Dialogical guidelines aided by knowledge acquisition

Enhancing the design of explainable interfaces and algorithmic accuracy

Gerdes, Anne

Published in:
Proceedings of the future technologies conference (FTC) 2020

DOI:
10.1007/978-3-030-63128-4_19

Publication date:
2021

Document version:
Accepted manuscript

Citation for pulished version (APA):
Gerdes, A. (2021). Dialogical guidelines aided by knowledge acquisition: Enhancing the design of explainable interfaces and algorithmic accuracy. In K. Arai, S. Kapoor, & R. Bhatia (Eds.), *Proceedings of the future technologies conference (FTC) 2020* (Vol. 1, pp. 243-257). Springer. https://doi.org/10.1007/978-3-030-63128-4_19

Dialogical Guidelines Aided by Knowledge Acquisition – Enhancing the Design of Explainable Interfaces and Algorithmic Accuracy

Anne Gerdes[1]

[1] University of Southern Denmark, Department of Design and Communication, Universitetsparken 1, 6000 Kolding, Denmark, gerdes@sdu.dk

Abstract. Understanding expert domain knowledge may inform the design of explainable interfaces that convey comprehensible information by 'mirroring' the explanation practice of domain experts. Likewise, scrutinizing expert domain knowledge is pivotal to guarantee data quality and enhance algorithmic accuracy, by zooming in on the types of data and information that constitute relevant and reliable representations in a given domain. Against this backdrop, the paper revitalizes the field of knowledge acquisition and presents easily applicable user-centered and value-oriented dialogical guidelines to unravel domain knowledge with the aim of enhancing the design of explainable interfaces and algorithmic accuracy. While it might seem counter-intuitive to revisit the field of knowledge acquisition in the era of machine learning and deep learning, there are plenty of cases in which AI systems, trained on biased data, have led to epistemological deficiencies with morally harmful consequences. In order to improve the data preparation and modelling stage in the development of ML models, this paper suggests that AI developers could benefit from the pragmatic application of manageable dialogical guidelines aided by knowledge acquisition to cultivate shared understanding between AI developers and domain expert end users.

Keywords: Dialogical Guidelines, Knowledge Acquisition, Epistemic Opacity, Ethics, Algorithmic Accuracy, Explainable AI.

1 Introduction

In the December 1966 SCI Newsletter of the University of Washington, Kehl had a short feature article under a polemical headline inspired by Hamming's renowned sentence [1] "The purpose of computing is insight, not numbers" [2]. Here, he doubted whether, "the reams of papers generated by two high-speed printers" at the university computing center could possibly be read by anyone, and he summarized his observations by stating, "I think, as the result of over-computing, analogues to the over-kill." The problem arose not only out of the huge amount of data, but also from the fact that a scientist is dependent on a programmer to carry out his or her research. Therefore, the Department of Physiology and Biophysics had bought a Raytheon 440 computer with 8K memory with programs allowing, "real-time data logging, reduction, and display during the course of each animal experiment". Writing about his experiences with the Raytheon, Kehl noted that he, "likes the programming simplicity, but it lacks the 'hands on' capability. There are no knobs to turn, or displays to observe the dynamic responses of variables".

Despite the straightforward programming, Kehl complained about not being able to interact with the system and follow its line of reasoning. To Kehl, lack of insight into the Raytheon's inner workings constituted an annoyance. Today, as artificial intelligence applications increasingly influence our lives and play a significant role in critical social domains, epistemological opacity is heavily intertwined with moral issues, since lack of insight into how results are achieved or decisions are made may undermine the justification of outcomes. Here, Danaher [3] has coined the term 'algocracy' as a neutral term to account for ways in which algorithmic decision-making may scaffold our actions for better or worse, depending on whether or not we take the opacity concern seriously.

As a prototypical example of moral wrongdoing caused by epistemic opacity, we have witnessed racially biased decisions in the criminal justice system, most notably with the example reflected by the Compass predictive scoring algorithm for assessing the risk of crime recidivism. Here, ProPublica [4] found that black defendants with criminal records for minor offenses were incorrectly profiled as being at high risk of committing future crimes, which was twice as often as white defendants with criminal records for serious offenses.

> As risk scores move further into the mainstream of the criminal justice system, policy makers have called for further studies of whether the scores are biased. When he was U.S. Attorney General, Eric Holder asked the U.S. Sentencing Commission to study potential bias in the tests used at sentencing. "Although these measures were crafted with the best of intentions, I am concerned that they inadvertently undermine our efforts to ensure individualized and equal justice".
> [4]

One of the main challenges for Artificial Intelligence (AI) concerns the normative implications of epistemic opacity in applications that are based on machine learning (ML). As the rationale behind black-box algorithms escapes human understanding, it becomes hard to legitimize system outcomes, and this again hampers efforts to establish trust in the performance of ML systems. The lack of algorithmic transparency in AI applications negatively affects credibility, safety and accountability [5].

