. Quais?

*Em relação às ferramentas sugeridas, acredita que elas têm utilidade na implementação de ética em IA? Comente.

Choose one of the following answers

○ Concordo completamente.

○ Concordo.

○ Neutro.

○ Discordo.

○ Discordo completamente.

Please enter your comment here:

*Selecione um ou mais Princípios que considere mais facilmente implementáveis:

Check all that apply

□ Transparency (Transparência)

□ Justice and fairness (Justiça e equidade)

□ Non-maleficence (Não-maleficência)

□ Responsibility (Responsabilidade)

□ Privacy (Privacidade)

□ Beneficence (Beneficência)

□ Freedom and autonomy (Liberdade e autonomia)

□ Trust (Confiança)

□ Sustainability (Sustentabilidade)

□ Dignity (Dignidade)

□ Solidarity (Solidariedade)

□ Nenhum

*Você considerou fácil a compreensão das perguntas apresentadas nas cartas do Guia?

❗ Choose one of the following answers

○ Concordo completamente.

○ Concordo.

○ Neutro.

○ Discordo.

○ Discordo completamente.

Please enter your comment here:

*Em relação às perguntas presentes nas cartas do guia, as questões respondidas pelo uso das cartas ajudaram a elicitar os requisitos éticos (i.e., criação de histórias de usuário)? Por favor, comente.

❗ Choose one of the following answers

○ Concordo completamente.

○ Concordo.

○ Neutro.

○ Discordo.

○ Discordo completamente.

Please enter your comment here:

*Você acha que o guia ajudou a equipe de desenvolvimento de software a identificar e elicitar os requisitos éticos? Descreva como foi essa ajuda.

❗ Choose one of the following answers

○ Concordo completamente.

○ Concordo.

○ Neutro.

○ Discordo.

○ Discordo completamente.

Please enter your comment here:

*O guia melhorou a sua consciência ética e o seu aprendizado? Se sim, por favor, descreva como.

❗ Choose one of the following answers

○ Concordo completamente.

○ Concordo.

○ Neutro.

○ Discordo.

○ Discordo completamente.

Please enter your comment here:

*Em qual fase do processo de desenvolvimento de software você considera mais viável a utilização do Guia?

❶ Check all that apply

□ Análise de requisitos

□ Projeto

□ Codificação

□ Teste

□ Implementação e manutenção

*Você usaria o RE4AI Ethical Guide na elicitação de requisitos? Por favor, explique sua resposta.

❶ Choose one of the following answers

○ Concordo completamente.

○ Concordo.

○ Neutro.

○ Discordo.

○ Discordo completamente.

Please enter your comment here:

Você tem alguma sugestão de melhoria para o Guia?

Appendix D

Ethical Perspectives in AI: A Two-folded Exploratory Study From Literature and Active Development Projects

Aiming to understand AI ethics in practice and its relation with requirements engineering, we explored a mixed method study, where a bibliometric review of the literature was performed using the TEMAC method, afterwards we explored GitHub repositories using the README files of each repository to map requirements with ethical principles. The objective was to obtain an overview of the current state of the literature and software projects on tools, methods and techniques used in practical AI ethics.

Proceedings of the 54th Hawaii International Conference on System Sciences | 2021

Ethical Perspectives in AI: A Two-folded Exploratory Study From Literature and Active Development Projects

José Antonio Siqueira de Cerqueira
University of Brasília(UnB)
Brasília, DF, Brazil
antonio.cerqueira@aluno.unb.br

Lucas dos S. Althoff
University of Brasília(UnB)
Brasília, DF, Brazil
lucasa@aluno.unb.br

Paulo Santos de Almeida
University of Brasília(UnB)
Brasília, DF, Brazil
paulo.almeida@redes.unb.br

Edna Dias Canedo
University of Brasília(UnB)
Brasília, DF, Brazil
ednacanedo@unb.br

Abstract

Background: *Interest in Artificial Intelligence (AI) based systems has been gaining traction at a fast pace, both for software development teams and for society as a whole. This increased interest has lead to the employment of AI techniques such as Machine Learning and Deep Learning for diverse purposes, like medicine and surveillance systems, and such uses have raised the awareness about the ethical implications of the usage of AI systems.* ***Aims***: *With this work we aim to obtain an overview of the current state of the literature and software projects on tools, methods and techniques used in practical AI ethics.* ***Method***: *We have conducted an exploratory study in both a scientific database and a software projects repository in order to understand their current state on techniques, methods and tools used for implementing AI ethics.* ***Results***: *A total of 182 abstracts were retrieved and five classes were devised from the analysis in Scopus, 1) AI in Agile and Business for Requirement Engineering (RE) (22.8%), 2) RE in Theoretical Context (14.8%), 3) Quality Requirements (22.6%), 4) Proceedings and Conferences (22%), 5) AI in Requirements Engineering (17.8%). Furthermore, out of 589 projects from GitHub, we found 21 tools for implementing AI ethics. Highlighted publicly available tools found to assist the implementation of AI ethics are InterpretML, Deon and TransparentAI.* ***Conclusions***: *The combined energy of both explored sources fosters an enhanced debate and stimulates progress towards AI ethics in practice.*

1. Introduction

There is an increasing number of software development teams building Artificial Intelligence (AI) based systems, and they are gaining popularity in our society at a fast pace [1, 2]. The use of AI techniques like Machine Learning (ML) and Deep Learning (DL) on diverse fields such as medicine, surveillance systems, business, transportation, and many other domains, have raised great awareness about the ethical implications of the use of such systems [3, 2], becoming subject of urging interest to the industry, researchers in academia, and the population at large [4].

Whilst AI popularization is growing, incidents related to those AI-based systems are also becoming more common [3]; several notorious incidents have led to public discussion on AI ethics. One such case is the Cambridge Analytic scandal, where data from Facebook users were obtained inappropriately and used to influence the result of an election [5]. Another example is a biased ML algorithm against women by Amazon, which led to more male candidates being hired [6]. Also, new threats rise concerning ethical misuse of AI bases system such as fake news with the use of deep-fake and AI-based voice technologies, where someone's face could be superimposed on videos and political leaders can be depicted inciting violence and panic, for instance, may be used to rig elections, change political opinions and spread misinformation in general [7].

