Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

A participatory data-centric approach to AI Ethics by Design

Anne Gerdes

To cite this article: Anne Gerdes (2022) A participatory data-centric approach to AI Ethics by Design, Applied Artificial Intelligence, 36:1, 2009222, DOI: 10.1080/08839514.2021.2009222

To link to this article: https://doi.org/10.1080/08839514.2021.2009222

APPLIED ARTIFICIAL INTELLIGENCE
2022, VOL. 36, NO. 1, e2009222 (778 pages)
https://doi.org/10.1080/08839514.2021.2009222

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

A participatory data-centric approach to AI Ethics by Design

Anne Gerdes

Department of Design and Communication, University of Southern Denmark, Kolding, Denmark

ABSTRACT

Data-driven artificial intelligence (AI) based on machine learning techniques (ML) has increasingly become an enabler in critical societal domains. However, the introduction of ML systems is often accompanied by unjustified, biased, and discriminated outcomes with severe consequences for the individuals affected. Consequently, in recent years value-based design methods have sought to anticipate and mitigate moral wrongdoing by drawing attention to ethical and epistemic challenges related to the design of AI systems. This article presents a participatory data-centric approach to AI Ethics by Design by promoting and refining insights from contributions within the family of value-sensitive design methods. The approach provides a practicable outlook on addressing epistemic and ethical issues related to data activities in early ML development project stages. Hence, the article seeks to enhance opportunities for ethically informed AI design by stressing the need for bridge building to cultivate a shared understanding among system developers and domain experts about a given data domain and its relatedness to a specific practice.

ARTICLE HISTORY
Received 3 June 2021
Accepted 17 November 2021

Introduction

AI is ubiquitous in everyday life as algorithmic decision-making has skyrocketed over the last decade due to the availability of enormous datasets, computer processing speed, and cloud storage. Low risk use scenarios combined with almost unlimited access to behavioral data from social media, sensors, and online interactions, have accelerated the performance of machine learning models. But the success has come at the expense of an awareness of issues concerning, e.g., validation, verification, and explainability. In low-risk scenarios, failure is not a catastrophe; profiling models improve by learning from mistakes. Therefore, recommender systems routinely encourage users to respond to misaligned profiles by clarifying their preferences. However, mainstream, data-driven AI based on ML techniques has increasingly been adopted in wide areas of critical societal domains to inform decision-making within justice, law

CONTACT Anne Gerdes gerdes@sdu.dk Department of Design and Communication, University of Southern Denmark, Universitetsparken 1, Kolding DK-6000, Denmark

This article has been republished with minor changes. These changes do not impact the academic content of the article.

enforcement, health care, education, and public administration. As a result, prediction and decision-making systems have entered domains with the potential to affect peoples' lives seriously (Ananny 2016; Elish and Boyd 2018; Mittelstadt et al. 2016), making it crucial to anticipate ethical and epistemic issues at an early developmental stage.

Data is a relatively limited resource in critical domains compared to datasets from mundane online activities, which means that training datasets for ML modelling must be carefully selected and curated to ensure algorithmic accuracy. For example, we do not have substantial datasets for child abuse. Therefore, creating a predictive risk model that may prevent child abuse with an acceptable threshold for false positives requires substantiated descriptions of indicators and necessitates the introduction of proxies for variables that signify characteristics that correlate with child abuse (Gillingham 2016).

The typical focus of ML projects, is on the developer's responsibility to select and curate data and take steps to mitigate risks from biased training datasets. However, this article emphasizes that data activities in ML projects can be improved by making room for dialogue to grow a mutual understanding between domain experts and ML developers about the domain being scrutinized. Such collaborative data activities can be facilitated by a bridge builder with inter-disciplinary competencies in the fields of computer science and ethics. Hence, we promote and refine existing contributions to ethical AI design and suggest a participatory data-centric approach to AI Ethics by Design.

The article is organized as follows: section two presents the field of value-based design and directs attention to shortcomings in the endorsed approach to AI development, namely the value sensitive design (VSD) methodology. Contemporary trends in VSD primarily focus on clarifying value issues at a conceptual level, which leaves AI practitioners with no clear operational guidelines besides high-level "best practice" principles that are hardly informative to system developers with limited knowledge of philosophy. The article emphasizes the importance of positioning domain experts at center stage and reestablishing genuine inter-disciplinary efforts. We extract best practices in the field of ethical AI design, exemplified by two prototypical approaches to AI VSD, namely the AI for Social Good VSD (AI4SG-VSD) method (Floridi et al. 2018, 2020; Umbrello and van de Poel 2021) and the AI Ethics by Design framework (Brey and Dainow 2021). Against this backdrop, section three suggests a participatory data-centric approach to AI Ethics by Design as a manageable remedy to pro-actively address epistemic and ethical issues in ML systems. Finally, section four concludes the article.

Introducing Selected Family Members in the Field of AI Value Sensitive Design

The early days' “design turn” in IT ethics suggested a pro-active stance to IT system development, prioritizing doing “front-loaden” ethics rather than post-hoc analyses of IT systems (van den Hoven 2007). Historically, Friedman et al. proposed VSD as an ethically focused alternative to related user-centered and participatory design methods in the fields of Human Computer Interaction (HCI) and Computer Supported Cooperative Work (CSCW) (Friedman and Kahn, 2003). The underlying intentions of VSD were promising and encouraged humanists and social scientists to engage in genuine inter-disciplinary design activities by collaborating with computer scientists and software engineers. Nissenbaum coined the term “engineering activism” to summarize the roles of the humanities and social sciences in IT system development:

