An Agile Framework for Trustworthy AI

Stefan Leijnen[1] and **Huib Aldewereld**[1] and **Rudy van Belkom**[2]
and **Roland Bijvank**[1] and **Roelant Ossewaarde**[1]

Abstract.

The ethics guidelines put forward by the AI High Level Expert Group (AI-HLEG) present a list of seven key requirements that Human-centered, trustworthy AI systems should meet. These guidelines are useful for the evaluation of AI systems, but can be complemented by applied methods and tools for the development of trustworthy AI systems in practice. In this position paper we propose a framework for translating the AI-HLEG ethics guidelines into the specific context within which an AI system operates. This approach aligns well with a set of Agile principles commonly employed in software engineering.

1 INTRODUCTION

Artificial intelligence has the potential to support the accomplishment of some of human society's deepest problems [25]. With AI systems, we are able to investigate options that we would normally consider naive, but which could unexpectedly lead to major breakthroughs, such as the recent discovery of two new planets from data from the NASA archives [6] or visual recognition algorithms outperforming human doctors at diagnosing breast cancer [16]. Simultaneously, AI has the potential to disrupt societies through its impact on existing economic and social structures. Risks involved in the deployment of this powerful technology include a reduction of control of the digital systems, the introduction of biases based on gender or race, and a radical increase of societal inequality, or accordingly to some, the end of the human race [7]. These risks may cause this key technological development that can aid humanity, to be nullified by fear and distrust.

1.1 Ethics Guidelines for Trustworthy AI

To exploit opportunities and prevent threats it is important to increase the trustworthiness of AI and monitor its development. Ethical guidelines are required for this. To this end, ethics codes and principles have been published [10] by governments (e.g. [24]), private sector (e.g. [18]) and research institutes (e.g. [1]). Despite the clear agreement that AI should be ethical, there is debate about what constitutes 'ethical AI' and what ethical requirements and technical standards are needed to achieve it [14].

The Ethics Guidelines for Trustworthy AI were presented by the AI-HLEG on April 8^{th} 2019 [12]. The report builds on a draft that was published in December 2018, on which over 500 comments were made following an open consultation. The guidelines state seven key requirements that AI systems must meet in order to be considered trustworthy: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) societal and environmental well-being, and (7) accountability. A 'Trustworthy AI Assessment List' was developed in order to determine to what extent an application meets the requirements. The AI-HLEG guidelines can be considered a primary ethics directive for the development of trustworthy AI systems, due to the thought and expertise that went into creating it and the support of the European Commission (EC) for a human-centered approach to AI.

As a step towards ensuring compliance with this directives, conceivably through future legislation, the EC issued a Whitepaper on Artificial Intelligence [5] on February 19^{th} 2020. In this whitepaper the EC sets out proposals for promoting the development of AI in Europe while ensuring fundamental human rights are respected. An important part of this white paper is the proposal to create a prior conformity assessment for high-risk AI applications, based on the Ethics guidelines of the AI-HLEG. This legal framework should address the risks for fundamental rights and safety.

1.2 Ex Ante Evaluation vs Continuous Design

With the choice for a prior conformity assessment, the EC opts for an ex ante approach to aligning systems with the ethics guidelines, i.e. it should be determined in advance whether AI applications are able to meet the guidelines. Particularly when exposing society to high-risk AI applications such as facial recognition or deep fake algorithms, thoughtful risk assessment and caution action is required [5]. However, in order for AI systems to conform ex ante with these guidelines, methods and tools need to be developed that allow these guidelines to be integrated during the development of the AI system. For example, full transparency of a decision model that has been trained using machine learning methods may not be feasible, but during the development cycle an understanding of what constitutes 'sufficient transparency' can emerge, given a functional, ethical and technical context. So in theory, it is possible to check off every key requirement of the list by conforming with each requirement to some extent thereby passing the ethical evaluation. In practice, the context and value of the key requirements become explicit in designing, developing, training, testing and using AI systems.

Moreover, although the seven key requirements are considered to be equally important [12] trade-offs can arise when integrating guidelines into practice. Beyond evaluating these trade-offs and documenting the considerations as suggested in the AI-HLEG guidelines, methods and tools are required to deal with these trade-offs during development. The term 'trade-off' suggests a compromise, but a de-

[1] Research group Intelligent Data Systems, HU University of Applied Sciences Utrecht, The Netherlands. Corresponding author: stefan.leijnen@hu.nl. All authors contributed equally.
[2] STT Netherlands Study Centre for Technology Trends, The Hague, The Netherlands.

sign choice does not necessarily need to constitute a zero-sum game where increasing the value of one element naturally decreases another. Colloquially, this is captured in Benjamin Franklin's statement that "Those who would give up essential liberty, to purchase a little temporary safety, deserve neither liberty nor safety." What presents itself as a balancing act between two juxtaposed values, may be a missed opportunity in design.

