EAI Endorsed Transactions
on Creative Technologies

Research Article **EAI.EU**

On Trusting a Cyber Librarian: How rethinking underlying data storage infrastructure can mitigate risks of automation

Maria Joseph Israel[1], Mark Graves [2], Ahmed Amer[1]

[1]Santa Clara University, Santa Clara, CA 95053, USA
[2]University of Notre Dame, Notre Dame, IN 46556 USA

Abstract

INTRODUCTION: The increased ability of Artificial Intelligence (AI) technologies to generate and parse texts will inevitably lead to more proposals for AI's use in the semantic sentiment analysis (SSA) of textual sources. We argue that instead of focusing solely on debating the merits of automated versus manual processing and analysis of texts, it is critical to also rethink our underlying storage and representation formats. Specifically, we argue that accommodating multivariate metadata is an example of how underlying data storage infrastructure can reshape the ethical debate surrounding the use of such algorithms. In other words, a system that employs automated analysis may typically require manual intervention to assess the quality of its output, or demand that we select between multiple competing NLP algorithms. Settling on whichever algorithm or ensemble can produce the best results, this is a decision that need not be made a priori at all.
OBJECTIVES: An underlying storage and representation system that allows for the existence and evaluation of multiple variants of the same source data, while maintaining attribution to the individual sources of each variant, would be an example of a much-needed enhancement to existing storage technologies, especially in anticipation of the proliferation of AI semantic analysis technologies.
METHODS: To this end, we take the view of AI in SSA as a sociotechnical system, and describe a possible novel solution that would allow for safer cyber curation. This can be done by allowing multiple different annotations to coexist within a single publishing ecosystem (whether those different annotations are the result of competing algorithmic models, or varying degrees of human intervention).
RESULTS: We discuss the feasibility of such a scheme, using our own infrastructure model (*MultiVerse*) as an illustrative model for such a system, and analyse the ethical implications.
CONCLUSION: Considering an underlying storage and representation system that allows for the existence and evaluation of multiple variants of the same source data, while maintaining attribution to the individual sources of each variant within a single publishing ecosystem helps mitigate risks of automation and enhances AI (semantic) explainability.

Received on 07 August 2021; accepted on 29 November 2021; published on 01 December 2021

Keywords: Intelligent systems, AI-Human problem, semantic sentiment analysis, artificial intelligence, ethics of AI, cyber curation of scholarship.

doi:10.4108/eai.1-12-2021.172359

1. Introduction

Artificial Intelligence (AI) is increasingly touching and structuring our lives. AI helps enhance our ordinary lives with tailored news, better traffic predictions, more accurate weather forecasts, better personal time management of meetings and email communications, and cost-efficient healthcare diagnosis. Though the moral nature of the personal use of AI in these examples is largely beneficial, implicit, and nominal, its impact becomes more direct and ethically ambiguous when employed to influence one's viewpoints. For example,

[*]Corresponding author. Email: misrael@scu.edu

tailored news can be potentially problematic when it filters information to accommodate a certain marketing agenda, turning an otherwise neutral process into a mechanism of distortion. But even if the promised benefits are realized, the impact of AI is not merely limited to beneficial new behaviors. It can change how we evaluate and judge others, and not simply enact our current judgement. This is true whenever AIs are employed to essentially judge people, and is important because it takes us from the question of whether AI can do what we want, to how it can affect our needs and judgements.