> Humanists and social scientists can no longer bracket technical details—leaving them to someone else—as they focus on the social effects of technology. Fastidious attention to the before-and-after picture, however richly painted, is not enough. Sometimes a fine-grained understanding of systems—even down to gritty details of architecture, algorithm, code, and possibly the underlying physical characteristics—plays an essential part in describing and explaining the social, ethical, and political dimensions of new information technologies.
>
> (Nissenbaum 2001)

Early work in VSD design stood out from other user-centered design methods by including direct and indirect stakeholders and presenting a so-called tripartite iterative and integrative VSD method paying particular attention to ethical and human values in design. The VSD process is iteratively organized around three types of investigations: conceptual, empirical, and technical. Typically, the starting point is at the conceptual level. Here, relevant values are found by applying philosophical inquiries, and stakeholders affected by the system are being identified. At the empirical level, designers engage with direct and indirect stakeholders to include their voices and value perspectives. The level of the technical investigation provides for the translation of values into technical requirements. In recent years, however, the tripartite iterative VSD method has not delivered on a research agenda informing technical investigations. Instead, typically, VSD emphasizes a philosophically rooted evaluation of existing technologies and tools (Bozdag and van den Hoven 2015) or the conceptual exploration of values at play (Hayes, van de Poel, and Steen 2020) together with stakeholders (Aizenberg and van den Hoven 2020; Zhu et al. 2018). VSD shows less commitment to collaborations with developers at the technical level to proactively embed values in IT systems.

Technical investigations are downplayed, viz. collaboration between system developers, social scientists, and humanists concerning how value insights can be translated into operational, technical guidelines. This is a pity because embedding values in AI systems requires attention to the "gritty details," as emphasized two decades ago in the quotation above by Nissenbaum. On that note, Veale and Binns (2017) call for interdisciplinarity in the field with an obligation to seek to comprehend the research perspectives of others. Consequently, AI developers need to grow awareness of ethical challenges in contextual settings. At the same time, social scientists and humanists must "grapple more rigorously with the technical proposals placed on the table and ensure that critiques with operational implications reach the ears of the computing community" (Veale and Binns 2017, 13). If social scientists and humanists draw back from technical involvement, they run the risk of "setting the 'moral background' for conversation about ethics and technology as being about abstract principles" (Tubella and Dignum 2019). Correspondingly, facilitating stakeholder-driven value elicitation requires attention to how values can be turned into technically realizable design requirements. Otherwise, in a worst-case scenario, the field of ethical AI design risks framing the contribution from social science and the humanities as a silo-derived stakeholder informed moral philosophical input, which is not easily made computationally operationalizable.

If social scientists and humanists commit to collaborating in genuine interdisciplinary settings they can provide conceptual clarification of values in a manner, which is informative to the technical design (see, e.g., Tubella and Dignum 2019). Moreover, when humanists and AI developers join forces, humanists may avoid the pitfalls of flawed, mythically hyped ideas about the challenges of AI. For example, the mainstream misconception that ML algorithms are self-learning algorithms and black boxes. But ML algorithms are well understood and fully engineered by hand. These algorithms produce models, i.e., multi-layered neural networks, which can be non-explainable. It is the cocktail of straightforward algorithms, such as optimization algorithms (goal: to minimize error and maximize prediction accuracy) working on complex data that produce complex and sometimes inscrutable models.

As an illustrative example of mythical hype, Elish and boyd (2018) draw attention to the widespread misunderstandings concerning the technical superiority of Cambridge Analytica. Among other stories, this case made Hillary Clinton claim that she had been a victim of 'weaponized' technology and lost the election to Trump on that account. Clinton's observation stands in stark contrast to her own campaign's experiences with an algorithmic decision-making system named Ada after the renowned Ada Lovelace, who invented the first programming language. During the campaign, Clinton learned the hard way that Ada was not a precise laser weapon:

> No one really knew exactly how Ada made her decisions, but they did know that she was a powerful computer program analysing an unimaginable amount of data. So, they trusted her . . . After the loss, Bill pointed his middle finger at the data wonks who put all their faith in a computer program and ignored the millions of working class voters who had either lost their jobs or feared they might lose their jobs. In one phone call with Hillary, Bill reportedly got so angry that he threw his phone out the window of his Arkansas penthouse.
>
> (Smith 2018, 5)

According to Elish and boyd (2018), the hyped assumption that Cambridge Analytica can be held accountable for manipulating elections has been significantly moderated after ML experts entered the scene and clarified the state of the art within the field.

A Closer Look at Selected AI VSD Methods

Against the backdrop of the challenges arising from misunderstandings about AI—primarily due to lack of knowledge about what is computationally feasible—we move on to present two prototypical approaches to ethical AI design represented by the Sienna project (Brey and Dainow 2021) and the AI4SG-VSD method (Umbrello and van de Poel 2021). Hence, the Sienna project uses VSD as a springboard for developing an AI Ethics by Design framework that covers the development and the deployment system life cycle. The authors outline a five-layer model moving from abstract value levels into concrete design requirements, so-called "ethical requisites", which enable AI developers to proactively instantiate values while balancing system functionality. Also, they provide advice on how to anticipate ethical challenges in the first place, e.g., how to document risk mitigation measures by outlining which standards have been followed to clean data and remove bias from training datasets. Finally, they provide guidance on the organizational coordination of follow-up activities in the deployment of such systems.

Insights from the outline of the five-layer model are incorporated in a generic system development model (exemplified by agile system development) and developed into a practical stepwise guideline with instructions and tasks to follow for AI developers in the different phases of a generic system development process, viz. specification of objectives, specification of requirements, high-level design, data collection and preparation, detailed design and development, testing and evaluation. Here, we outline these phases paying particular attention to data activities.