Recently, a variety of initiatives and methodological frameworks have emerged, which, in more or less detailed ways, draw attention to and account for values and normative concerns in AI. *The Asilomar AI Principles* [6], for example, which were negotiated at the 2017 Asilomar conference, confirm increased attention to ethical concerns within the scientific community. Similarly, the IEEE framework for ethically aligned design [7] and reviews of AI ethics tools [8] direct attention to value alignment and the development of standards for ethically informed engineering in AI. Guidance for developing agile activities for fostering ethical awareness in AI development has gained momentum [9], and companies such as DeepMind's Ethics and Society Group are increasingly addressing moral issues.

Against this setting, the purpose of this paper is to contribute to the field from a philosophically anchored and user-centered perspective. The paper provides pragmatic guidelines for unravelling domain knowledge with the aim of improving the design of explainable interfaces and enhancing algorithmic accuracy by pro-actively facilitating a shared understanding between AI developer and domain expert end users.

By revisiting the field of knowledge acquisition, the paper proposes dialogical guidelines to support the design of AI applications through conceptual clarification of domain knowledge within the field under investigation. In so doing, the paper places a particular focus on prerequisites for facilitating discussion and value-oriented reflection with domain experts in order to pro-actively determine the explicative relevance and appropriate interpretation of core concepts and terms. This would promote a shared conceptual understanding and prevent misunderstandings about the status of notions and data sources.

Section 2 outlines current studies in explainable interfaces for improving interaction, understanding, and trust in AI systems [10, 11]. This is followed in section 3 by an account of encounters with predictive risk models in the field of the social care, which is used to illustrate the epistemic and ethical challenges that arise when adopting predictive machine learning in a context in which decision-making relies on situated knowledge with ambiguous terms [12, 13]. In section 4, the paper clarifies the need to facilitate dialogues between AI developer and domain expert end users. In addition, this section positions the field of knowledge acquisition in relation to a prototypical approach adopted in AI development and data science management. This springboard creates the context necessary to revisit the field of knowledge engineering (sec. 5 and 5.1) in order to find a way to create space for user-centered dialogue and value-oriented reflections about the conceptualization of central notions and relevant terms in a given domain. Finally, the conclusion emphasizes the importance of exploring domain knowledge together with domain expert end users in order to pro-actively enhance algorithmic accuracy and to improve the design of explainable interfaces.

2 Explainable AI

Among the most prevalent and least explainable classes of algorithms, we find unsupervised deep learning algorithms, i.e. algorithms engineered without hand. Probably, the most famous is DeepMind's AlphaGo Zero program, which overnight trained itself to play Go and later went on to defeat the South Korean master Lee Sedol. Recently, a revised version of it, AlphaZero has mastered chess, Go and Shogi. As Silver et al. say, "AlphaZero replaces the handcrafted and domain-specific augmentations…with deep neural networks, a general-purpose learning algorithm, and a general-purpose tree search algorithm" [14]. This breed of deep learning system builds models from trial-and-error learning with little or no human oversight, and as learning increasingly gets delegated to the algorithm, AI developers should ensure the development of deep learning "techniques that learn more explainable features or representations" [10]. This point stresses the need to advance research in explanation generating algorithms in order to train deep learning algorithms to select 'learning routes' that not only enhance technical interpretability but also contribute to the design of explainable user interfaces as a prerequisite for fostering trust in AI [15]. If we are going to succeed in delegating important tasks to AI systems, we have to be able to understand how decisions are reached.

In a similar manner, but in the field of computer simulations, Duran and Formanek [16] introduce the notion of *essential epistemic opacity* (OOE) to account for any process whose justifying steps cannot be assessed or surveyed by a human agent. In the face of such epistemic uncertainty, they argue for the possibility of justified belief in computer simulations (viz. computational reliabilism) on condition that we can specify what it takes to rely on such computational processes. They suggest a relaxation of the demand for accessibility and surveyability and argue that,

> researchers are justified in believing the results of their simulations when… the probability that the next set of results of a reliable computer simulation is trustworthy is greater than the probability that the next set of results is trustworthy given that the first set was produced by an unreliable process by mere luck.
> [16]

To introduce a distinction between trustworthiness and randomness in this manner presumably makes sense for simulation models that depend on pre-defined methods that can be examined. But this is not the case for ML models. In their article 'Research Priorities for Robust and Beneficial Artificial Intelligence' Russel et al. [17] note that verification models are challenged by AI systems, since they do not operate with departure in "a fixed machine model" [17]. Instead, they function in a dynamic environment, which is not known in its entirety by the AI developer. As noted above, the predictive power of AI systems often escapes meaningful human interpretation, as "... there is no theory correlating input variables to things humans understand as causal or even as 'things'" [18]. On the same note, Burrell explains how even simple spam classifiers work:

> Humans … evaluate spam according to genre: the phishing scam, the Nigerian 419 email, the Viagra sales pitch. By contrast, the 'bag of words' approach breaks down texts into atomistic collections of units, words whose ordering is irrelevant. The algorithm surfaces very general terms characteristic of spam emails, often terms that are (in isolation) quite mundane and banal. My semantic analysis attempted to reconcile the statistical patterns the algorithm surfaces with a meaning that relates to the implicit strategy of the text as a whole, but this is decidedly *not* how the machine 'thinks.'
> [19]

On top of that, complex algorithmic eco-systems, consisting of communities of coders and complex coding architectures create diminishing opportunities to survey and understand system-dependencies [3]. Yet, lack of technical insight into how, for example, a deep learning network establishes a model of a given domain (interpretable

AI), does not imply that such systems cannot be explainable in the sense that we are able to understand a given result via post hoc interpretations. Here, Preece [20] notes that human explanations are post hoc interpretations anyway, implying that "[i]n a sense, seeking fully-transparent interpretations from a deep learning based AI system is holding the system to a higher standard than the one to which humans can be held" [20]. Elaborating on the same note, Lipton [21] reminds us that human decision-makers cannot reveal "the mechanisms by which the brain works".

So-called model-agnostic approaches or black box explainers may, therefore, serve to present a textual or visual explanation of outputs. For instance, such techniques may display visualizations of how features were weighted, illustrate pixel importance in relation to image recognition, or present ways to display how alternation of input data may affect the outcome [22]. However, model-agnostic approaches are not useful for deep learning techniques that rely on complex multi-layered neural networks due to their inherent complexity. As Oxborough puts it, "this algorithm is the least explainable because each hidden node represents a non-linear combination of all the previous nodes" [23].

In the context of these interrelated epistemic and normative challenges, it seems odd that algorithms have become "fundamental enablers in modern society" [24]. The observation that the best performing AI systems are often the least transparent underscores the need for research to cater for the field of explainable AI. One result of this is DARPA's explainable AI program, which seeks to understand the psychological mechanisms behind an effective explanation. This is done with the aim of enhancing interpretability at the model level in tandem with improvement in the affordances of explainable interfaces [10]. Hence, system interpretability is promoted by developing techniques that may scaffold the model construction of deep learning systems in their internal representations by forcing deep learning systems to adapt to and generate explainable learning routes for their models. Here, Deep XAI refers to "modified deep learning techniques to learn explainable features" [10].

The design of explainable user interfaces looks to usability studies and the field of HCI for inspiration. Furthermore, the approach is facilitated by a framework for explanation evaluation to measure the extent to which the explanation is understandable, helpful, and justified. Here, Gunning and Aha [10] identify two main problem areas within XAI, namely the challenges relating to data analytics and to autonomy. Hence, when relying on (semi-)autonomous systems in, for example, the context of warfare, it is essential that users can understand and foresee a system's behavior. Similarly, relying on decision support based on big data analytics requires attention to computational challenges related to the perfection of algorithmic accuracy and data quality.

It is against this backdrop that the paper identifies various means to clarify domain knowledge and to cultivate shared understanding between AI developers and domain experts to anticipate design challenges relating to epistemic uncertainty and opacity. Before moving on, however, I am going to present a case in more detail to illustrate how lack of informed dialogue between AI developers and domain experts may result in flawed outcomes with morally harmful consequences. The case that follows represents a prototypical example and reflects ethical challenges related to confusion at the developmental stage about the role of certain terms and data. In this case, a combination of initial misconceptions and subsequent lack of epistemic transparency in a ML system for prediction-based decision-making makes a bad cocktail.

3 Challenges Applying Predictive Risk Modelling in domains with epistemic uncertainty

Predictive risk modelling with preventive purposes are gaining momentum in the public sector, notable in the field of social welfare. However, skeptical voices warn about the risks of an "actuarial paradigm" in which predictive risk assessment methods replace "the passion that understands the pulse of life beneath the official version of events" [25]. The actuarial turn implies that people are not characterized by their actual behavior but on the basis of statistical methods assessing their connection to others with whom they share a profile [26]. Some would argue that predictive risk models are beneficial for vulnerable families in need of social services, who do not reach out for help to the authorities. Nevertheless, being tracked down by a set of statistics may count as an intimidating experience and so constitute an obstacle to trustful collaboration between social workers and families. In addition, "Having a high risk score may result in stigma for already stigmatized populations, and thus reinforce structural inequalities" [13] – a point also reflected in Eubanks' investigation of the effects of automated decision-making on the poor population in America, for whom she coins the term "digital poorhouses". Here, Eubanks draws attention to the way in which "poverty management" through profiling negatively distorts and impacts upon the lives of poor people, who are singled out as individuals responsible for their own misery with little or no attention paid to wider responsibilities for fighting poverty at a societal and governmental level [25]. Clearly, caution is required when applying predictive risk modelling in addressing the needs of the most vulnerable groups in society.