Various ethical guidelines and principles for AI have been proposed by organizations, commissions, institutes and the industry. Those propositions, however, do not meet the demands from real world development of ethical AI-based systems, as these ethical principles are often too high level, abstract and general [8, 9, 10] and pose no real evidence that they can influence ethical decision making [11]. As a result, those in charge of developing such AI-based systems who are also concerned with the ethical questions that come up have become frustrated by how little help is offered by the highly theoretical texts provided by the principles and codes available [2]. Hence, developing ethical AI is an overwhelmingly defying and complicated task [7]. Most of the studies found in the literature focus to a large extent on theoretical and conceptual principles and guidelines, not providing an effective and realistic framework on how to implement ethics in AI [3, 2]. Therefore, a deeper focus on technological details of the various methods and technologies in the AI and

URI: https://hdl.handle.net/10125/71257
978-0-9981331-4-0

ML area is needed; in other words, currently there is a need to decrease the distance between abstract values (principles, guidelines and codes) and technical implementations [8, 2].

The main objective of this work is to identify tools, methods and techniques already publicly available to assist practitioners involved in the development of AI-based systems to implement ethical principles in their work, hence shedding some light on the topic of applied ethics in AI and bridging the gap between said principles and practice. We devised five classes from our scientific database analysis in Scopus, 1) AI in Agile and Business for Requirement Engineering (RE) (22.8%), 2) RE in Theoretical Context (14.8%), 3) Quality Requirements (22.6%), 4) Proceedings and Conferences (22%), 5) AI in RE (17.8%). The combined collaboration of both scientific and open-source sources fosters a broadened debate on this topic.

2. Background and Related Work

Various institutions and organizations from public and private sectors presented guidelines and principles towards ethical AI-based systems. Two of them will be highlighted in this part of the study: 1) The Ethically Aligned Design [12] by the Institute of Electrical Electronics Engineers (IEEE), whose first edition was published in 2019, presenting analyses and recommendations as a guidance for governments, business and public to take as consideration when dealing with the advancement of AI for the benefit of humanity; 2) Trustworthy AI by the High-Level Expert Group on AI from the European Commission, which was presented on April 2019 [13], showing a set of 7 key requirements that AI systems should meet in order to be considered trustworthy: a) Human agency and oversight; b) Technical robustness and safety; c) Privacy and data governance; d) Transparency; e) Diversity, non-discrimination and fairness; f) Societal and environmental well-being; g) Accountability.

In the academia there are several authors who are currently researching on ethical principles and guidelines. In other words, plenty of reviews and surveys about ethical guidelines, frameworks and principles are available in the current literature. Jobin et al. [14] introduced a comprehensive mapping of the current AI ethics landscape on 84 guidelines proposed worldwide by the private and public sectors providing an overview of the most relevant principles among them. They argue that there is a fast increase in the number and variety of documents that evinces the growing interest by the international community for ethics in AI, but the proposed principles and guidelines have a significant divergence on how to achieve ethical AI. A major cause of divergence among them is how they should be implemented.

Floridi et al. [15] provided a synthesis of six sets of guidelines, extracting 47 principles that overall have a great degree of coherence and overlap among them. The authors state that the four core principles commonly used in bioethics: 1) beneficence, 2) non-maleficence, 3) autonomy, 4) justice and a fifth one, explicability, are greatly adapted to ethical challenges in AI. They seized the significance of each of the 47 principles, forming an ethical framework, within which they offer a list of recommendations with 20 items. The work, however, is majorly conceptual and oriented to government policies, not providing technical solutions to developers. Rothenberger et al. [16] presented a survey and evaluation of guidelines for AI ethics, providing a sum of 5 principles and a ranking of these principles through interviews with 51 experts. They argue that responsibility was ranked first, but respondents asked who would be responsible for the actions of an AI.

Hagendorff [8] analyzed and compared 22 guidelines, finding that almost all guidelines suggest that technical solutions exist, yet not providing technical explanations. Moreover, he states that to deduce concrete technological implementations from very abstract ethical principles is a major problem. Therefore, he considers that there is an urge to close the gap between ethics and technical discourses. McNamara et al. [11] surveyed 63 software engineering students and 105 professional software developers to understand the impact of the Association for Computing Machinery (ACM) Code of Ethics and Professional Conduct [17]. Surprisingly, the authors concluded that there is no evidence that the ACM code of ethics influences ethical decision making in software development. Thus, we consider that discussion regarding theoretical ethics in AI is already consistent and further discussion on this theme will not be approached in this study, as there seems to be an agreement concerning theoretical ethical aspects [2]. On the other hand, there is an urge to perform the translation between the 'what' and the 'how' in AI ethics [2].

A limited number of works available argue about practical ethical AI. Most of literature do not propose a method or tool to implement ethics in AI, rather they survey available tools or perceptions from practitioners. Morley et al. [2] presented a typology and a catalog of available tools and methods to translate principles into practices, in the ML field. However, most of the tools found lack good documentation and focus on small portions of the software development process. Hence, despite being promising, they demand extra work before use. In addition, the authors pinpointed some possible opportunities to researchers such as provide an assessment of the catalogued tools and examples of its usage.

Vakkuri et al. [3] surveyed 211 software companies in an attempt to understand the current state of practice of ethics in AI in the industry, arguing that it is still in its early stages. The authors suggest that a starting point to implement ethics in AI is by the use of the guidelines available. Nonetheless, those are not practical for developers, thus requiring additional work before they can be carried into a real system. The authors concluded that practitioners have a key role in implementing ethics in AI, once activities in AI software projects are nearly the same as in any other software project. In addition, AI ethics implementation from a software development viewpoint could be seen as a non-functional requirement of an AI based system, and that the Product Owner has the responsibility to ensure ethical User Stories in sprint backlogs. Finally, the authors indicated three issues to be avoided regarding applied AI ethics: 1) Do not outsource ethics in AI software development; 2) Do not assume that ethics can implemented without being systematically done; 3) Do not delegate ethics implementation to a single person.