Today AI is being used to predict one's ethnicity [1], credit-worthiness for a loan or mortgage [2, 3], academic grade [4], or political leanings [5]. More recently, the literal judgement in court sentencing is increasingly influenced by "risk assessment" AI with potentially dire consequences to these developments [6–8]. Although seemingly innocuous, the application of algorithms for the micro-evaluations of a text demands moral explication. The use of Natural Language Processing (NLP) algorithms varies from its application to judge the veracity of a text's authorship, to the assessment of a written work's sub-text such as the writer's sentiment in the piece [9, 10]. Because automated sentiment analysis and similar textual processing become more efficient as one increases the data available for the AI, it seems unlikely that this practice will cease, indeed it may be the only way to handle the exponential volume of automatically generated text flooding online media channels. The interconnectedness of data sets used by the AI mean there are no neutral or bias-free domains of knowledge. The need to automatically identify bad actors posting online news [11, 12] or social media [13, 14], can wrongfully limit an individual's freedom of speech or be gamed effectively by deliberate bad actors or states. These situations contextualize the ethical and professional domain of the hypothetical cyber-archivist, the AI librarian or scholarly assistant who processes written data and annotates it for further analysis or classification.

AI's usefulness for all such cyber-archivist tasks is undeniable, given its ability to quickly sift through massive datasets and to detect and trace patterns that would be impossible for a human to process with any efficiency [15, 16]. For example, given human limitations and financial considerations, combing through online media posts to detect trends in public sentiment, or to detect spam in individual post comments, would require more personnel hours than could reasonably be brought to bear by any individual party or organization. As more and more data about our world becomes available and meets computing power to process it as never before, this apparent usefulness can only grow. But whether or not such usefulness is truly beneficial, or merely an invitation to hand over human judgment to fallible algorithms, given the potential for bias and error, is a topic of intense debate [17–21]. And when AI is used to process and pass judgment upon large data sets, attempts to improve the quality of an AI solution may be hindered by the very nature of the data that leads us to embrace such solutions – specifically, its vastness. For example, if an AI model that has processed vast volumes of data is found to be flawed, then correcting such a flaw and embracing a new model may be impossible without entirely reprocessing the vast datasets involved. This could mean that opportunities to embrace new, more trustworthy, AI models (or to simply tweak existing models to correct a minor flaw), would be lost to us without sufficient information being preserved regarding more than simply the results of prior processing.

To debate the merits and perils of applying such technologies without consideration for how the underlying technological infrastructure could be changed to promote or discourage risks, is a necessary ongoing ethical conversation, for any blinkered views could lead to an inaccurate and potentially harmful AI model. Given the fundamental nature of this problem for all AI models we will consider the role of automated algorithms in rendering judgment without reference to a specific domain, that is, in its most general form as a processor of data that mimics human judgment. More specifically, we look to how artificial automation is analogous to an archivist or librarian citing, archiving, and scholarly critiquing data. We are, therefore, dealing with the question of whether or not a cyber-archivist can be both useful and safely trusted.

In deciding whether or not to place AI technology in a position of trust, the question is not merely whether the AI can be trusted to offer good judgments, but also critical is how that technology, and the judgements it makes, is integrated into the broader system. The questions of whether or not an AI's judgment can be trusted is not therefore our focus, but rather we look at the manner in which it is best applied. We illustrate the potential to overlook this by illustrating how underlying infrastructure can impact the amount of trust placed in AI, and we do this by describing our system, *MultiVerse*[1], which allows us to support

[1] The term "*Multiverse*" is widely used in different domains to describe different concepts. In science, it refers to everything that exists in totality [22] - as a hypothetical group of multiple universes. In quantum-computation, it refers to a reality in which many classical computations can occur simultaneously [23]. In a bibliographic-archival system, referred to as "Archival Multiverse", it denotes "the plurality of evidentiary texts (records in multiple forms and cultural contexts), memory-keeping practices and institutions, bureaucratic and personal motivations, community perspectives and needs, and cultural and legal constructs" [24](Pluralizing the Archival Curriculum Group). In Information Systems, it deals with

the coexistence and processing of multiple (competing, and potentially conflicting) decisions within the same archive. In other words, we argue that the ethical dilemma posed by whether or not AI can be trusted in roles of judgment can be mitigated by building better technological infrastructure underlying such AIs and affecting how AI and humans interact and collaborate. Specifically, we use the analogy of a flawed cyber-archivist, being trusted thanks to the construction of a suitably resilient library, rather than being the subject of attempts to create a flawless AI to serve as a trustworthy cyber-archivist.