Focusing on challenges in the development of predictive models, Gillingham [12] emphasizes how progress in data mining has led to "the application of actuarial risk assessment without some of the uncertainties that requiring practitioners to manually input information into a tool bring" [12]. At the same time, however, he points to its downsides, namely that development of predictive tools for use in social care, more specifically in the child protection service, may not be a straightforward task because of the characteristics of the work practices involved in service activities. Here, attention to feeding systems with data is not always prioritized, and social work activities take place in a domain influenced by uncertainties, in which conclusions about a socially constructed phenomenon, such as what counts as child maltreatment, are drawn against the backdrop of a situated understanding. As a result, the introduction of substantiated descriptions of child maltreatment might turn out to be problematic. This suggests that it can become hard to establish quantifiable data of high quality within the domain. Hence, in the practice of child protection, biased decision-making processes may carry over to predictive risk models, as noted in the introductory example from the criminal justice system. Following up on that line, Eubanks notes that "[t]he digital poorhouse replaces the sometimes-biased decision-making of frontline social workers with the rational discrimination of high-tech tools" [25].

Furthermore, it appears that computer-based assessments are given quasi-oracular status and placed beyond question. Eubanks refers to case studies in the social care sector illustrating that social care workers, either on their own initiative or because they were so instructed, tended to align their scores with computer based assessments in case of conflicting scores and despite the fact that the systems did not necessarily provide flawless predictions[25]. This tendency to automation bias is well described in human factors engineering and human-computer interaction (see, e.g., [27]). Automation bias implies that people tend to favor erroneous system information without paying attention to contradictory sources of evidence, even when these are correct or signal what people would usually judge to be correct based on their professional expertise.

To summarize, creating a predictive risk model that can anticipate child maltreatment requires clearly defined terms, reliable training data sets, and reliable outcome variables to ensure that a machine learning algorithm builds a model that may precisely correlate the relations between features that construct the category of children who are at risk of being maltreated, without over fitting, i.e. producing false positives by including children not at risk. Here, both Gillingham [12] and Eubanks [25] note that in areas with sparse data material, such as terror or child maltreatment, it is not feasible to build a statistically sound model without introducing proxies for variables, or including substantiated analysis of what indicates terror or child maltreatment. In such cases, it is worth noting that "A model's predictive ability is compromised when outcome variables are subjective" [25]. Prediction-based methods may be applied more successfully in domains that are characterized by precisely defined terms and rely on empirical testable 'objective' knowledge. Moreover, introducing the term 'substantiation' in the training of an ML system invites flawed predictions about whether children are at risk of maltreatment or not. In the specific case, Gillingham notes that the system developers applied a working definition of substantiation, which was flawed due to misunderstandings – "they were not aware that the data set provided to them was inaccurate and, additionally, those that supplied it did not understand the importance of accurately labeled data to the process of machine learning" [12]. This led Gillingham to emphasize the need to gain insight into the given domain to anticipate what information it takes to develop a predictive risk model.

4 Clarifying the Need to Facilitate Deliberations about Domain Knowledge in AI System Development

As illustrated above, failing to properly account for domain knowledge may give rise to epistemic and ethical challenges, which may result in moral wrong. To clarify normative and epistemological issues relating to the ethics of algorithms, Mittelstadt et al. [28] present a conceptual map to assist structuring. They single out three epistemic concerns. Firstly, as a result of poor data quality, misguided evidence may lead to biased decision-making with unfair outcomes. Secondly, inscrutable knowledge generation processes, i.e. black box algorithms, can negatively affect the accessibility and comprehensibility of information. Thirdly, inconclusive evidence may be derived through the lack of causal evidence due to (sometimes spurious) correlations and associations between data. Next, they highlight two normative concerns regarding ways in which unfair algorithmic outcomes may foster discrimination, or have transformative consequences rooted in subtle manipulative ways to alter our preferences and choices, such as behavior and decision-shaping personalization algorithms. Finally, 'traceability' is mentioned as an overarching concern centered on the complexity of clarifying issues of accountability and moral responsibility in technologically complex settings with lack of human oversight and intervention.

Clearly, it makes good sense to maintain a distinction between normative and epistemic challenges, as normative challenges may arise independently of epistemic issues, as when unfair outcomes are produced using high-quality data and transparent algorithmic decision-making processes.

However, in the initial stages of trying to apply knowledge derived from practices whose nature is uncertain, it is often the case, in the development of ML models, that epistemic and normative issues become entangled. This paper, therefore, explores ways in which domain knowledge may appropriately be accounted for with the aim of improving both the data preparation and modeling stages in ML system development and the design of explainable interfaces.

In what follows, then, this paper turns to the field of knowledge acquisition for inspiration to present guidelines that may facilitate AI developers and end-users' deliberations about expert domain knowledge and decision-making practices.

There are, of course, guidelines for developing AI systems, which emphasize the importance of scrutinizing machine-learning algorithms, data sources, and data models to enhance system performance and ensure that systems align with the domain of application. While these methods are useful for scaffolding data science projects by giving overall guidelines concerning how to manage the different stages of a project, they do not involve a user-centered perspective or account for normative concerns arising in the wake of epistemic uncertainty. Nevertheless, they should be praised for addressing the general challenges in data science projects.