From the outset, it is crucial to give voice to stakeholders to inform design choices when specifying objectives. This overall recommendation points to the importance of establishing conditions for mutual understanding among stakeholders and developers as a prerequisite for subsequent data explorative activities. The *specification of requirement* phase presents existing tools to support data activities. Hence, datasheets provide a thorough description of dataset characteristics and recommended use scenarios, which help dataset creators reflect on the different phases in a dataset lifecycle, i.e., data understanding, data preparation (collection, cleaning, labeling, preprocessing), modeling, evaluation, deployment, and maintenance. In a similar vein, Bendner and Friedmann (2018) introduce Data Statements as "a characterization of a dataset that provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems build on the software". Such tools for documentation may facilitate data governance and auditing. Likewise, while data sheets focus on data for model training, model cards outline the "performance characteristics" of ML-models within given domains (Mitchell et al. 2019). Model cards 'tag' ML models by specifying relevant application fields to prevent models from being uncritically transferred to domains outside their intended use-context. The *high-level design* phase also points to these tools to ensure transparency and the ethical compliance of the overall system architecture. Moreover, non-technical requirements are needed to facilitate organizational procedures that may govern the development process, e.g., assessment, data protection audits, testing regimes, documentation of data activities, and organizational procedures settling conflicts between an ethical governance authority and AI developers.

Furthermore, the *data collection and data preparation* phase is singled out as critical to ensure fairness and data accuracy. The authors warn that data typically reflect societal biases — "data can never be assumed to be accurate, representative or neutral; it must be demonstrated that it is". In addition, when selecting data and training datasets, it is pivotal to account for data protection measures, ensure transparency, and establish means to mitigate the risk from potential harmful bias. In the *detailed design and development* phase, the ethical requisites of the system design are fleshed out in detail, and the design is evaluated against the overall ethical guidelines. The *testing and evaluation* phase "uses the project's ethical requisites document to design a testing regime to test the system's compliance with its ethical requirements" (Brey and Dainow 2021, 35). Here, stakeholder input is needed to include their view on whether the ethical requisites have been appropriately integrated into the system.

Summing up, ethical guidelines for implementation, deployment, and use are described with suggestions for follow-up activities, such as risk assessments and methods to ensure ethically compliant deployment and to adequately address potential unethical changes in the embedded "ethical characteristica of the system". Follow-up activities are essential, as the system may transform during the deployment phase if for example, it is deployed with new datasets (Brey and Dainow 2021, 37). The AI Ethics by Design framework reaches beyond system functionality and includes attention to deployment and organizational processes affected by the re-organization of workflows as well as broader organizational cultural issues. However, it seems odd that Brey and Dainow denote VSD as "the globally recommended approach for AI development" because their AI Ethics by Design methodology develops a framework in which VSD is collapsed into a supplementary or 'add-on' tool to system development methods.

A generic, VSD-based approach to ethical AI design is reflected in work by Umbrello and van de Poel (2021). Their approach provides conceptually applicable guidelines by modifying VSD to accommodate specific value challenges related to AI. The authors build their modification of VSD, the AI4SG-VSD methodology, on the observation of two specific, interdependent challenges associated with AI, which VSD does not account for. First, AI systems may acquire knowledge in ways that are opaque or incomprehensible to humans and thereby challenge the legitimization of decision-making and obscure accountability. Second, AI systems may go astray and learn in ways that transgress or overrule the values embedded in the system in the first place. The emergence of biased learning paths with negative consequences may be subtle, unforeseeable, and inscrutable to humans and thus imply uncontrollable outcomes leading to moral wrongdoing. Therefore, it becomes pivotal not only to anticipate ethical problems to avoid moral wrong (the main claim in VSD), but also to ensure that AI systems are beneficial and promote values that contribute to social good. Here, the authors introduce the sustainable development goals (SDG) as a globally shared conception of valuable social ends. Finally, like the Sienna Project, they extend the VSD approach to include the AI system deployment life cycle.

To begin with, they suggest replacing the VSD heuristic value list, containing 13 prototypical values that often deserve attention in ICT system design (Friedman and Kahn, 2003), with "a set of AI-specific design principles." Here, they point to the values distilled by the EU High-Level Expert Group on the Ethics of AI, namely, *respect of human autonomy, prevention of harm, fairness*, and *explicability*. User-driven bottom-up value elicitation might, of course, be relevant. However, according to Umbrello and van de Poel, a generic list from "an AI-specific entity" is needed to avoid overlooking prototypical AI-ethical issues. Therefore, their point of departure is seven ethical principles rooted in the values mentioned above and seen as necessary to create beneficial AI4SG (Floridi et al. 2020). Hence, *falsifiability and incremental deployment* concern system reliability and AI trustworthiness and imply that it should always be possible to formally verify system states.

Contrary to ordinary SW-systems with a fixed SW-architecture, self-learning systems may change their models to align with, e.g., new data in a dynamic environment. As such, it is crucial to be able to investigate whether or not the system satisfies given constraints, that is, the AI-system operates as expected (Russell, Dewey, and Tegmark 2015). Consequently, in safety-critical domains, incremental deployment backed up by falsifiable hypotheses is pivotal.

Safeguards against the manipulations of predictors mitigate perils from algorithmic gaming as well as from overestimating the role of non-causal patterns in data analytics. *Receiver-contextualized intervention* addresses the need to balance human autonomy and decision power with machine interventions in a respectful and supportive manner. For example, profiling tools may enhance data-driven contextualization of future intervention based on my revealed preferences, as long as this is done without intruding on my autonomy. Also, it is important that "users can ignore intervention, but accept subsequent, more appropriate interventions [. . .] later on" (Floridi et al. 2020). *Receiver-contextualized explanation and transparent purposes* imply avoiding opaque operations that are inscrutable to humans and further stress the need for explainable interfaces to convey domain-relevant, adequate explanations.