Another recent research by Vakkuri et al. [18] surveyed applied ethics in AI in the start-ups context. The authors discovered that several Software Engineering practices – well established and in existence – can be used to implement AI ethics, e.g., documentation. Besides, despite practitioners having ethical concerns, AI ethics is largely not being implemented, partially being a result of a lack of formal methods and tools to implement it. Despite the existence of the catalogued tools in Morley et al. [2], the authors focused on small parts of the development process, have little usability, and for what concerns AI ethics, there is a need for it to be addressed from the beginning of development, that is, from the requirements elicitation stage [3].

Requirements Engineering (RE) is seen as the first stage of the software development life cycle that deals with the elicitation, analysis, specification, and validation of software requirements as well as the management and documentation of requirements. To start a discussion about ethical requirements in AI, first we need to address RE in AI in a general manner. Vogelsang and Borg [1], in an attempt to understand the perspectives from data scientists regarding RE for ML, stated that there is not much work on RE for ML systems, while literature suggests that RE is the most difficult activity for the development of ML-based systems. Their main findings include that requirements engineers needs to be conscious about new requirements introduced by the ML paradigm, which are explicability and freedom from discrimination. The first type of requirement was mentioned as important quality requirement in their interviews. However, the second type seems more problematic once ML algorithms are designed to identify recurring patterns in data applying these patterns to judge about concealed data. They pointed two reasons for this last statement, first that discrimination is more implicit in ML systems and second that ML algorithms enlarge discrimination bias in the data during the training process. Moreover, all interviewees mentioned challenges concerning ethics and legal aspects.

Belani and Car [19] proposed a RE4AI taxonomy, that is, RE for AI, bringing forth an overview of challenges posed to RE towards building AI-based complex systems. We highlight that, for elicitation activity of RE, the authors defined regulation – ethics – not clear, as a problem related to the system to be. Kostova et al. [20] stated that ”RE is the only place to address problems related to the use of AI based systems due to its interdisciplinary nature, with a strong technical emphasis”. Their work identified two faces of RE discipline in AI, first AI tools are used more frequently during the RE process (AI4RE), second the RE process for AI based systems is different (RE4AI). In this research, our interest resides in the second aspect.

Few works discuss about RE for ethical requirements of AI. Guizzardi et al. [9] presented a definition of Ethical Requirements as the ones derived from guidelines and ethical principles. And the key concept behind it is of Runtime stakeholders, defined as persons that are using, are affected by, or influencing the results and outcomes of an AI based system while in operation. The authors argued that Runtime stakeholders are often ignored in traditional RE. More importantly, they stated that ”Ethical requirements are functional and quality requirements elicited from Runtime stakeholders in accordance with the five ethical principles – beneficence, non-maleficence, autonomy, justice, explicability” [9], the same principles provided by Floridi et al. [15]. Their main goal is to use traditional RE techniques to derive ethical requirements to the case of driver-less cars, to make sure they comply to ethical principles. However, they do not present a systematic methodology employed to do so, neither a validation of the technique.

Vakkuri et al. [21] introduced a tool as a starting point for implementing ethics in AI, named ECCOLA. The proposed tool is a deck of cards to raise awareness of AI ethics in a development team, once the team produces documentation of their ethical decision-making, as for example in the form of non-functional requirements in product backlog. However, there is no validation method of the proposed tool yet. Aberkane [10] in his Master's thesis performed a systematic literature review of ethics and requirements engineering in IEEE Xplore. Using the ACM/IEEE-CS Code of Ethics as a guide the author presented an extensive list of ethical issues identified in the literature. The author does not address in-depth analyses on ethical issues in AI through his study.

Our paper aims to broaden the discussion on ethics in AI by exploring a scientific database (Scopus), also a database for open source projects (GitHub), in order to understand the ethical requirements of AI in academia and the present state of such requirements in publicly available software projects. To the best of our knowledge, no study has approached GitHub for ethical requirements in AI. Moreover, we found that it was already conducted a bibliometric research in Web of Science database to explore sustainable requirements in AI [22]; thus we choose Scopus, as it is a large and well known scientific database, with different sets of criteria and analyses.

3. Methodology

Our research strategy is based on two different approaches. The first one is based on the Theory of Consolidated Meta-Analytic Approach (TEMAC) method [23], from which we only implement few steps aimed at extracting the fundamental literature inside the scope of our study, and to retain the main features about the topography of this field of research. The second one is based on mining GitHub for projects related to ethics in AI, where we explore practical implementations and relate them to the findings inside the field of research. In order to explore a scientific database to extract valuable information from bibliometric data we will use the Theory of Consolidated Meta-Analytic Approach, that provides an objective technique that allows metrics to be established between literary researches in the same field through rigid systematization of research meta-data. We will not provide a full analysis over the TEMAC method, rather we will highlight important steps used in this research.

The whole method is comprised of three stages. The first stage (Preparing the research) is to define the correct keyword, the year range, the scientific databases for extraction, and the area of knowledge, to the study. The second stage (Presenting and inter-relating the data) is to extract bibliometric information from the databases and the application of bibliometric laws to analyse relations between them. The third and last stage (Integrating and validating models) to be used on the evidence acquired from Citation and Bibliographic Coupling mapping study. Highlighting main studies, approaches and lines of research – VosViewer and Iramuteq tools are used to present graphical analyses. In this paper, we will prune parts of TEMAC, which is advised by its creators. In other words, we will present analyses over the first stage, due to its nature of preparing the research, while in the second stage, where there is a wide set of possible analyses, we will focus on a small set suitable to the aims of this research. In addition, the third stage presents an important in-depth study over the data obtained, hence leading to a better understanding.