As an example, the so-called Cross Industry Standard Process for Data Mining (CRISP-DM) describing a standard life cycle in data management projects is also often referred to in data science projects. The CRISP-DM standard operates with six stages, namely, business understanding, data understanding, data preparation, modeling, evaluation, and deployment. In the first two stages, the data scientist sets out to define the project goal in alignment with business needs while making sure that relevant data are, or can be made, available. Next, the data preparation stage covers the creation and quality assessment of the data sources that are to be used for data analysis. The importance of "getting the right data for a project" cannot be underestimated, as Kelleher & Tierney point out,

> a survey of data scientists in 2016 found that 79 percent of their time is spent on data preparation…[people] imagine data scientists spend their time building complex models to extract insights from the data. But the simple truth is that no matter how good your data analysis is, it won't identify useful patterns unless it is applied to the right data.
> [29]

In the modeling stage, ML algorithms are used to sift useful patterns from data to create models, which may encode the extracted patterns. Experiments are carried out to see which algorithms provide the most accurate models when run on the data. Most often, model testing will reveal data errors, which require the data scientist to return to the data preparation stage and refine the data set. In a broader context, the evaluation stage assesses the model's performance in relation to overall business goals. Finally, the deployment phase deals with the integration of "the selected models into the business environment" [29] and undertakes regular reviews to ensure that the model is still pursuing specific business goals.

These data science guidelines address critical steps in ML development. They explicitly point to the importance of data preparation. However, the possibility of moral wrongdoing caused by biased data in algorithmic decision-making has not been dismissed. The overall orientation in ML development is empirical, and interventions are typically *post hoc* and consist of tuning models by automatically testing inferences against cases. In addition, when applying deep learning techniques, there seems to be no need for involving domain experts at all. Hence, as deep learning models are "gathering knowledge from experience, this approach avoids the need for human operators to formally specify all the knowledge the computer needs" [30]. Notwithstanding, as mentioned above, there are plenty of good reasons to re-instantiate human oversight to fight epistemological uncertainty and ensure that the rationale behind decisions can be revealed in a manner comprehensible to humans. This is especially important when systems are going to be deployed in critical social domains.

5 Revisiting the Field of Knowledge Acquisition

Knowledge engineering has been defined by Edward Feigenbaum [31] as a discipline that seeks to integrate knowledge into computer systems for the purpose of solving problems that normally would take human expertise. Moreover, Feigenbaum realized that "the problem of knowledge acquisition is the critical bottleneck problem in artificial intelligence". Feigenbaum wanted to accentuate the need for automated or "at least semi-automated" knowledge acquisition, as he complained that knowledge was acquired in "a painstaking way that reminds one of cottage industries".

For the purpose of this paper, I am going to revitalize communicatively oriented approaches in the field of knowledge acquisition (KA) [32] and translate them into the context of contemporary AI. These methods may be useful in helping AI developers to come to grips with the domain they are designing for. In most knowledge acquisition standards, the major part of the work is devoted to the analysis stage, before the implementation and testing stages. This is the case, for example, in the so-called *Knowledge Acquisition and Documentation Structuring methodology* (KADS). Here, elicitation of data concerning domain knowledge is analyzed and transformed into an "interpretative framework", which underlies the knowledge representation architecture of an expert system [33]

The field of knowledge engineering was modernized in the early 90's [34] and turned into a "modeling activity", which could be described as a cyclical process that emphasized the usage context and the given organizational setting. Here, knowledge engineering was supported by holistic methodologies, such as the *CommonKADS*, which provides six models – Organization, Task, Agent, Knowledge, Communication, Design – as guidelines for identifying, structuring, and modelling knowledge before implementing it into knowledge-based systems.

It is worth noting that back in the eighties, the field of knowledge acquisition was also motivated by the problem of explainability in expert systems design and pursued KA techniques to advance the identification and description of problem-solving tasks confronting domain experts. As Kidd tells us,

> When presented with a given AI tool…, we are still hard pressed to say what kind of problems it can be used to solve well…the emphasis has been biased too much toward a "performance" approach to AI..., where research has been oriented entirely toward achieving impressive levels of performance... Consequently, we are now in a position where we often cannot explain why any system does, or does not, perform successfully on a particular task. [33]

To sum up, KA techniques provide tools for eliciting and analyzing expert knowledge for the purpose of creating a conceptual model of their domain knowledge. Clearly, when undertaking KA activities, we face the epistemological challenge that expert knowledge most often is tacit and experience-based knowledge, situated in a given practice [35]. However, the purpose of this paper is not to translate human expertise into machine expertise. Instead, and more modestly, I turn to KA as a source of inspiration for improving evaluative dialogues about the explicative relevance and appropriateness of conceptual terms and data within a given domain of knowledge.

The aim of this paper is, then, to let insights from the field of KA supplement existing tools, to help safeguard algorithmic reliability and improve the design of explainable interfaces.