Moreover, as data-driven AI systems feed on behavioral data, *Privacy protection and data subject consent* are essential topics of attention in AI4SG. Also, *situational fairness* needs to be accounted for to face the challenge from biased data training sets, which may otherwise lead to biased decision-making resulting in unfair outcomes, stigmatization, or discrimination. Finally, *Human-friendly semanticisation* refers to AI-mediated enhancement of human sense-making with attention to eliminating random AI-driven meaning-making, which does not align with human sense-making.

In light of the observations mentioned above concerning shortcomings in VSD and the introduction of AI-specific values, the authors outline the AI4SG-VSD method consisting of four iterative phases. Here, the *context analysis* follows VSD by drawing attention to direct and indirect stakeholders and socio-technical problems embedded in a given use context. The second, and non-empirical phase, *value identification*, distinguishes between SDG values promoted by the design, AI-specific values respected by the design, and conceptual exploration of contextually relevant values. In the third phase, *design requirements* are formulated based on insights from phases one and two. To translate abstract values into design requirements, the authors suggest using a value hierarchy tool to visualize "potential design pathways". For example, at the top level of the hierarchy, the value "nonmaleficence" may be translated into norms concerning "privacy protection and data subject consent" (found in the list of design principles) at a lower level in the hierarchy. At the lowest practical level, design requirements are stated, such as "clear terms of use" and "pseudonymization of data subject information (e.g., GDPR 2016/679[recital 28]" (Umbrello and van de Poel 2021). The authors state that "Visualisation helps determine how related values can produce technical design requirements", and the *value elicitation* phase aim "to help designers begin to design *for* various values more effectively".

The fourth phase, *prototyping*, explores the value-resilience of the system design in practice. Traditionally, user-centered and participatory design methods use mock-ups and prototypes to investigate system affordances and system impacts with end-users in the usage context (Derboven et al. 2010). Low-fidelity mock-ups (e.g., drawings, Lego, sketches, simple models) are useful at an early stage of the system development process in tandem with, e.g., future workshops to open up the design space. High-fidelity prototypes are helpful at a later stage to explore functionality-related problems and the broader contextual and organizational implications of the system design.

Umbrello and van de Poel (2021) understand prototyping not as a technique but elevate it to an independent design phase, which assists in anticipating the design's ethical and societal effects and reveals ways in which the design may influence values. Prototyping is viewed as a means to mitigate value problems and risks during the system development process and the systems' entire life cycle. The authors suggest that it "is [. . .] advisable to go through a number of trials for such apps [(ed.) an illustrative example of a Covid 19-contact tracking app], scaling up from very small-scale testing with mock-ups to test settings of increasing size (not unlike what is done in medical experiments with new drugs)".The assumption is that prototype testing may lead to new design iterations both during the design process and during the deployment of the system. However, prototypes, and most certainly mock-ups, lack the precision needed to properly inform AI ethical system design and handle issues concerning ML maintenance during the deployment phase. As noticed below by Sculley et al. (2015), relying on these tools is presumably not the remedy you are looking for to meet fundamental challenges in AI design and bring clarity to the table:

> It is convenient to test new ideas in small scale via prototypes. However, regularly relying on a prototyping environment may be an indicator that the full-scale system is brittle, difficult to change, or could benefit from improved abstractions and interfaces. Maintaining a prototyping environment carries its own cost, and there is a significant danger that time pressure may encourage a prototyping system to be used as a production solution. Additionally, results found at small scale rarely reflect the reality at full scale.
>
> (Sculley et al. 2015, 6)

Moreover, the outline of the AI4SG-VSD method does not explicitly address the importance of bringing different fields of expertise onboard, which is a pity as inter-disciplinarity is a prerequisite for a successful AI4SG-VSD design process. It is hard to imagine a team of AI developers who would be able to follow the method without guidance from a bridge builder. Engaging in such multi-disciplinary efforts does not, of course, rule out mono-scholarly expertise. For example, the value elicitation phase must be philosophically informed as fairness "is not just an abstract constrained optimization problem. It is a messy, contextually embedded, and necessarily socio-technical problem and needs to be treated as such" (Veale and Binns 2017, 13). However, value-based design requirements also need to be negotiated and computationally calibrated to serve their purposes. For example, to prevent unintended value outcomes caused by opaque self-learning algorithms, Umbrello and van de Poel (2021) suggest telling a tax fraud detection algorithm "to optimize itself not only in terms of effectiveness [. . .] but also in terms of fairness". But this conceptually one-step solution is not straightforward computationally feasible. The cost of fairness is paid by accuracy because we now restrict the learning process by introducing a fairness goal: minimize error with the constraint not to disrupt a given notion of fairness. Our optimization algorithm now provides models that are fairness-compliant but less precise in predicting tax frauds. The point is that trading off accuracy for ethically behaving algorithms will adversely affect performance, which means that we introduce new negative value consequences, which we need to settle by pro-actively agreeing on the threshold for acceptable trade-offs within the given context. In this step, we must balance fairness against accuracy and negotiate an adequate level of detecting fewer tax frauds or identifying false-positive tax frauds. The task is computationally doable, but the steps are more demanding than indicated by the conceptual proposal above. Still, it is helpful to apply value hierarchies, which translate values and norms into design requirements, as a stepping-stone for further inter-disciplinary discussion about how ethical values can be embedded in the design of AI systems.