The second approach explores GitHub – a major platform that hosts open source software projects – that became the largest open source community in the world. In each repository in GitHub a README file is present – one of the first documents that developers sees when coming into contact with a new repository – informing other people the usefulness of a project, what they can do with it, and how they can use it [24]. Looking forward to track traces of how AI projects assimilate ethical principles and concepts we perform a qualitative analyses over GitHub README files. With the study of Portugal et al. [25] as basis, in this work a similar approach is devised adapting their proposed methods to our goal. Further, we detail steps that will be performed on the Section (4) to perform our investigation and analyses of README files from GitHub:

1. Retrieve a corpus of README files from GitHub corresponding to a query using a keyword. We used the tool Corpus Retrieval, presented by Portugal et al. [26]; 2. Manually categorizing the README files into different types; 3. Highlight tools found in previous step; 4. Discover the most relevant keyword in the corpus set. This is done filtering out supplementary forms and unmeaningful active forms, and through the generation and analysis of a word cloud; 5. Expand analysis through the use of POS-Tagging technique in the whole corpus, retrieving frequency of verbs that are commonly related to requirements – RE patterns candidates; 6. Create a separate sub corpus for each RE pattern; 7. Extract information concerning requirements, by manual reading, and listing, for each sub corpus produced; 8. Categorize found requirements into ethical principles in AI. In order to provide a better visualization, Figure 1 shows how we approach exploring GitHub's README files.

4. Results and Analyses

An initial step to perform bibliometric analyses over a scientific database is the preparation and definition

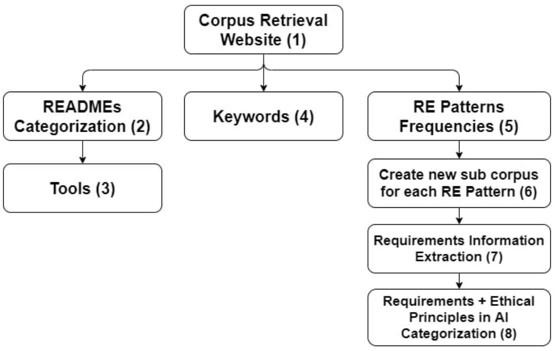

Corpus Retrieval Website (1)
READMEs Categorization (2)
Keywords (4)
RE Patterns Frequencies (5)
Tools (3)
Create new sub corpus for each RE Pattern (6)
Requirements Information Extraction (7)
Requirements + Ethical Principles in AI Categorization (8)

Figure 1. Methodology for exploring GitHub

of parameters for the research. To reduce meaning suppression as well as to capture the majority of works associating AI, ethics and software requirements we designed a search pattern divided in these three branches in the Scopus database: ((”ethic” OR ”ethics” OR ”ethical” OR ”moral” OR ”code of conduct” OR ”transparency” OR ”security”) AND (”requirement analysis” OR ”requirements specification” OR ”requirement elicitation” OR ”software development” OR ”requirement engineering”) AND (”artificial intelligence” OR ”machine learning” OR ”deep learning” OR ”predictive model” OR ”DL” OR ”ML” OR ”AI”)).

Each branch is composed by a set of interchangeable words selected to retain all aimed subtopics inside the database. The conduction of this research was from 10th of May to 14th of June. The year range was limited to the last ten years. Neither area of knowledge nor kind of document were limited, since we do not want to filter the applicability of the research. The query retrieved 182 results.

To have a broader view from the evolution of the research area we expand the year range to 15 years with a total of 209 publications, as shown in Figure 2. Regarding the evolution of publications over time, it can be seen that the last three year contains 69% of the scientific production in the area. Citation evolution also reflects the attention this topic is gaining in the last years, as shown in Figure 3. Different groups of research in a wide range of countries and institutions are active in research areas that expose ethics requirements within AI context. Regarding publications by country, as shown in Figure 4, the USA stand out with over two times India’s and UK production; we also highlight the presence of Brazil with the 10th production. Even though USA has the aforementioned high production, Italy has a higher concentrated, since the top institution in production, as show in Figure 5. Figure 4 and 5 only depict the top 10 countries and institutions, respectively.

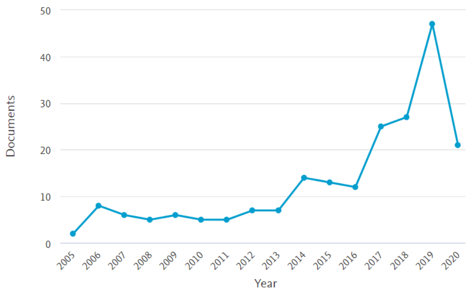


Figure 2. Number of publications in Scopus by year.

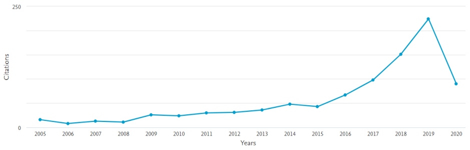


Figure 3. Evolution of amount of citations.

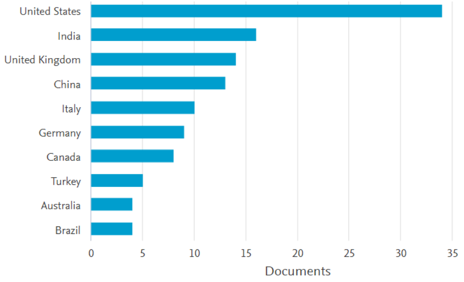


Figure 4. Records by country.

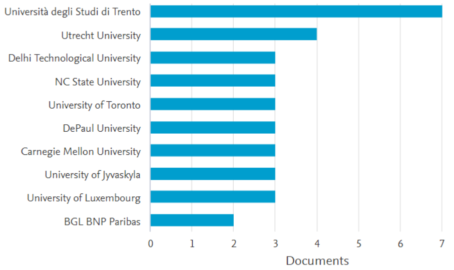


Figure 5. Records by institution.