5.1 Dialogical Guidelines Aided by Knowledge Acquisition

In what follows, the paper outlines guidelines to enhance communication between the knowledge engineer, who is an expert in AI, and one or more domain experts, who understand the application field. The purpose is to raise awareness of the importance of focusing on expert domain knowledge and ways in which experts arrive at explanations as situated actors in a given practice. The paper is not claiming that one can 'dig' knowledge out of the heads of domain experts as a kind of 'jewel mining', nor that experts are able to verbalize all the tacit knowledge upon which their expertise relies. Nevertheless, it still makes sense to engage in dialogue and presumably also to include observational studies of the expert's performance situated in his or her practice.

As a first step, AI developers ought to take the time to explore whether or not it is at all feasible to employ AI solutions in practices characterized both by epistemic uncertainty and by lack of financial resources, which is typically the case in the public sector. The field of maintainable machine learning is still underdeveloped, and *hidden technical debt* may accumulate unnoticed in complex systems. An ML infrastructure is highly complex and, besides attention to the code level, it requires attention to debt at the overall system level, especially data dependency debt. Here, "it can be inappropriately easy to build large data dependency chains that can be difficult to entangle" [36]. Of course, it takes courage to confront management with the fact that they do not have the economic muscles to benefit from an AI solution, as the maintenance costs are ongoing and high. Nevertheless, this is a problem that has to be confronted, especially with the contemporary hype about AI.

With these reservations in mind, understanding expert knowledge may assist the design of explainable interfaces that convey comprehensible information about how the system reaches results and decisions by 'mirroring' the explanation practice of the domain experts. Likewise, scrutinizing expert domain knowledge is crucial to guarantee data quality and enhance algorithm accuracy, by zooming in on the types of data and information that ought to be considered relevant and reliable representations in the given domain. Here, it is important to critically assess how the decision-making practice is backed up in the professional domain one seeks to model.

As an example, a supervised learning algorithm depends on a labelled outcome variable (serving as "teacher"/reference point) to train the algorithm and build up a reliable model. The reliability of the outcome variable is pivotal to safeguarding the predictive precision of the algorithm. Consequently, it becomes essential to ensure that AI developers have the means to judge the reliability of the variables used to train the algorithm.

Despite many years of user-involvement, failure to communicate is often the biggest hurdle in system development. Probably, because communication or user investigation enterprises are typically seen as a one-way project, which objectifies the user or domain expert. Hence, successful participatory projects are often rooted in the academic field of HCI [37], whereas private companies still resort to more cost-effective and less user-centered methods. Using this as its background, this paper takes a pragmatic standpoint and presents manageable KA guidelines based on qualitative methods for exploring domain expertise and supplements them with value-oriented approaches to design.

Knowledge acquisition lies in the field of cognitive science. Still, before moving on, I would like to situate KA dialogical activities in the overarching frame of Habermas' ideas on discourse ethics, more specifically by briefly turning to his theory of communicative rationality. Here, Habermas emphasizes communication as emancipation by prioritizing communicative understanding-oriented action over strategic instrumental thinking [38].

Ideally speaking, communicative actions imply that participants are always to be seen as equals engaged in seeking mutual understanding, without manipulative or strategic purposes. In that respect, Habermas presents four validity claims as prerequisites for understanding-oriented actions – the comprehensibility of the utterance, the truthfulness of its propositional content, the truthfulness of the utterance, guaranteed by the speaker's commitment, and the validity of the utterance in relation to values and norms. Clearly, in real life settings, strategic considerations may be present, and here Habermas reminds us that it might have been otherwise.

The paper is not trying to complicate things unnecessarily, but deliberations with domain experts and end-users about their practice do not take place in a power neutral environment. As such, it is crucial to underscore the obligations of AI developers to consider the impact on their work of calculative strategic actions, which may unfold in an organizational setting.

Having raised awareness of the importance of engaging in understanding-seeking communication with the purpose of reaching consensus among equals, I return to the field of KA. Here, in order to establish fruitful conditions for the process of knowledge acquisition, the knowledge engineer must bridge the language gap between her field of expertise in AI and the end-user's field of domain expertise. There is no need of a sophisticated conceptual modelling language to interpret the end-user's expert knowledge. Rather, this can be achieved by introducing straightforward ways of examining central concepts, presuppositions, and causal complexes in the given domain. Clearly, the conceptual schemes of domain experts may be more or less elaborated. In some domains, there is a high degree of logical order, as, for example, in the health care and medical field, whereas the above-mentioned practice of social care workers is rooted in a field with a lower degree of conceptual order. To uncover the interrelatedness between central terms and the role they play in practice, dialogue can be facilitated by applying simple tools for conceptual sketching, such as visualization tools, sketches, or mind maps. To include value-oriented reflections, the field of value sensitive design has developed envisioning cards, which may be used to "raise awareness of long-term and systematic issues in design" [39]. The agile practice of "consequence scanning" also invites AI developers and stakeholders to consider intended as well as unintended consequences of a technology and to deliberate about its value implications without demanding expertise in ethics in advance [9]. The agile approach with instructive guidelines is easy to integrate into projects. However, it is a disadvantage that end users are only indirectly represented through user advocates.