To summarize, the presented AI VSD methods draw attention to crucial AI-specific value issues concerning the system development stage and the overall system life cycle. However, the realization of the AI4SG-VSD method presupposes a inter-disciplinary setting. Still, Umbrello and van de Poel ignore the importance of inter-disciplinarity and do not address prerequisites for establishing participation. The authors' neglect of these issues is problematic, especially because they outline an abstract and highly conceptual approach facilitated by a notorious explorative tool, namely prototyping. On the other hand, although Brey and Dainow (2021) emphasize the role of inter-disciplinarity, they suggest anchoring VSD in a classical system development framework. Thereby, they underestimate the fact that AI projects differ from traditional software projects by requiring data-oriented architectures, which implies that AI project activities cannot be facilitated by prescriptions outlined in traditional system development methods:

> The main differences arise from unique activities like data discovery, dataset preparation, model training, deployment success measurement etc. Some of these activities cannot be defined precisely enough to have a reliable time estimate, some assume huge potential risks, and some make it difficult to measure the overall added value of the project. Therefore, ML deployments do not lend themselves well to widespread approaches to software engineering management paradigm, and neither to common software architectural patterns.
>
> (Paleyes, Urma, and Lawrence 2021, 14)

In this setting, this paper argues that bridge building is essential for establishing strong communication channels and facilitating data exploration activities between AI developers and domain experts. Consequently, a participatory data-centric approach to AI Ethics by Design can engage domain experts during the system design process and, furthermore, help raise organizational awareness of the challenges related to data-driven knowledge generation.

The Role of Bridge Building in Participatory Data-Centric AI Ethics by Design

Focusing on contemporary challenges in data science, Ng stresses the need to "shift our mindset toward not just improving the code but toward a more systematic way of improving the data" (Sagar 2021). Likewise, Kim et al. (2018) empirically investigate challenges and best practices among data scientists and point out that "factors that complicate data understanding include lack of documentation, inconsistent schemas and multiple possible interpretations of data labels" (Kim et al. 2018, 1031). Initiatives such as the previously mentioned datasheets for datasets provide standardized guidelines for dataset documentation, which may improve transparency and accountability and "facilitate better communication between dataset creators and dataset consumers" (Gebru et al. 2020, 1). Correspondingly, in the field of HCI, human-centered approaches to data science are starting to gain traction (Aragon et al. 2016). For example, Seidelin, Dittrich, and Grönvall (2020) show how "data may be foregrounded as an explicit element of design". The authors outline co-design activities which are didactically designed to facilitate collaborative workshops, which support domain experts in understanding and critically reflecting on data and data structures in a specific database. In this way, co-design activities in collaborative settings serve to empower domain experts and enhance data literacy. However, the authors present a case with challenges related to databased services. Here, domain experts explore and negotiate the meaning of data and data dependencies with the help of a data notation consisting of simple icons representing entities in a database. The co-design activities focus on helping domain experts understand the role of data entities and the information architecture of a database. This contribution is less helpful when tackling challenges that arise in a data-driven ML developmental context. Here, defining the right dataset for an ML project determines whether the project succeeds or fails. Nevertheless, Seidelin, Dittrich, and Grönvall (2020) provide examples of how domain expert empowerment can be facilitated by collaborative activities and stresses the importance of positioning users at the center of system development.

The life cycle of an ML project can, roughly, be described by the following activities: project definition, data collection and preparation, model training, and model deployment in practice. The suggested participatory data-centric design approach to AI Ethics by Design focuses on data activities and includes attention to the deployment stage but leaves out issues concerning model training and verification. Our focus is motivated by the fact that ML developers spend 80% of their time on data preparation (Kelleher and Tierney 2018). Likewise, an investigation of ML practitioners' real-world needs reveals a misalignment between fair ML research and challenges in ML development practice. The authors describe how research literature emphasizes the development of algorithmic de-biasing solutions at the expense of paying attention to the role of the dataset. Nevertheless, they observe that "many of our interviewees reported that their teams typically look to their training datasets, not their ML models, as the most important place to intervene to improve fairness in their products" (Holstein et al. 2019). Consequently, a thorough conceptualization of data in a given domain provides a solid foundation for the subsequent stages concerning model learning and verification.

In the context of these observations, ML developers are often faced with the challenge of interpreting and understanding nuances in data without domain knowledge to assist them (Kim et al. 2018). On top of that, end-users do not always realize the importance of communicating whether and how domain-specific data reflects reliable substantiated descriptions of a given practice. This problem is not as big in logically ordered domains as in domains characterized by less formalistic procedures. For example, in a health-care context, medical experts' labeling of model training data is pivotal to ensure the performance of a breast-cancer-detecting image analysis system, and the predictive accuracy of the ML algorithm depends on the quality of the labeling work. Here, a bottleneck situation may challenge an ML project in case it turns out to be hard to recruit experts to carry out the labeling work. But generally speaking, the health domain is characterized as a highly conceptual-ordered field with precisely defined work flows and objective diagnosis criteria, which provides a solid basis for separating reliable data in the domain. However, in other domains, such as in the field of public administration concerning social care or education, accurately labeling data is challenging as data reflects socially constructed phenomena. For instance, to decide what constitutes "educational excellence" (O'Neil 2016, 52), proxies that correlate with success have to be defined with the help of domain experts, who also need to be knowledgeable of the mechanisms behind data-driven knowledge generation to make a qualified and informed decision about which proxies to select. Likewise, substantiated descriptions of, for example, what counts as 'disadvantaged individuals' may be included in statistics without attention to the potential fuzziness of substantiation in the first place. Algorithms trained on statistics based on such substantiated data might make inaccurate predictions about vulnerable individuals resulting in unjustified interventions. Hence, “the challenge of deciding what can be quantified in order to generate useful predictions [. . .] should not be underestimated” (Gillingham 2016, 1053).