To detail the topography of the research area we applied the author co-citation map analysis, as shown in Figure 6. This map has roughly 3 clusters, indicating three group of authors sustaining this research area. The most critical author are Giorgini, P. who proposed the ”STS-tool” in 2012, which helps engineers to specify Socio-technical Security Requirements through social commitments. This kind of work are directly related to Mylopoulos, J.’s works in security and privacy requirements specifying ownership, permission, delegation and others. Other important cluster are represented by Cleland-huang, J. with studies in automated tracing, product listing and recommendation, such works are related to data security and transparency. Finally, the cluster represented by Tan, T., Singh, R. and Sun, Z. can be considered as the biometric branch of the research topography, with studies in iris recognition and localization, closely related to ethics and policy of biometric systems because such systems are associated with spoofing attacks and have implications in security requirements.

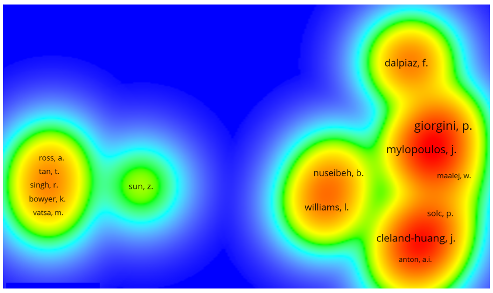


Figure 6. Co-citation density map. Source: Scopus. Map generated by VOSViewer.

To answer what are the tendencies of this research area we applied the Bibliographic Coupling map analysis. This analysis, as shown in 7, is based on the number of citation counted within the set of selected articles, mapping the currently most important sources for research considering only the last three years. In overall, 15 clusters can be observed.

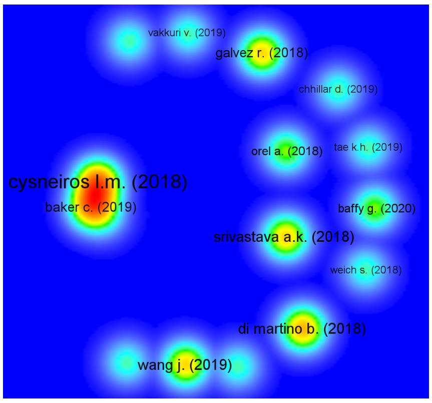


Figure 7. Bibliographic coupling of authors density map. Source: Scopus. Map generated by VOSViewer.

The most prominent cluster represented by Cysneros (2018) and Backer (2019), in general, presenting approaches for classification and identification of security requirements. This reveals the great concern of the research in privacy threats and security risks, eventually related to applications such autonomous driving, linguistic analysis. This cluster is followed by Di martino's (2018) cloud services research and Wang (2019), each of them with 4 accumulated citations.

To perform the integrating model we conducted an analysis based on the Descending Hierarchic Classification that proposes to identify main classes on requirements in ethics in AI research. This analysis examined 182 abstracts and found 1019 text segments. The text segments were organized into five classes: Class 1 with 22.8%, Class 2 with 14.8%, Class 3 with 22.6%, Class 4 with 22% and Class 5 with 17.8%, as shown in Figure 8. In Class 1 the most representative work is Intelligent Software Mining with Business Intelligence Tools for Automation of Micro services in SOA: A Use Case for Analytics [27], where authors explored the automated process of mining software engineering data for useful business applications. Analysing the words that represent the class such as Decision, Organization, Business, Agile, System, Development and Principle, it is possible to note that they point to the use of AI-based systems as a decisive factor in organizations. Correspondingly the class is called AI in Agile and Business for RE.

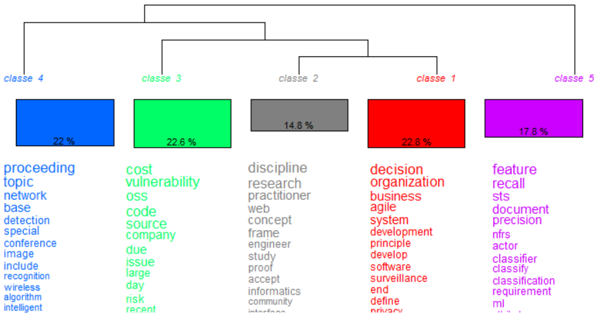


Figure 8. Descending hierarchic classification dendrogram. Generated with the use of Iramuteq.

In Class 2 the most representative work is Requirements We Live By [28] where the author called researchers attention for reflections upon RE as a discipline in light of new technologies as AI. Analysing the words that represent the class such as Discipline, Research, Practitioner, Web, Concept, Frame and Engineer, it can be seen that they all point towards RE research as a discipline and a concept, along with practitioners and engineers in this field with a theoretical approach. Accordingly the class is called RE in Theoretical Context. In Class 3 the most representative works are: 1) Enhancing Offshore Safety Culture Through Continuous Management of Barriers and Success Paths [29], in this study decision support systems are explored to assess safety culture and control room management in the context of offshore operations. 2) The correlation between OSS project and organizational performance [30] in this research authors proposed policy directions to improve awareness of Open Source Software in their company. Analysing the words that represent the class such as Cost, Vulnerability, OSS, Code, Source and Company, it can be seen that they all point towards quality factors as cost, vulnerability, of source code, in the context of companies using OSS or not. Accordingly the class is called Quality Requirements.

In Class 4, the most representative work is actually an abstract of a Proceeding focused on the research of sensors, and, analysing the words that represent the class such as Proceeding, Topic, Network, Base, Detection, Special, and Conference, it can be seen that they all point towards Proceedings and Conferences without much connection to RE and AI, mainly due to the fact that no area of knowledge was filtered. Accordingly the class is called Proceedings and Conferences.