The rest of the paper is organized as follows: Section 2 discusses the related work covering the efforts in tackling the trustworthiness of automated systems and the importance of human-computer interaction. Section 3 further leads the ethical discussions of AI/ML as understood by the proponents and opponents of cyber-archivists. Section 4 discusses philosophical ethics and pragmatic approach on trust within the context of AI automated content. Section 5 briefly describes our project, *MultiVerse*, as an illustrative example to discuss the importance of the underlying data storage infrastructure of an automated system, and broader ethical concerns. In particular, our focus in this paper is on the broader conflicting ethical implications that can be impacted by such focus on systems infrastructure (e.g., data privacy versus veracity, accuracy versus authenticity, efficiency versus transparency, and the ongoing need for more explainable AI). Section 6 offers a summary of the paper while also describing further applications of *MultiVerse*.

2. Related Work: The Problem of Flawed Librarians

With our use of a library analogy and its focused use of text analysis and annotation, it is necessary to acknowledge the efforts that lead us to this work. In particular, there is a large body of works on automating the processing of textual data and considerable recent efforts in tackling the trustworthiness of such automated systems. One particularly promising approach has been to consider how humans and AI can most beneficially interact. Our proposal, to focus more on the underlying storage infrastructure as a means of mitigating potential problems, builds upon our ongoing work, and a considerable body of prior research, in the domain of data provenance.

Tools and techniques in automating data science, also known as AutoML/ AutoAI, are the subject of research in many companies and open source communities[26, 27]. Given the speed and cost-effectiveness of AI for such tasks, there is optimism in the industry that AI/ML systems can eventually replace the thousands of human workers who are currently involved in making decisions, for example, automated comments moderation on social media [28]. Other examples of automated ML and NLP techniques for semantic sentiment analysis include: financial microblogs and news [29], twitter [30–32], big social data [33], clinical analytics [34], specific language-based literature [35–37], and publishing domains [38–40]. These systems have the potential to perform moderation much faster than human moderators, which is attractive for more than simple performance/cost reasons (since removal of harmful content quickly can reduce the harm it causes). Automating humanly laborious tasks not only facilitates scalability, it is also promoted for its potential to introduce consistency in performing allocated tasks/decisions. But this is not necessarily a good thing, if an error or a bias is consistently and reliably propagated across vast volume of data and large number of people.

Despite the many benefits of automated ML and NLP techniques, their use introduces new challenges. In an AI-automated system, identifying tasks that should be automated and configuring tools to perform those tasks is crucial. Perhaps there are those who view the biggest hurdle in accepting AI-generated models to be the lack of trust and transparency, given the potential for large-scale harm due to errors [27]. Attempting to understand an intelligent agent's intent, performance, future plans, and reasoning process is a challenge. Accurate automated systems are not an easy task. These challenges place a greater emphasis on how AI and humans interact, and prior research on this point – Computer Supported Cooperative Work (CSCW) research – has established that a fundamental socio-technical gap exists between how individuals manage information in everyday social situations versus how this is done explicitly through the use of technology [41, 42]. Often, technical systems fail to capture the flexibility or ambiguity that is inherent in normal social conditions [43]. Concurrently, research findings reveal the deficiencies of AI in making decisions that require it to be attuned to the sensitivities in cultural context or to the differences in linguistic cues[40, 43, 44]. These failures to detect individual differences of context and content can have serious consequences, for example, in failing to distinguish hate speech and misinformation from newsworthiness in automated news feeds can have serious consequences. In fact, these failures to address context issues and misinformation on automated *Facebook* or *WhatsApp* content regulation

the complexity, plurality, and increasingly post-physical nature of information flows [25]. Our use of the term *"MultiVerse"* with a capitalized 'V' denotes a version of our proposed digital infrastructure for a richer metadata representation, which captures the nature of representing multiple versions of a source data object, and was named partially due to the system's earliest tests being focused on translated poetry verses.

arguably contributed to violence in Myanmar [45]. Overcoming these obstacles requires human ingenuity and the moral to engage artificial intelligent systems.