To systematically improve the data activities, we suggest participatory activities in collaboration with domain experts. Historically, participatory design practices in HCI have been inspired by the Scandinavian tradition of system development, which democratized system development, viewed users as “competent practitioners” (Greenbaum and Kyng 1991, 3), and engaged them in the development of ICT systems in workplace settings. In participatory design, the design space is opened up by establishing space for design dialogs between system developers and users. Here, inspired by Wittgenstein’s notion of language as language games, Ehn and Sjögren notice that “by understanding the design process as a process of *creating new language games* that have a family resemblance with the language games of both users and designers, we have an orientation for doing design as skill based participation” (1991, 253).

In most cases, the performance of an ML model depends not on the modeling algorithm but on the data preparation, which can be qualified if the person doing it is knowledgeable of the domain or capable of bringing in domain expertise. Hence, “a model is not better than the predictor variables input to it,” and domain knowledge “facilitates the derivation of powerful predictor variables from the existing variables. [. . .] There is simply no substitution for domain knowledge” (Nisbet, Elder, and Miner 2009, 30). This point emphasizes the importance of enhancing a shared understanding of a given data domain. Hence, a bridge builder can facilitate participatory design activities and establish space for “a meeting of language games” (Ehn and Sjögren 1991) by introducing dialogical guidelines for the acquisition of domain knowledge (Gerdes 2021).

To make room for mutual understanding, the bridge builder can facilitate collaborative workshops (Seidelin, Dittrich, and Grönvall 2020) and introduce tools, such as, conceptual sketching, mind mappings, and card sorting, to reveal the domain knowledge against which the data has to be understood. Also, the domain expert is presumably feeling out of her comfort zone in an ML development context. In this situation, starting by exploring domain expert knowledge may strengthen the domain expert's self-confidence. From the developer’s perspective, gaining insight into the domain expert’s knowledge will, at a later stage, enhance her opportunities for interpreting nuances in data. Such insight may make her more alert to whether the datasets provided accurately reflects the domain or if the datasets are inaccurate because those supplying it are unaware of the importance of delivering reliable substantiated data (Gillingham 2016).

Clarification of domain expert knowledge also serves to find common ground before entering into the more demanding exploration of values and power structures. Hence, the investigations of the values at stake aligns with the value elicitation phases in the AI VSD design approaches discussed in section two. Value elicitation activities can be introduced with the help of tools, such as value hierarchies (Umbrello and van de Poel 2021) and value sketches (Friedman, Hendry, and Borning 2017), making room for reflections on ambiguous insights concerning value issues in data. Also, semi-structured value interviews and agile consequential scanning for responsible innovation (Consequence Scanning – an agile practice for responsible innovators | doteveryone n.d.) can support value elicitation activities and lay the groundwork for discussions concerning ethical and epistemic challenges related to data-driven knowledge generation in the domain being scrutinized.

Although the AI VSD methods presented above pay attention to societal bias and introduce de-biasing strategies, these approaches do not address the overarching power structures surrounding a data environment, besides mentioning alterations in organizational workflows. It is a reasonably straightforward task to see how gender and racially-biased image datasets reflect and reinforce structural discrimination. On top of that, we emphasize that the bridge builder has an obligation to direct ML developers' attention to complex socio-political issues by raising awareness of how power mechanisms may affect ground truth in data.

In an organizational setting, data is wired into complex interacting societal and organizational structures, and the manners in which power structures influence data are subtle. As an example, of this, data behind decision-making support systems in the elderly care sector fail to capture the tacit dimension of the care practice that unfolds and thereby disvaluing that which cannot be quantified. As such, the datafication of the elderly care sector reflects a political quest for documentation and standardization, which furthermore "lends itself to an instrumental practice not supportive of growing competencies within the field of caregiving" (Gerdes 2008, 46). Moreover, it is often the case that data reliability is low because the demand for documentation negatively impacts the overall workload of caregivers, who consider such activities time-consuming, bureaucratic procedures that capitalize on primary job duties. In a recent study, Petersen, Christensen, and Hildebrandt (2020) demonstrate that algorithmic decision-making is challenged in public administration as social caseworkers are unwilling to feed sensitive data about clients into AI systems. They have moral concerns because formal documentation is decontextualized and fails to provide insight into the real-world framing of situated decision-making processes. In addition, they fear that sensitive data may adverserly frame and affect their clients' future opportunities. This concern is also shared by Eubanks, who notices that "justice demands the ability to evolve, but the digital poorhouse locks us into patterns of the past" Eubanks (2018, 19).

There are no simple techniques or quick fixes for resolving the political clash between datafication and practice. Still, the bridge builder is responsible for drawing attention to these issues to help empower ML developers anticipating the broader implications of their work. At a more general level, within the community of ML developers, we must facilitate moral self-cultivation through educational initiatives and seek to cultivate a practice that advances ethics as second nature (Gerdes 2018).

Following up on that note, we must be aware that engaging in engineering activism requires an awareness of power structures and global challenges. ML projects are typically framed to favor profit-maximizing goals at the expense of peoples' shared interest in a sustainable future. Presumably, the time has come to "revitalize participation by changing [participatory design] so that it may again become a tool to help people influence important matters in their lives" (Bødker and Kyng 2018). With this call, the HCI community is encouraged to engage in design activities for a sustainable style of living rather than contributing to "the Competition State." Phrased otherwise, this ambitious goal is also reflected in the AI4SG-VSD design method (Umbrello and van de Poel 2021).