In Class 5, the most representative works are: 1) Hidden in plain sight: Automatically identifying security requirements from natural language artifacts [31] where the authors used ML techniques to develop a tool-assisted process taking as input a set of natural language artifacts to aid requirements engineers in producing a more comprehensive and classified set of security requirements. 2) Extracting Quality Attributes from User Stories for Early Architecture Decision Making [32] aimed to automatically identify quality attributes from user stories (functional user requirements). 3) Automatically Classifying Functional and Non-functional Requirements Using Supervised Machine Learning [33], in this study authors used supervised ML to automatically classify requirements as functional and non-functional. Analysing the words that represent the class such as Feature, Recall, STS, Document, Precision, NFRS and Requirement, as well as the most representative works, it is observed that they converge to automatically classify requirements by using AI techniques. Thus, the class is called AI in Requirements Engineering.

4.1. GitHub

In order to discover a satisfactory string to be used in the query for GitHub repositories, a preliminary research was conducted. The number of results for each search pattern, are presented in Table 1.

Table 1. Number of repositories found with different search strings.

String	Number of repositories
“ethical artificial intelligence”	19
ethic artificial intelligence	36
ethical artificial intelligence	461
artificial intelligence ethics	501
ethics in ai	589

With the use of the string “*ethical artificial intelligence*” with quotation marks produced a set of 19 repositories only about curated lists of courses, books, video lectures and papers about AI. While using the string *ethic artificial intelligence* without quotation marks resulted in 36 repositories, it is too small for proper analyses. Thus, exploration of GitHub README files was performed using the string *ethics in ai* without quotation marks in the search field of the Corpus Retrieval web based tool, providing us with a set of 589 README files, the largest amount of repositories retrieved between the tested search strings.

Further, a manual reading of the selection of all 589 README files from the corpus was performed. We found out the following filtered categories: 486 (82.5%) Reference lists (e.g., curated lists, lectures and course materials, assignments, conferences materials, software lists), 78 (13.2%) AI applications for end users, 21 (3.5%) tools for implementing AI ethics. Only 4 (0.7%) were not found. Analyzed data is available in https://zenodo.org/record/4284782.

From the tools found in GitHub that actually assist ethical AI implementation we highlight. 1) InterpretML: a package that incorporates ML interpretability techniques that explains blackbox systems, thus it is possible to understand the reasons behind individual predictions. 2) Deon: is a command line tool to assist in easily adding ethics checklist to a data science projects and 3) TransparentAI: wraps a mature tool named SHAP to give simple visualization solutions for AI-based systems, in face of TrustworthyAI requirements by the European Union, associated to transparency of model and datasets applied in some project. 4) XAI: designed based on the 8 Principles for Responsive Machine Learning, enables analysis and evaluation of data models, having AI explainability as main player.

To get a qualitative insight from GitHub corpus, we filtered out supplementary forms and unmeaningful active forms (e.g., td, https, href, nbsp, javascript and format types). Following, a word cloud is generated to visualize most frequent words present in the filtered corpus, as shown in Figure 9. From Figure 9 we can observe that most of repositories available are related to learning aspects of ethics in AI (e.g., courses and books), and concerning file repositories (e.g., assignments and courses materials). With the totality of README files found (589), an analysis on Iramuteq is performed

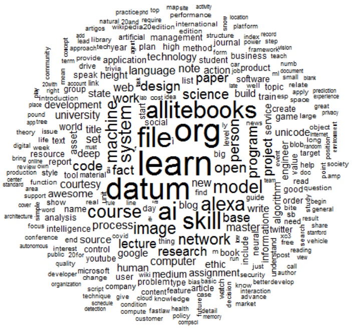


Figure 9. Word cloud from github project README files.

regarding requirements in publicly available projects in GitHub. Using some of candidate RE patterns provided by Portugal et al. [34] we are able to extract information regarding requirements in AI projects and then relate them to ethical principles. Table 2 shows the relation between verbs – RE patterns – and their frequencies.

Table 2. Frequency of selected RE patterns.

Verb	Frequency
allow	528
enable	509
provide	1379
able	583
create	2093

We highlight some requirements found, and classify them according to Floridi et al. [15] principles: 1) beneficence, 2) non-maleficence, 3) autonomy, 4) justice and 5) explicability. Regarding **allow** RE pattern: **1)** "allowing for user input and classifications ... users will have more control over how they are being represented and classified". Principle: autonomy; **2)** "allowing you to also review its code for unknown adversarial bias". Principle: explicability; **3)** "allowing calculation of relative importance of varying features and attributes to customers". Principle: explicability; **4)** "allow robots to perform complex tasks like navigating an environment and detecting pedestrians". Principle: non-maleficence; **5)** "should allow the user to identify global contextual and collective outliers artificial adversary". Principle: autonomy; **6)** "will allow the algorithm to correctly determine the output for inputs that were not a part of the training data". Principle: not concerning ethics; **7)** "allow the application for read only access to google drive the account profile and offline access on behalf of one of your google accounts". Principle: autonomy; **8)** "should be allowed to unblur or identify the patient they are speaking to". Principle: not concerning ethics; **9)** "algorithms are allowed to take certain protected categories into account when making predictions". Principle: justice, non-maleficence; **10)** "algorithms which are used to predict loan eligibility or risk of recidivism should not be allowed to base predictions off of gender or race". Principle: justice, non-maleficence; **11)** "a new approach to training machine learning models that decentralizes the training process allowing for users privacy to be maintained by not needing to send their data to a centralized server". Principle: autonomy.