To overcome these challenges and to boost user's morale to act upon an artificial intelligent system requires human intervention. The Human-in-the-loop system or Human-guided machine learning [42] taps the speed and processing power along with human intuition and morality. Hybrid AI-Human systems forge a strong collaboration between artificial and organic systems and this opens a way to solve difficult tasks that were once thought to be intractable. To be ethical, this man-computer symbiosis must be characterised by the cooperation of machines with humans. The machine and AI systems should not be designed to replace the natural skills and abilities of humans, but rather to co-exist with and assist humans in making their work and lives more efficient and effective. Fortunately, some progress towards this goal has been made. Some works that combine human-in-the-loop collaboration with AI for solving difficult problems include, but not limited to: image classification [46], object annotation [47, 48], protein folding [49, 50], disaster relief distribution [51], galaxy discovery [52], and online content regulation [53].

Human-Computer Interaction (HCI) and in particular Computer-Supported Cooperative Work (CSCW) are not radically new concepts in spite of their current urgency. The concept of symbiotic computing has been around since the early 1960s "Man-Machine Symbiosis" work by J. C. R. Licklider [54]. Licklider envisioned computers serving as partners whose interactive design as intelligent agents would collaborate with human beings to solve interesting and worthy problems in computing and society. This view can be universally applied to any technologies that extend or enhance humans abilities to interact with their environments, and can therefore be considered a persistent question surrounding our interaction with AI.

More generally, as long as human operators and new automated systems simultaneously adapt, they will co-evolve. However, it remains important to remember that the socio-technical gaps that CSCW problems generalize, are never completely resolved and continued efforts to "round off the edges" [43] of such coevolution is necessary. Given the shortcomings of automated tools and the required careful human administration of these tools, we propose that instead of developing fully automated systems that require perfection for complete autonomy, researchers and designers should make efforts to improve the current state of mixed-initiative regulation systems where humans work alongside AI systems.

Since automated tools are likely to perform worse than humans on cases where understanding nuance and context is crucial, perhaps the most significant consideration is determining when automated tools should perform certain tasks by themselves and when results of such tasks need to be reviewed by human actors. We echo calls by previous studies for building systems that ensure that the interactions between automation and human activities foster robust communities that function well at scale [44].

We specifically focus on our own proposed *MultiVerse* which is an example of a broader research into the maintenance and preservation of richer semantic metadata. Our own work falls under the larger project of *MetaScriptura* that focuses on the infrastructure that is needed to maintain richer semantic metadata including the ability to preserve provenance information and to present annotated and multivariate data. *MultiVerse* specifically focuses on the presentation of the issue of multivariate data, but our work builds on prior work in data provenance and immutable data storage. Examples of such works include scientific workflow management (such as Kepler [55], Vistrails [56], Taverna [57], etc.), graph database storage (such as Neo4j[2], AgensGraph[3], TigerGraph[4], LightGraph[5]), blockchain [58], FreeHaven[6] [59], Haven [60], glacier [61], etc. Scientific workflow management and graph database storage systems help in generating and maintaining data provenance and make use of either resource description framework (RDF) or labeled property graph to model data structure and storage. Blockchain technology, FreeHaven, Haven, and glacier deal with anonymity, immutability, and persistence to provide highly durable decentralized data storage by which they enable trust platforms and protect data storages from potentially accidental catastrophic failures or malicious adversaries who may attempt to erase data from a distributed storage system. We leverage some of these immutable storage and provenance tracking systems to enable *MultiVerse* to be part of the system that provides a trust platform that prevents people from altering what was written/stored in a distributed system.