Finally, in the deployment life cycle, the AI VSD methods presented above suggest carefully monitoring the performance of AI systems on an ongoing basis by applying impact assessment tools and by considering whether and how AI implementation causes workflow and organizational transformations. It is equally important to pay attention to the ongoing maintenance issues arising from hidden technical debt in ML systems. Hence, in the deployment of traditional software systems, the term *hidden technical debt* was coined by Cunningham as an analogical reference to fiscal debt in emphasizing how a balanced technical debt may be both unavoidable and beneficial but also cause serious problems if left unmaintained. The field of maintainable machine learning is still underdeveloped, and hidden technical data debt is particularly challenging (Sculley et al. 2015). An ML infrastructure is highly complex and requires additional ML-specific configurations on top of configurations ordinarily applied in the development of regular software systems, making ML systems more vulnerable to configuration debt than ordinary software systems. Also, besides attending to the code level, ML systems require attention to technical debt at the overall system level, especially external data dependency debt can accumulate unnoticed due to inconspicuous feedback loops between systems indirectly affecting one another (Sculley et al. 2015). The renowned case of the Google Flu Trend algorithm (GFT) successfully predicted the spread of flu across the United States. Yet, a later version of the GFT algorithm made false predictions as the improvement of the model now caused users to click more on flu-related content, thereby tunneling their search results into flu-searches, which lead to the overfitting of the GFT algorithm (Lazer et al. 2014). Moreover, as noted in the quotation below, although machine learning might perform well, there are no free lunches:

> [. . .] mature systems may have dozens or hundreds of models running simultaneously [. . .] this raises a wide range of important problems, including the problem of updating many configurations for many similar models safely and automatically, how to manage and assign resources among models with different business priorities, and how to visualize and detect blockages in the flow of data in a production pipeline.
>
> (Sculley et al. 2015, 4)

Consequently, decision-makers who consider applying ML systems should at the outset be made aware of the kind of organizational commitment it takes to manage the burden of allocating resources to taming and paying down technical debt in such systems.

Concluding Remarks

In summary, using insights from contemporary AI VSD approaches as a steppingstone, this article suggests a participatory data-centric approach to AI Ethics by Design. Here, the role of a bridge builder, with inter-disciplinary competences in computer science and ethics, is seen as essential to facilitate the cultivation of a shared understanding between stakeholders. Hence, with a data-centric perspective as the point of departure, it becomes possible to engage ML developers and domain experts in collaborative activities centered around a specific data domain and its relatedness to a given practice. Such a data-centric setting serves as a concrete scene for more abstract value reflections with the purpose of proactively addressing ethical and epistemic challenges in the design and deployment of ML systems. As such, collaborative data activities constitute a practical point of departure for doing AI Ethics by Design. Consequently, bridge building is essential for creating strong communication channels and orchestrating different types of domain-specific knowledge in ML projects. The participatory data-centric design approach to AI Ethics by Design presented here represents a manageable contribution for accomplishing just that.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Anne Gerdes http://orcid.org/0000-0002-2991-5074

References

Aizenberg, E., and J. van den Hoven. 2020. Designing for human rights in AI. *Big Data & Society* 7 (2):2053951720949566. doi:10.1177/2053951720949566.

Ananny, M. 2016. Toward an ethics of algorithms: Convening, observation, probability, and timeliness. *Science Technology and Human Values* 41 (1):93–117. doi:10.1177/0162243915606523.

Aragon, C., C. Hutto, A. Echenique, B. Fiore-Gartland, Y. Huang, J. Kim, G. Neff, W. Xing, and J. Bayer. 2016. Developing a research agenda for human-centered data science. Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion, 529–35. doi:10.1145/2818052.2855518.

Bender, E. M., and B. Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6:587–604. doi:10.1162/tacl_a_00041.

Bødker, S., and M. Kyng. 2018. ACM transactions on computer-human interaction. *Participatory Design that Matters; Facing the Big Issues* 25 (1):4:1–4:31. doi:10.1145/3152421.

Bozdag, E., and J. van den Hoven. 2015. Breaking the filter bubble: Democracy and design. *Ethics and Information Technology* 17 (4):249–65. doi:10.1007/s10676-015-9380-y.

Brey, P., and B. Dainow. 2021. Ethics by design and ethics of use in AI and robotics. The SIENNA project - Stakeholder-informed ethics for new technologies with high socio-economic and human rights impact. Accessed April 26, 2021. https://www.sienna-project.eu/digitalAssets/915/c_915554-l_1-k_sienna-ethics-by-design-and-ethics-of-use.pdf .

Consequence Scanning – an agile practice for responsible innovators | doteveryone. n.d. Accessed March 21, 2021. https://www.doteveryone.org.uk/project/consequence-scanning/ .

Derboven, J., D. De Roeck, M. Verstraete, D. Geerts, J. Schneider-Barnes, and K. Luyten. 2010. Comparing user interaction with low and high fidelity prototypes of tabletop surfaces. Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries, 148–57. doi:10.1145/1868914.1868935.

Ehn, P., and D. Sjögren. 1991. From system description to scripts for action. In *Design at work - Cooperative design of computer systems*, ed. J. Greenbaum, and M. Kyng, 241–69. Hillsdale, New Jersey: Lawrence Erlbaum Associates Inc.

Elish, M. C., and D. Boyd. 2018. Situating methods in the magic of Big Data and AI. *Communication Monographs* 85 (1):57–80. doi:10.1080/03637751.2017.1375130.

Eubanks, E. 2018 Automating Inequality - How high-tech tools profile, police, and punish the poor (New York: Picador)

Floridi, L., J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, et al. 2018 December. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines* 1–24. doi:10.12932/AP0443.32.4.2014.

Floridi, L., J. Cowls, T. C. King, and M. Taddeo. 2020. How to design AI for social good: Seven essential factors. *Science and Engineering Ethics* 26 (3):1771–96. doi:10.1007/s11948-020-00213-5.