Regarding **enable** RE pattern: **1)** "skater is a unified framework to enable model interpretation for all forms of model to help one build an interpretable machine learning system often needed for real world use cases ... towards to enabling faithful interpretability for all forms models". Principle: explicability, justice; **2)** "enables developers or auditing entities to discover and test for unwarranted associations between an algorithm's outputs and certain users sub-populations identified by protected features explanation explorer". Principle: explicability, justice; **3)** "it improves the interference of manoeuvres reducing rate of false positives in the detection of lane change manoeuvres and enables the exploration of situations in which the surrounding vehicles behave dangerously not possible if relying on safe generative models such as idm". Principle: non-maleficence, justice; **4)** "enabling safe and effective learning in autonomous driving model based real life methods that employ constraints to keep the agent close to the training data for the model". Principle: non-maleficence; **5)** "working group on AI for COVID-19 project enable the secure and rapid transfer of information about hospital bed capacity and availability of critical resources during public health emergencies". Principle: beneficence, justice; **6)** "enables easy visualization and analysis of models and comparison across training algorithms". Principle: explicability, autonomy; **7)** "which enables users to view explanations of individual instances under different contexts". Principle: explicability, autonomy; **8)** "enabling users to seamlessly test models for several bias and fairness metrics in relation to multiple population sub groups". Principle: explicability, autonomy, beneficence, justice; **9)** "enable researchers and practitioners alike to quickly grasp capabilities and limitations of a particular explainable method one explanation does not fit all". Principle: explicability, autonomy.

Regarding **provide** RE pattern: "to provide explanations and analyse the fairness and robustness of black box models". Principle: explicability, beneficence. Finally, we argue that, although tools are available in GitHub, they are centered in providing explicability, serving as a tool to show a black box ML algorithm, or other AI technique, as a white box. Revealing, again, that this principle is an enabler for the others, that is, removing it would destroy all the concept of ethical AI. Being able to audit, detect anomalies, understand how things are working behind the curtains is the main aspect involving ethics in AI that should be protected and broadened.

5. Final Remarks

There is an increasing discussion in academia and in industry on ethics in AI. Several researchers have an agreement on theoretical principles that guide ethics in AI. However, such principles are not easy to implement, thus, there is a need to translate ethical principles into practice. This study attempted to provide an overview analysis on the topic of ethics in AI both in a scientific database and a repository of open source projects. Although it is seen from our analysis that the scientific community focus on AI methods or tools to assist requirements engineers on requirements analysis, in fact GitHub projects benefits from scientific community, and vice-versa, as published papers commonly open-source their codes, and scholars explore available tools in their researches. It was produced five classes from our scientific database analysis in Scopus, 1) AI in Agile and Business for RE (22.8%), 2) RE in Theoretical Context (14.8%), 3) Quality Requirements (22.6%), 4) Proceedings and Conferences (22%), 5) AI in Requirements Engineering (17.8%).

The requirements found within GitHub through the used method may not reflect only results for AI projects: a considerable number of projects in GitHub are not focused on AI-based systems, many being actually courses, references or such that reference the area. Projects found were not necessarily built with the considerations of those principles, but we classified the requirements according to them because they are already being used in the literature to address ethical requirements in AI as in [9]. It is uttermost that the main focus of regulations should be situated on the work of the software developer, in consonance with Antonov and Kerikmäe [35]. Implement ethics in AI is not an easy task, sustained dedication is needed. However, it is crucial to internalize ethical AI development urgency. Highlighted tools found in GitHub such as InterpretML, Deon, XAI and TransparentAI are some examples of active development projects from the Open Source Community that can enable transparency, a highly recurrent principle from guidelines. The combined energy of both scientific and open-source sources fosters an enhanced debate and stimulates progress towards AI ethics in practice.

References

[1] A. Vogelsang and M. Borg, "Requirements engineering for machine learning: Perspectives from data scientists," in *2019 IEEE 27th International Requirements Engineering Conference Workshops (REW)*, pp. 245–251, IEEE, 2019.

[2] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, "From what to how: An initial review of publicly available ai ethics tools, methods and research to translate principles into practices," *arXiv preprint arXiv:1905.06876*, 2019.

[3] V. Vakkuri, K.-K. Kemell, J. Kultanen, and P. Abrahamsson, "The current state of industrial practice in artificial intelligence ethics," *IEEE Software*, 2020.

[4] A. Hagerty and I. Rubinov, "Global ai ethics: A review of the social impacts and ethical implications of artificial intelligence," *arXiv preprint arXiv:1907.07892*, 2019.

[5] T. N. Y. Times, *Cambridge Analytica and Facebook: The Scandal and the Fallout So far*, Apr. 2018. `https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html`.

[6] Reuters, *Amazon scraps secret AI recruiting tool that showed bias agaisnt women*, Oct. 2018. `https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G`.

[7] K. Siau and W. Wang, "Artificial intelligence (ai) ethics: Ethics of ai and ethical ai," *Journal of Database Management (JDM)*, vol. 31, no. 2, pp. 74–87, 2020.

[8] T. Hagendorff, "The ethics of ai ethics–an evaluation of guidelines," *arXiv preprint arXiv:1903.03425*, 2019.

[9] R. Guizzardi, G. Amaral, G. Guizzardi, and J. Mylopoulos, "Ethical requirements for ai systems," in *Canadian Conference on Artificial Intelligence*, pp. 251–256, Springer, 2020.

[10] A. Aberkane, "Exploring ethics in requirements engineering," Master's thesis, 2018.

[11] A. McNamara, J. Smith, and E. Murphy-Hill, "Does acms code of ethics change ethical decision making in software development?," in *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 729–733, 2018.

[12] I. S. Association, *The IEEE global initiative on ethics of autonomous and intelligent systems*, Apr. 2019. `https://standards.ieee.org/industry-connections/ec/autonomous-systems.html`.

[13] E. C. High-Level Expert Group on Artificial Intelligence, *Ethics guidelines for trustworthy AI*, Apr. 2019. `https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai`.

[14] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[15] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, *et al.*, "Ai4peoplean ethical framework for a good ai society: Opportunities, risks, principles, and recommendations," *Minds and Machines*, vol. 28, no. 4, pp. 689–707, 2018.

[16] L. Rothenberger, B. Fabian, and E. Arunov, "Relevance of ethical guidelines for artificial intelligence - a survey and evaluation," in *27th European Conference on Information Systems - Information Systems for a Sharing Society, ECIS 2019, Stockholm and Uppsala, Sweden, June 8-14, 2019* (J. vom Brocke, S. Gregor, and O. Müller, eds.), 2019.