Similarly to valuing ethical requirements individually, this balancing act between conflicting key requirements also becomes explicit in the context where they are applied. Consider for example the apparent trade-off between privacy (key requirement 2) and safety (key requirement 3): a customs officer using a smart scanner to inspect your travel bag at an airport will typically yield a different level of cooperation than a baker using the same scanner to inspect your shopping bag. While we may value safety over privacy at an airport, in a different context we may feel the same act violates our privacy. Different privacy standards may exist within contexts, as do the design opportunities to prevent conflicts or trade-offs between values. Design decisions would thereby be ideally made in the most specific context where values can be maximized in relation to each other.

Finally, the design context matters because trustworthiness is a human concern. Although technology can be labeled as trustworthy, it is the context in which this technology is placed that reflects on whether the technology is trustworthy and for the benefit of the human actors concerned.

We therefore argue for integrating the ethical key requirements into the process of developing AI systems, making a translation of the general but abstract AI-HLEG ethics guidelines into specific but custom product requirements that shape or constrain the design (*ethical drivers*), taking into account concerns from direct and indirect human actors, and advocating that ethical design should take place in an applied context.

2 AN ETHICS DESIGN APPROACH

There are different approaches to software design, ranging between the Waterfall model where the design is determined up front, to Agile approaches where there is continuous planning, designing and learning during development [2]. As pragmatics dictate, some larger projects start with iterations involving more design work, gradually transitioning to iterations that involve more implementation work (cf. Fig.1).

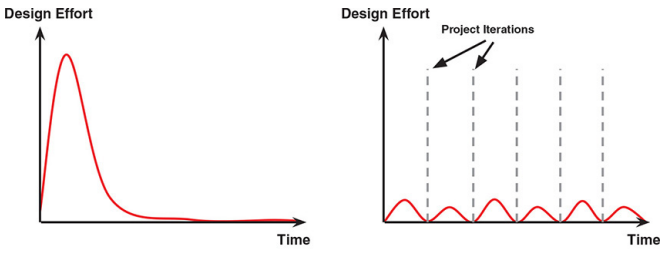

Figure 1. Two contrasting approaches to system design: the Waterfall model (left) and the Agile model (right) [4].

Ethical drivers are special in two ways. First, when designing a product, ethical considerations are usually not directly represented in the form of a stakeholder. Therefore, it will typically involve special focus to guard that the guidelines are given proper attention during design activities. Second, the ethical perspective (cf. [19]) typically affects the product as a whole, which implies that decisions based on ethical drivers are best made in the early stages of software design, when more global decisions are made. Both the extra design effort and the global impact of ethical drivers create potential tension with Agile development approaches.

Existing approaches to ethical development (e.g., Value-sensitive design (VSD) [11] or Design for Values [23]) focus primarily on the Waterfall model, where the design is known before engineering starts. High-level requirements are consequently translated into design requirements, for instance through the use of value hierarchies [22]. These abstract design requirements are the starting point of the engineering development cycle, and are then typically translated into product features. It should be noted that VSD focuses on high-level value conflicts, and tries to determine the possible (technical) design space allowed by the combination of values of the stakeholders. Tensions between the value conceptions at later stages of the development should be evaluated according to VSD, but it is not clearly described how this should be done, as [13] already noticed VSD is lacking a clear ethical perspective.

Also, ethical considerations made in the abstract design can be (unintentionally) overturned in the design decisions made by the development team during the development sprints. Sufficient support for ethical design has to be provided to the design phase as well. This can be tackled in keeping sufficient focus on the ethical guidelines during the development process.

Moreover, current engineering practice focus on functional and non-functional aspects of the system from a system point of view. Creating Human-centered AI systems requires keeping a clear focus on the points of view concerning human actors. Apart from the already mentioned Value-sensitive design method, there are few to none development methodologies that focus on this form of ethical engineering. In principle, Agile methodologies such as Scrum allow for ethical engineering processes, as they are flexible enough to include any steps regarding value discovery, or weighing stakeholder concerns. Although the discovery of non-functional requirements in Agile software processes has received attention from other researchers [9], it is still unclear how to prevent the neglect of the ethical concerns of some stakeholders.