Friedman, B., D. G. Hendry, and A. Borning. 2017. A survey of value sensitive design methods. *Foundations and Trends in Human-Computer Interaction* 11 (2):63–125. doi:10.1561/1100000015.

Friedman, B., and P. H. Kahn. 2003. Human values, ethics, and design. In *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications*, Jacko, J., and Sears, A., 1177–1201. Mahwah: L. Erlbaum Associates Inc.

Gebru, T., J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford. 2020. Datasheets for datasets. ArXiv:1803.09010 [Cs]. http://arxiv.org/abs/1803.09010 .

Gerdes, A. 2021. Dialogical guidelines aided by knowledge acquisition: enhancing the design of explainable interfaces and algorithmic accuracy. In *Proceedings of the Future Technologies Conference (FTC) 2020, Volume 1* Virtual event, eds. K. Arai, S. Kapoor, and R. Bhatia, 243–57. Springer International Publishing.

Gerdes, A. 2008. The clash between standardisation and engagement. *Journal of Information, Communication and Ethics in Society* 6 (1):46–59. doi:10.1108/14779960810866792.

Gerdes, A. 2018. An inclusive ethical design perspective for a flourishing future with artificial intelligent systems. *European Journal of Risk Regulation* 9 (4):677–89. doi:10.1017/err.2018.62.

Gillingham, P. 2016. Predictive risk modelling to prevent child maltreatment and other adverse outcomes for service users: Inside the 'black box' of machine learning. *The British Journal of Social Work* 46 (4):1044–58. doi:10.1093/bjsw/bcv031.

Greenbaum, J., and M. Kyng. 1991. *Design at work: Cooperative design of computer systems.* New Jersey: LEA.

Hayes, P., I. van de Poel, and M. Steen. 2020. Algorithms and values in justice and security. *AI & SOCIETY* 35 (3):533–55. doi:10.1007/s00146-019-00932-9.

Holstein, K., J. Wortman Vaughan, H. Daumé, M. Dudik, and H. Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need? Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–16. doi:10.1145/3290605.3300830.

Kelleher, J. D., and B. Tierney. 2018. *Data science*. Cambridge, MA: The MIT Press.

Kim, M., T. Zimmermann, R. DeLine, and A. Begel. 2018. Data scientists in software teams: State of the art and challenges. *IEEE Transactions on Software Engineering* 44 (11):1024–38. doi:10.1109/TSE.2017.2754374.

Lazer, D., R. Kennedy, G. King, and A. Vespignani. 2014. The parable of Google Flu: Traps in big data analysis. *Science* 343 (6176):1203–05. doi:10.1126/science.1248506.

Mitchell, M., S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. 2019. Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–29. doi:10.1145/3287560.3287596.

Mittelstadt, B. D., P. Allo, M. Taddeo, S. Wachter, and L. Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3 (2):205395171667967--205395171667967. doi:10.1177/2053951716679679.

Nisbet, R., J. F. Elder, and G. Miner. 2009. *Handbook of statistical analysis and data mining applications*. Amsterdam, Boston: Academic Press/Elsevier.

Nissenbaum, H. 2001. How computer systems embody values. *Computer* 34 (3):120–119. doi:10.1109/2.910905.

O'Neil, C. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown.

Paleyes, A., R.-G. Urma, and N. D. Lawrence. 2021. Challenges in deploying machine learning: A survey of case studies. ArXiv:2011.09926 [Cs]. http://arxiv.org/abs/2011.09926 .

Petersen, A. C. M., L. R. Christensen, and T. T. Hildebrandt. 2020. The role of discretion in the age of automation. *Computer Supported Cooperative Work (CSCW)* 29 (3):303–33. doi:10.1007/s10606-020-09371-3.

Russell, S., D. Dewey, and M. Tegmark. 2015. Research priorities for robust and beneficial artificial intelligence. *Ai Magazine* 36 (4):105–14. doi:10.1609/aimag.v36i4.2577.

Sagar, R. 2021. Big data to good data: Andrew Ng urges ML community to be more data-centric and less model-centric. Analytics India Magazine. Accessed May 17, 2021. https://analyticsindiamag.com/big-data-to-good-data-andrew-ng-urges-ml-community-to-be-more-data-centric-and-less-model-centric/ .

Sculley, D., G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison. 2015. Hidden technical debt in machine learning systems. In *Advances in neural information processing systems*, ed. C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, vol. 28, 2503–11. Curran Associates, Inc. http://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf .

Seidelin, C., Y. Dittrich, and E. Grönvall. 2020. Foregrounding data in co-design–An exploration of how data may become an object of design. *International Journal of Human-Computer Studies* 143:102505. doi:10.1016/j.ijhcs.2020.102505.

Smith, G. 2018. *The AI delusion*. Oxford: Oxford University Press.

Tubella, A. A., A. Theodorou, V. Dignum, and F. Dignum. 2019. Governance by glass-box: Implementing transparent moral bounds for AI behaviour. *arXiv Preprint arXiv:1905.04994*.

Tubella, A., and V. Dignum. 2019. The glass box approach: Verifying contextual adherence to values. AISafety 2019. Macao, China, August 11–12. http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-160949 .

Umbrello, S., and I. van de Poel. 2021. Mapping value sensitive design onto AI for social good principles. *AI and Ethics* 1:283–96. doi:10.1007/s43681-021-00038-3.

van den Hoven, J. 2007. ICT and value sensitive design. *IFIP International Federation for Information Processing* 233:67–72.

Veale, M., and R. Binns. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society* 4 (2):2053951717743530. doi:10.1177/2053951717743530.

Zhu, H., B. Yu, A. Halfaker, and L. Terveen. 2018. Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on Human-Computer Interaction* 2 (CSCW):1–23. doi:10.1145/3274463.