[17] A. for Computing Machinery, *ACM code of ethics and professional conduct*, Aug. 2018. `https://www.acm.org/binaries/content/assets/about/acm-code-of-ethics-booklet.pdf`.

[18] V. Vakkuri, K. Kemell, M. Jantunen, and P. Abrahamsson, ""this is just a prototype": How ethics are ignored in software startup-like environments," in *Agile Processes in Software Engineering and Extreme Programming - 21st International Conference on Agile Software Development, XP 2020, Copenhagen, Denmark, June 8-12, 2020, Proceedings* (V. Stray, R. Hoda, M. Paasivaara, and P. Kruchten, eds.), vol. 383 of *Lecture Notes in Business Information Processing*, pp. 195–210, Springer, 2020.

[19] H. Belani, M. Vukovic, and Z. Car, "Requirements engineering challenges in building ai-based complex systems," in *27th IEEE International Requirements Engineering Conference Workshops, RE 2019 Workshops, Jeju Island, Korea (South), September 23-27, 2019*, pp. 252–255, IEEE, 2019.

[20] B. Kostova, S. Gurses, and A. Wegmann, "On the interplay between requirements, engineering, and artificial intelligence," in *Joint Proceedings of REFSQ-2020 Workshops, Doctoral Symposium, Live Studies Track, and Poster Track co-located with the 26th International Conference on Requirements Engineering: Foundation for Software Quality (REFSQ 2020), Pisa, Italy, March 24, 2020* (M. Sabetzadeh, A. Vogelsang, S. Abualhaija, M. Borg, F. Dalpiaz, M. Daneva, N. Condori-Fernández, X. Franch, D. Fucci, V. Gervasi, E. C. Groen, R. S. S. Guizzardi, A. Herrmann, J. Horkoff, L. Mich, A. Perini, and A. Susi, eds.), vol. 2584 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020.

[21] V. Vakkuri, K.-K. Kemell, and P. Abrahamsson, "Eccola–a method for implementing ethically aligned ai systems," *arXiv preprint arXiv:2004.08377*, 2020.

[22] S. Larsson, M. Anneroth, A. Felländer, L. Felländer-Tsai, F. Heintz, and R. C. Ångström, "Sustainable ai: An inventory of the state of knowledge of ethical, social, and legal challenges related to artificial intelligence," 2019.

[23] A. M. Mariano, A. C. B. Reis, L. dos Santos Althoff, and L. B. Barros, "A bibliographic review of software metrics: Applying the consolidated meta-analytic approach," in *International Joint conference on Industrial Engineering and Operations Management*, (Cham), pp. 243–256, Springer, Springer International Publishing, 2018.

[24] G. A. A. Prana, C. Treude, F. Thung, T. Atapattu, and D. Lo, "Categorizing the content of github readme files," *Empirical Software Engineering*, vol. 24, no. 3, pp. 1296–1327, 2019.

[25] R. L. Q. Portugal, M. A. Casanova, T. Li, and J. C. S. do Prado Leite, "GH4RE: repository recommendation on github for requirements elicitation reuse," in *Proceedings of the Forum and Doctoral Consortium Papers Presented at the 29th International Conference on Advanced Information Systems Engineering, CAiSE 2017, Essen, Germany, June 12-16, 2017* (X. Franch, J. Ralyté, R. Matulevicius, C. Salinesi, and R. J. Wieringa, eds.), vol. 1848 of *CEUR Workshop Proceedings*, pp. 113–120, CEUR-WS.org, 2017.

[26] R. L. Q. Portugal, H. Roque, and J. C. S. do Prado Leite, "A corpus builder: Retrieving raw data from github for knowledge reuse in requirements elicitation," in *Proceedings of the 3rd Annual International Symposium on Information Management and Big Data - SIMBig 2016, Cusco, Peru, September 1-3, 2016* (J. A. Lossio-Ventura and H. Alatrista-Salas, eds.), vol. 1743 of *CEUR Workshop Proceedings*, pp. 48–54, CEUR-WS.org, 2016.

[27] D. P. Wangoo, "Intelligent software mining with business intelligence tools for automation of micro services in soa: A use case for analytics," in *2020 7th International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 98–101, IEEE, 2020.

[28] B. Nuseibeh, "Requirements we live by," in *2019 IEEE 27th International Requirements Engineering Conference (RE)*, pp. 1–1, IEEE, 2019.

[29] W. R. Nelson, A. I. Ahluwalia, *et al.*, "Enhancing offshore safety culture through continuous management of barriers and success paths," in *SPE Health, Safety, Security, Environment, & Social Responsibility Conference-North America*, Society of Petroleum Engineers, 2017.

[30] J.-B. Kim and H. Park, "The correlation between oss project and organizational performance," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 4S2, pp. 72–78, 2019.

[31] M. Riaz, J. King, J. Slankas, and L. Williams, "Hidden in plain sight: Automatically identifying security requirements from natural language artifacts," in *2014 IEEE 22nd International Requirements Engineering Conference (RE)*, pp. 183–192, IEEE, 2014.

[32] F. Gilson, M. Galster, and F. Georis, "Extracting quality attributes from user stories for early architecture decision making," in *2019 IEEE International Conference on Software Architecture Companion (ICSA-C)*, pp. 129–136, IEEE, 2019.

[33] Z. Kurtanović and W. Maalej, "Automatically classifying functional and non-functional requirements using supervised machine learning," in *2017 IEEE 25th International Requirements Engineering Conference (RE)*, pp. 490–495, IEEE, 2017.

[34] R. L. Q. Portugal and J. C. S. do Prado Leite, "Extracting requirements patterns from software repositories," in *2016 IEEE 24th International Requirements Engineering Conference Workshops (REW)*, pp. 304–307, IEEE, 2016.

[35] A. Antonov and T. Kerikmäe, "Trustworthy ai as a future driver for competitiveness and social change in the eu," in *The EU in the 21st Century*, pp. 135–154, Springer, 2020.