2.1 Agile Development

Agile development focuses on short and iterative cycles to improve agility and flexibility in the development process [2]. Scrum [21] is the dominant Agile methodology, especially in the segments served by smaller development teams [3]. Development in Scrum is done in short cycles (*Sprints*) in which concrete parts of the design are implemented into a usable product, while keeping a backlog of work that remains to be done. The main driver of development within sprints are so-called *User Stories*.

User Stories express desired system functionality from the perspective of a particular user, expressing a particular desire in a given context. Two examples are shown in figure 2. A User Story is typically formulated in the following template:

"**As a** ... [actor] **I want** ... [functionality] **in order to** ... [desire] **given** ... [context]."

User Stories have several characteristics:

- they allow large projects to be divided into smaller parts that can be developed independently;

- they are typically short and contain only those development steps that can be made in a short amount of time (i.e. days rather than weeks);
- they are especially useful for projects where requirements and desires change rapidly, or where these are misunderstood; and
- they facilitate time estimation of tasks.

The high-level design is created and formulated in *Epics*. An Epic is a User Story that is too big to fit in a single development cycle [20], and therefore has to be broken down (either at the start of the project, or in a later iteration of development) into smaller, more concrete User Stories. This breakdown is done through a process of User Story Mapping [17]. In later phases of development, User Stories can be broken down further into even more contextualized features, if deemed necessary. Finally, at the start of each Sprint the priority of the User Stories is determined, for instance, by means of planning poker [15]. Thereby the developers resolve the importance of each User Story, determining in which order they will be implemented in the project. This priority determines which User Stories are implemented in the current Sprint, and which User Stories will have to wait for a future Sprint.

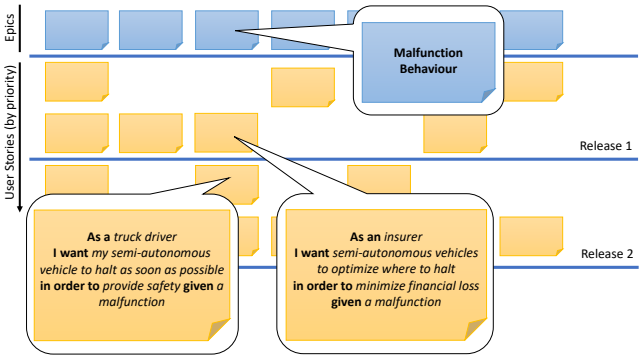

Figure 2. Example layout of a backlog for the development of a semi-autonomous truck, with high-level Epics and prioritized User Stories.

At any stage during a project, the relation between the Epics and the User Stories is kept visible on the Scrum board (see figure 2). The top row shows the Epics that have been identified for the current project; below each Epic are the related User Stories, indicating which User Stories are deemed necessary to create a releasable version of the project.

2.2 An Agile Approach to Trustworthy AI

It is a key agile principle to postpone design decisions until as late as possible, allowing for just-in-time but well-informed decisions. However, some overarching requirements, such as those sourced from AI-HLEG guidelines, are better formulated once all major design decisions are made. This requires careful consideration of where and how to apply these requirements.

Figure 3 shows the translation of high-level Epics to contextualized User Stories in action while involving the AI-HLEG ethics guidelines at each step in the process.

User stories take a central place as a design artifact in Scrum. The structure that is commonly used for user stories in Scrum closely aligns with the requirement for a specific and contextualized design

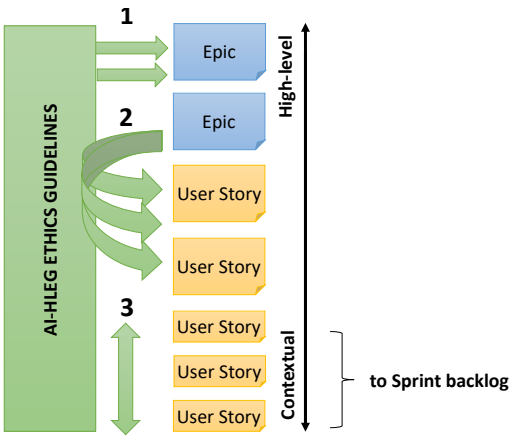

Figure 3. At several steps during the development of AI systems, the Agile design process should be informed by the HLEG-AI ethics guidelines: (1) creating Epics; (2) User Story mapping; (3) prioritization of User Stories.

element, in order to make the (comparative) value of the ethical requirements explicit; this is provided by the '**given** ... [context]' part of the User Story template. In user stories, the roles of the direct stakeholders (e.g. users, subjects) and indirect stakeholders (e.g. society, future generations) can be naturally represented in the '**As a** ... [actor]' part of a User Story. The ethical requirement is naturally referred to, as all considerations that add value in user stories, in the goal-part of the user story, described by the '**in order to** ... [desire]'-phrases.