MultiVerse looks at how an AI's improved infrastructure, for the preservation of both source data and its annotations (including AI generated annotations), can help grant greater resilience to decisions making capacities of AI-human systems. Our approach simplifies these decisions, as well as, allots for their safe reversal or delaying their implementation. In this way, a boon is made for explanatory data that supports these decisions of critical importance in the creation of accessible AI that also complies with the legislative demands for transparency like the EU's General Data

[2] Neo4j. https://neo4j.com/

[3] Agensgraph. https://bitnine.net/

[4] Tigergraph. https://www.tigergraph.com/

[5] Lightgraph. https://fma-ai.cn/

[6] FreeHaven. https://www.freehaven.net/

Protection Regulation (GDPR) [62–64]. It does so by preserving more data regarding how annotations (i.e., automated results and judgments), were produced. It also supports the preservation of multiple versions of such results (which would also be needed in explainable AI approaches that use a black box neural network alongside a more explainable transparent box - like a decision tree - to provide explainability for results). These features combine to grant greater flexibility in how humans verify the results or describe its data sources or when the results require explanation. To offer such a richer storage infrastructure, we leverage a novel architecture built upon our own extensions of data provenance research. Data provenance research is focused on the preservation and presentation of the origins and transformations of stored data, and has typically been narrowly employed for the management of project data like scientific workflow or code management [65–69].

3. The Proponents and Opponents of Cyber-Archivists

Opposing Camps of AI:. While AI systems present enormous potential benefits, they are not without problems. As a result, there are opposing camps arguing extreme views on the acceptance or rejection of AI. The optimists of AI, like Ray Kurzweil, an inventor and futurist [70] and other AI enthusiasts [71], predict a utopian future of immortality, immense wealth, and all-engaging robotic assistants to humans, ushered in with the singularity AI help. These techno-optimists believe that Genetics, Nanotechnology and Robotics (GNR) with ‘strong AI’ will revolutionize everything “allowing humans to harness speed, memory capacities and knowledge sharing ability of computers and our brain being directly connected to the cloud” [70]. On the other hand, there are those who argue AI risks and its potential dystopian consequences. The critics of strong AI include the likes of Bill Joy, a computer engineer, co-founder of Sun Microsystems, and venture capitalist [72], Stephen Hawking, a theoretical physicist [73], and Nick Bostrom, a philosopher at the University of Oxford [74]. They believe that AI is “threatening to make humans an endangered species and second rate status” [71]. But there are others like Sam Altman, an entrepreneur and CEO of "OpenAI" and Michio Kaku, a theoretical physicist and futurist, who believe that AI could be controlled through “openAI” and effective regulation [75]. They believe that humans could learn to exploit the power of the computers to augment their own skills and always stay a step ahead of AI or at least not be at a disadvantage. The spectrum on this is expansive as it ranges between the extremes of reactive fear and complete embrace of AI. Both accounts fail to make a rational and ethical assessment of AI. The practical debate, the real question, is not *whether* AI technologies should be adopted, but *how* they can be most beneficially, and most safely, adopted.

Algorithmic Transparency:. How algorithmic decisions are embedded in a larger AI system is difficult and specialized area of study. When an AI system produces outputs that can lead to harm, the likelihood of realizing that, let alone remedying it, can often be blamed on a lack of transparency regarding how the outcomes were reached. This has led to increasing demands for algorithmic transparency. But the immediate claim that these problems can be remedied by greater algorithmic transparency offers little more than the self-evident. Basically, any process or technology that does not offer perspective on its manner of operation is inherently suspect, and unlikely to be trusted. There is, of course, a place to discuss the philosophical notion of transparency as an ideal. Indeed, it can be argued that the genealogy for any one practical instantiation of the transparent is ultimately found in epistemological speculation concerning the nature of truth.