Moreover, explorative interviews, both individual and group-based, may be helpful. Here, LaFrance [40] walks the reader through different types of question in explorative interviews to account for the domain expert's knowledge and professional language, including her use of heuristics. Clearly, the endeavor is purpose-driven, though LaFrance's overall observations include attention to communication as first and foremost engaging in an understanding-oriented rather than a utility-oriented process:

> Experts and knowledge engineers alike have different experiences and goals as well as preferred ways of thinking and problem-solving. If these are ignored, the knowledge engineer's understanding of a particular domain may be incomplete or significantly misconducted.
> [40]

With *Grand Tour questions,* the knowledge engineer invites the domain expert to engage in overall reflections about the domain and its scope. Here, it is important to explore the boundaries of the domain, the domain expert's perspectives and goals, and the cultural background framing her reflections. The *Grand Tour questions* are thematically structured. Hence, under the theme *Cataloging the categories,* "the expected outcome is an organized taxonomy of the expert's terms and concepts" [40]. In an AI-developmental context, this may be useful to ensure mutual understanding of the conceptualization and interdependencies between terms and concepts in the domain. The themes *Ascertaining the attributes* and *Determining the interconnections* sort out distinguishing features and interrelationships at an even more detailed level. Moreover, the dialogue may be facilitated by inviting the domain expert to explore hypothetical or counterfactual scenarios. For example, in the above-mentioned case concerning predictive risk models for preventing child maltreatment, a hypothetical case could go through a 'least likely' case considering a child in a well-functioning family rather than a typical situation and follow the steps by which the social workers would substantiate a description in dealing with such a case. Likewise, a counterfactual scenario could review a real case occurrence and introduce changes in the state of events to explore alternative versions of the given case.

I domains with highly contingent data and uncertainty, AI developers might benefit from engaging in dialogue with domain experts to establish a shared conceptual understanding to help ensure that data suppliers understand the importance of accuracy and that AI developers' working definitions of relevant terms are accurate and backed up by appropriate data sources.

6 Conclusion and Future Work

To some, it might seem counter-intuitive to revitalize the field of knowledge acquisition in the era of machine learning and deep learning, as these systems learn from experience by building complex conceptual hierarchies with little or no human intervention. Hence, there seems to be little need for knowledge acquisition activities to account for domain knowledge. However, there are plenty of cases in which AI systems, trained on biased data, have led to epistemological deficiencies with morally harmful consequences. Consequently, in order to improve the data preparation and modelling stage in the development of ML-models, AI developers could benefit from the pragmatic application of manageable dialogical guidelines aided by knowledge acquisition to cultivate mutual understanding between AI developers and domain expert end users. Hence, the suggested dialogical guidelines may provide an articulated picture of the domain under scrutiny with attention being paid to value-laden issues. Seen in this perspective, insights from the field of KA may supplement existing tools and help improve the design of explainable interfaces and enhance algorithmic accuracy. Future work in this area involves theoretical refinement as well as evaluation of the value-oriented dialogical guidelines in the practical context of ML system development.

References

1. Hamming, R.W.: Numerical methods for scientists and engineers. McGraw-Hill, New York (1962).
2. Kehl, T.: The Purpose of Computing Is Insight, Not Numbers. Simulation. 7, 280–280 (1966). https://doi.org/10.1177/003754976600700605.