Scrum also offers methods for trade-offs between guidelines to be mapped. We propose to investigate the use of planning poker [15] as a method to weigh, compare and prioritize user stories (as in step 3 of figure 3), rather than an estimate of the size of user story, the ethical impact could be considered to award points and open a discussion between developers. This approach also allows for technical knowledge to inform the ethical discussion, i.e. a developer may be aware of a recent technology developed capable of resolving two (or more) conflicting values, or the reverse could happen, where the technological roadmap is driven by the need to resolve two conflicting values.

As a next step, the Scrum approach to designing trustworthy AI systems according to the ethics guidelines needs to be investigated further, empirically tested and adapted based on the results. In the next section we show a preliminary example how the ethics guidelines could be operationalized in Scrum.

2.3 Example of Ethics Guidelines in Agile Design

One of the design concerns of AI systems is their behavior in exceptional circumstances, such as when component failures or adverse external factors cause malfunctions. Consider the context of the semi-autonomous truck that is capable of assisting its human driver by making decisions in emergency situations. In our scenario, the truck must make a decision about bringing the vehicle from a state of moving to a full stop in the event of a system malfunction. The risk of an adverse event regarding the safety of the driver is smaller for a halted truck than for a truck that is moving along with traffic. Therefore, the default strategy of the AI may be to stop the truck as soon as possible in case of malfunction. However, it may be in the interest of some stakeholders to override this default strategy. For example, an insurer may want the truck to stops at a point where the risk of economic

loss of the truck is minimized, whereas it may be in the best interest of the driver to stop as soon as possible regardless of the hazard this creates for other drivers of other vehicles or costs of repairing the truck.

As step 1 in figure 3 shows, the AI-HLEG guidelines can be used to evaluate the ethical impact for each of the stakeholders when the specifying the Epics. This may yield new insights of conflicting values and requirements, or new stakeholders (such as insurers or other drivers) that the design should take into account. During User Story mapping, one User Story may put forward the driver's perspective, and another may take the insurer's perspective; the ethical considerations for each individual User Story can be aligned with the guidelines as soon as it is created, cf. step 2. Conflicting interests emerge most clearly at this stage, prompting a resolution mechanism in order to prioritize the items, cf. step 3. One such resolution mechanism is Scrum planning poker, where the impact of the key requirements is comparatively evaluated and discussed by the development team.

In the emergent design approach that underlies agile software development, general requirements are stepwise translated into scenarios of increasing specificity in the development cycles. The default strategy (*stop as soon as possible*) will be revisited and reconsidered when User Stories are created from Epics. This leads to a continuous need to reapply the AI-HLEG guidelines. Conflicting values will emerge as the set of user stories grows, exemplified by the sidecar position of the AI-HLEG ethics guidelines in figure 3. The iterative application of guidelines stands in contrast with the typical large upfront designs of VSD and other waterfall strategies which are considered an anti-pattern for Scrum [8]. The arrows in Fig. 3 exemplify that the ethical guidelines play a role at different levels of design abstraction: at the highest level, when Epics are produced (arrow 1); during subsequent detailing, when User Stories are produced (arrow 2); and at the lower levels, when User Stories are prioritized and the Sprint Backlog is organized (arrow 3).

3 DISCUSSION

The analysis provided and methods proposed in this position paper are part of ongoing applied research towards operationalizing ethical guidelines for AI into the practice of developing AI systems. More (empirical) research is needed in order to validate and extend the Agile framework for trustworthy AI presented here.

We propose that describing requirements as user stories have the advantage of placing the human actor central in designing the *how*. This approach bridges the gap from the moral and regulatory functions of ethics guidelines to the daily practice of implementing software, by carrying over the key ethics requirements to the fine-grained context where the *what*, *why*, and *for whom* can take on a meaning that is not evident at the abstract level of an AI system's comprehensive design.