Recently, transparency has once again taken a prominent place in public governance systems, where social activists strive for greater government accountability. In AI, as with these practices, transparency is touted as a way to disclose the inherent truth of a system. In the context of AI, it is understood as taking a peek inside the black-box of algorithms that enable its automated operations. However, we view transparency for AI systems more broadly, not as merely seeing phenomena inside a system, but rather, across the system, as argued by Ananny and Crowford, and Crawford [76, 77]. That is, not merely as code and data in a specific algorithm, but rather to see “transparency as socio-technical systems that do not contain complexity, but enact complexity by connecting to and intertwining with assemblages of humans and non-humans” [76]. In other words, it is better to take account of the more complete model of AI and this includes a comprehensive view of how humans and algorithms mutually intersect within the system [77]. Without a sound understanding of the nature of algorithmic transparency and decision making, a false conflation of the "algorithmic operation" and human policy failings is possible. This is an especially troubling occurrence when inherent bias in an AI model is applied to the judicial system as evident in the the scandalous *COMPAS* revelations about the Correctional Offender Management Profiling for Alternative Sanctions algorithm [6–8].

Accountability Beyond Algorithmic Transparency:. In the ideal, algorithms are transparent when they are predicative, enable benefits given they are fundamentally neutral, unbiased. As stated previously, it is logically possible that deterministic, flawed or discriminatory algorithms may on occasion produce equitable outcomes – an AI system must be continuously evaluated [78]. On this reality, Dwork and Mulligan state concerning AI "the reality is a far messier mix of technical and human curating" [78]. AI has moral implications, but never in isolation of the context in which it is applied. When AI has a negative impact, the assumption of fault and responsibility differs based on your perspective and role.

If algorithms are presented as an open book, then the developers of algorithms have less responsibility when they are misapplied. On the other hand, if algorithms are constructed as a black-box, or an autonomous agent operating with an opaque logic, then the users are denied accountability for how algorithms make decisions that affect them. In essence, the developers of such systems are asking that their judgment be trusted blindly, and would therefore be expected to shoulder more responsibility for any future problems.

There are also different default assumptions depending on the role one plays. Generally speaking, the present legal system does not hold firms responsible for the misuse of algorithms they develop [8, 27], but they can be held responsible for systems they sell. From the perspective of software developers, their algorithms are neutral and so a failure is more likely assumed to be due to users' thrusting algorithms into fallible contexts of biased data and improper use. At the users' end, algorithms are difficult to identify and comprehend and therefore they aren't typically held accountable for the ethical implications of their use [27]. [79] and [80] suggest that as algorithms seem to be unpredictable and inscrutable, assigning the responsibility to developers or users is ineffective and even impossible, but firms could be better held responsible for the ethical implications of their products' use. The author [27] conceptualizes algorithms as value-laden in that algorithms create moral consequences, reinforce or undercut ethical principles, and enable or diminish stakeholder rights and dignity. In other words, ascribing responsibility for algorithms resulting in harm is very tricky. This lack of clarity is a hurdle to responsible and ethical adoption of algorithms in critical roles, e.g., when they are placed in roles that require them to pass judgment. But it is insufficient to say that these risks need only greater transparency of the algorithm, for the algorithm alone is never responsible for the outcome, and transparency needs to expose more than the workings of an individual algorithm to offer the most resilience and trust possible. Moreover, an algorithm's transparency and one's relevant faith in it involves the quality of data it processes, the structure of the AI from which it operates and larger socio-cultural considerations introduced with human involvement.