3. Danaher, J.: The threat of algocracy: Reality, resistance and accommodation. Philos. Technol. 29, 245–268 (2016).
4. Larson, J., Mattu, S., Kirchner, L.: Machine bias. ProPublica. (2016).
5. Lepri, B., Oliver, N., Letouzé, E., Pentland, A., Vinck, P.: Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The Premise, the Proposed Solutions, and the Open Challenges. Philos. Technol. 31, 611–627 (2018). https://doi.org/10.1007/s13347-017-0279-x.
6. AI Principles, https://futureoflife.org/ai-principles/, last accessed 2020/03/21.
7. How, J.P.: Ethically Aligned Design. IEEE Control Syst. 38, (2018). https://doi.org/10.1109/MCS.2018.2810458.
8. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. Sci. Eng. Ethics. (2019). https://doi.org/10.1007/s11948-019-00165-5.
9. Consequence Scanning – an agile practice for responsible innovators | doteveryone, https://www.doteveryone.org.uk/project/consequence-scanning/, last accessed 2020/03/21.
10. Gunning, D., Aha, D.W.: DARPA's Explainable Artificial Intelligence Program. AI Mag. 40, 44–58 (2019).
11. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144. Association for Computing Machinery, San Francisco, California, USA (2016). https://doi.org/10.1145/2939672.2939778.
12. Gillingham, P.: Predictive Risk Modelling to Prevent Child Maltreatment and Other Adverse Outcomes for Service Users: Inside the 'Black Box' of Machine Learning. Br. J. Soc. Work. 46, 1044–1058 (2016). https://doi.org/10.1093/bjsw/bcv031.
13. Keddell, E.: The ethics of predictive risk modelling in the Aotearoa/New Zealand child welfare context: Child abuse prevention or neo-liberal tool? Crit. Soc. Policy. 35, 69–88 (2015). https://doi.org/10.1177/0261018314543224.
14. Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., Hassabis, D.: A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science. 362, 1140–1144 (2018). https://doi.org/10.1126/science.aar6404.
15. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A Survey of Methods for Explaining Black Box Models. ACM Comput. Surv. 51, 93:1–93:42 (2018). https://doi.org/10.1145/3236009.
16. Durán, J.M., Formanek, N.: Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism. Minds Mach. 28, 645–666 (2018). https://doi.org/10.1007/s11023-018-9481-6.
17. Russell, S., Dewey, D., Tegmark, M.: Research priorities for robust and beneficial artificial intelligence. Ai Mag. 36, 105–114 (2015).
18. Edwards, L., Veale, M.: Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. Duke Tech Rev. 16, 18 (2017).
19. Burrell, J.: How the Machine "Thinks:" Understanding Opacity in Machine Learning Algorithms. Ssrn. 1–12 (2015). https://doi.org/10.2139/ssrn.2660674.
20. Preece, A.: Asking 'Why' in AI: Explainability of intelligent systems – perspectives and challenges. Intell. Syst. Account. Finance Manag. 25, 63–72 (2018). https://doi.org/10.1002/isaf.1422.
21. Lipton, Z.: The mythos of model interpretability. Commun. ACM. 61, 36–43 (2018). https://doi.org/10.1145/3233231.
22. Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., Kieseberg, P., Holzinger, A.: Explainable AI: The New 42? In: Holzinger, A., Kieseberg, P., Tjoa, A.M.,

and Weippl, E. (eds.) Machine Learning and Knowledge Extraction. pp. 295–303. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-99740-7_21.
23. Oxborough, C., Cameron, E., Rao, A., Birchall, A., Townsend, A., Westermann, C.: Explainable AI: Driving business value through greater understanding. Retrieved PWC Website Httpswww Pwc Co Ukaudit-Assur.-Ai Pdf. (2018).
24. Algo Aware – informed debate on algorithmic decision-making, https://algoaware.lpweb.eu/, last accessed 2020/01/25.
25. Eubanks, V.: Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Publishing Group (2018).
26. Harcourt, B.E.: Against Prediction: Sentencing, Policing, and Punishing in an Actuarial Age. Social Science Research Network, Rochester, NY (2005).
27. Sarter, N.B., Schroeder, B.: Supporting decision making and action selection under time pressure and uncertainty: the case of in-flight icing. Hum. Factors. 43, 573–583 (2001). https://doi.org/10.1518/001872001775870403.
28. Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., Floridi, L.: The ethics of algorithms: Mapping the debate. Big Data Soc. 3, 2053951716679679 (2016).
29. Kelleher, J.D., Tierney, B.: Data Science. The MIT Press (2018).
30. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016).
31. Feigenbaum, E.A.: Knowledge engineering. The applied side of artificial intelligence. Ann. N. Y. Acad. Sci. 426, 91–107 (1984). https://doi.org/10.1111/j.1749-6632.1984.tb16513.x.
32. Neale, I.M.: First generation expert systems: a review of knowledge acquisition methodologies. Knowl. Eng. Rev. 3, 105–145 (1988). https://doi.org/10.1017/S0269888900004288.
33. Kidd, A.L., SpringerLink (Online service): Knowledge Acquisition for Expert Systems: A Practical Handbook. Springer US, Boston, MA (1987). https://doi.org/10.1007/978-1-4613-1823-1.
34. Schreiber, G., Wielinga, B., de Hoog, R., Akkermans, H., Van de Velde, W.: CommonKADS: a comprehensive methodology for KBS development. IEEE Expert. 9, 28–37 (1994). https://doi.org/10.1109/64.363263.
35. Dreyfus, H.L., Dreyfus, S.E.: Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer. Free Press (1986).
36. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., Dennison, D.: Hidden Technical Debt in Machine Learning Systems. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., and Garnett, R. (eds.) Advances in Neural Information Processing Systems 28. pp. 2503–2511. Curran Associates, Inc. (2015).
37. Bødker, S.: Third-wave HCI, 10 years later---participation and sharing. Interactions. 22, 24–31 (2015). https://doi.org/10.1145/2804405.
38. Habermas, J.: The theory of communicative action. Beacon, Boston (1984).
39. Friedman, B., Hendry, D.: The envisioning cards: a toolkit for catalyzing humanistic and technical imaginations. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1145–1148. Association for Computing Machinery, Austin, Texas, USA (2012). https://doi.org/10.1145/2207676.2208562.
40. LaFrance, M.: The Knowledge Acquisition Grid: a method for training knowledge engineers. Int. J. Man-Mach. Stud. 26, 245–255 (1987). https://doi.org/10.1016/S0020-7373(87)80094-9.