We argue that some conflicting ethical concerns only become more visible at the lower levels of design abstractions. Because Scrum cyclically combines high level design with low level design activities in each sprint, the relevance of the HLEG guidelines remains constant throughout the development process. It is also at this implementation level where the trade-offs between different key requirements can be weighed against each other, and sometimes resolved informed by technological possibilities.

REFERENCES

[1] AI Asilomar. Principles. future of life institute, 2018.

[2] Kent Beck, Mike Beedle, Arie Van Bennekum, Alistair Cockburn, Ward Cunningham, Martin Fowler, James Grenning, Jim Highsmith, Andrew Hunt, Ron Jeffries, et al. The agile manifesto, 2001.

[3] A. Begel and N. Nagappan, 'Usage and perceptions of agile software development in an industrial context: An exploratory study', in *First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007)*, pp. 255–264, (Sep. 2007).

[4] Humberto Cervantes and Rick Kazman, *Designing Software Architectures: A Practical Approach*, SEI Series in Software Engineering, Addison-Wesley, Boston, 2016.

[5] European Commission, 'Whitepaper on artificial intelligence - a european approach to excellence and trust', *B-1049 Brussels*, (2020).

[6] Anne Dattilo, Andrew Vanderburg, Christopher J Shallue, Andrew W Mayo, Perry Berlind, Allyson Bieryla, Michael L Calkins, Gilbert A Esquerdo, Mark E Everett, Steve B Howell, et al., 'Identifying exoplanets with deep learning. ii. two new super-earths uncovered by a neural network in k2 data', *The Astronomical Journal*, **157**(5), 169, (2019).

[7] Charles J Dunlap Jr, 'Accountability and autonomous weapons: Much ado about nothing', *Temp. Int'l & Comp. LJ*, **30**, 63, (2016).

[8] V. Eloranta, K. Koskimies, T. Mikkonen, and J. Vuorinen, 'Scrum anti-patterns – an empirical study', in *2013 20th Asia-Pacific Software Engineering Conference (APSEC)*, volume 1, pp. 503–510, (Dec 2013).

[9] Weam M Farid, 'The normap methodology: Lightweight engineering of non-functional requirements for agile processes', in *2012 19th Asia-Pacific Software Engineering Conference*, volume 1, pp. 322–325. IEEE, (2012).

[10] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar, 'Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai', *Berkman Klein Center Research Publication*, (2020-1), (2020).

[11] Batya Friedman and David G Hendry, *Value sensitive design: Shaping technology with moral imagination*, Mit Press, 2019.

[12] AI High-Level Expert Group, 'Ethics guidelines for trustworthy AI', *B-1049 Brussels*, (2019).

[13] Naomi Jacobs and Alina Huldtgren, 'Why value sensitive design needs ethical commitments', *Ethics and Information Technology*, 1–4, (2018).

[14] Anna Jobin, Marcello Ienca, and Effy Vayena, 'Artificial intelligence: the global landscape of ethics guidelines', *arXiv preprint arXiv:1906.11668*, (2019).

[15] Viljan Mahnič and Tomaž Hovelja, 'On using planning poker for estimating user stories', *Journal of Systems and Software*, **85**(9), 2086–2095, (2012).

[16] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg C Corrado, Ara Darzi, et al., 'International evaluation of an ai system for breast cancer screening', *Nature*, **577**(7788), 89–94, (2020).

[17] Jeff Patton and Peter Economy, *User story mapping: discover the whole story, build the right product*, O'Reilly Media, Inc., 2014.

[18] Sundar Pichai, 'Ai at google: our principles', *Google blog*, (2018).

[19] Nick Rozanski and Eóin Woods, *Software Systems Architecture*, Addison Wesley, Upper Saddle River, NJ, 2 edn., 2011.

[20] Kenneth S Rubin, *Essential Scrum: A practical guide to the most popular Agile process*, Addison-Wesley, 2012.

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

[22] Ibo Van de Poel, 'Translating values into design requirements', in *Philosophy and engineering: Reflections on practice, principles and process*, 253–266, Springer, (2013).

[23] Jeroen Van den Hoven, Pieter E Vermaas, and Ibo Van de Poel, *Handbook of ethics, values, and technological design: Sources, theory, values and application domains*, Springer, 2015.

[24] Cédric Villani, Yann Bonnet, Bertrand Rondepierre, et al., *For a meaningful artificial intelligence: Towards a French and European strategy*, Conseil national du numérique, 2018.

[25] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini, 'The role of artificial intelligence in achieving the sustainable development goals', *Nature Communications*, **11**(1), 1–10, (2020).