Without striving for transparency beyond the specific algorithm, i.e., striving for a broader, more holistic view of the system, we may miss opportunities to build better and more resilient AI-enhanced systems. Returning to our analogy of a cyber-archivist, we would argue that simply offering a view of the workings of a particular instance of such an AI is to pass on the opportunity to really understand the overall system and lessen later opportunities to harden it against failures. Specifically, imagine if one particular algorithm for processing a large dataset was deemed to be the best, and was employed for several years with acceptable performance (including full transparency regarding its implementation), but that it was discovered that its outputs were flawed for certain edge cases that could have been caught with a superior algorithm. The only way to remedy this, would seem to be to reprocess the entire dataset (assuming it is still available), and to compare the outputs of the algorithms. But if the data storage infrastructure had the facility to support the operation of both algorithms, and the maintenance of the provenance of their outputs, then this process would be feasible without a reprocessing of the potentially vast datasets (assuming they are still available). It's exactly this kind of increased accountability and accounting that is possible if we aim for transparency that goes beyond the algorithm alone, and is enabled with infrastructure that can support such a goal. Our *MultiVerse* system is an example of such an infrastructure.

4. Understanding Trust

Here we do not attempt to define the concept of trust, rather we explore it in the broader context of digital data preservation, given the proliferation of digital information generation and dissemination. Therefore, before probing the proposed *MultiVerse* system, we believe that it is essential to take a brief look at the term "trust" as understood in moral philosophy and practical contexts. Though many scholars agree on the importance of trust in individual endeavours and in larger society, there is no precise universal definition of it. Trust is often understood as an important element in building interpersonal and group behavior [81]. It is also key for managerial effectiveness and socio-political cohesiveness [82]. This understanding of trust is extensively examined in organizational theory. However, there is considerable uncertainty on the conditions or determinants of trust. Trust is generally assumed as an optimistic expectation, or a willing cooperation, on the part of an individual on the ultimate benefits resulting from an event or the behavior of a person (or group or institution). In contrast, [83] argues, "trust is based upon an underlying assumption of an implicit moral duty with a strong ethical component owed by the trusted person to the trusting individuals" (p.381). Furthermore, there are many different approaches and contexts in which the concept of trust could be explained. For example, as observed by [81], they include: (a) individual expectations, (b) interpersonal relationships, (c) economic exchanges, (d) social structures, and (e) ethical principles. Trust could also be understood from the perspective of philosophical ethics or pragmatic implementation approach. The former approach deals with the ideal and abstract notions of trust, while the latter concerns itself with practical implications of trust as experienced in different contexts, or kinds of trust that can be offered from a system.

4.1. Philosophy and Ethics on Trust

The concept of trust in moral philosophy (Western), is discussed in conjunction with the ultimate goal of reaching a “first principle” upon which all other rules can be based, and that would lead to a “good” society. The ideal first principle, or decision rule, has not been found. Instead there are now a number of alternative decision rules or principles that provide different perspectives or views of moral problems, and that are applied in sequence to gain understanding and insight [84]. Therefore, the concept of trust in philosophical ethics is understood as the result of a given decision or action that recognizes and protects the rights and interests of other people through an application of the ethical principles of analysis. For example, as stated by [81], “trust is the result of ‘right,’ ‘just,’ and ‘fair’ behavior, that is, morally correct decisions and actions based upon the ethical principles of analysis-that recognizes and protects the rights and interests of others within society.”

Trust can be further discussed based on principles/perspectives in traditional moral philosophy. Each of the first principles or decision rules or alternative perspectives from the classical ethicists asserts the following, as summarized by [75]: that a "good" person should act not for his or her short-term self gain only, but for a mixture of that gain together with his or her vision of the future (Protagoras), his or her sense of self-worth and personal virtues (Plato and Aristotle), his or her goal of community and religious injunctions (St. Augustine), his or her fear of retribution (Thomas Hobbes), his or her government requirements (John Locke), his or her calculation of social benefit (Jeremy Bentham and James Mill), his or her understanding of universal duty (Immanuel Kant), or his or her recognition of individual rights and social contracts (Jean-Jacques Rousseau and Thomas Jefferson), his or her notion of distributive justice (John Rawls), his or her application of contributing liberty (Robert Nozick). All of these normative rules, designed to take the legitimate interests of others into account, were assumed by moral philosophers to encourage greater trust among, and to improve cooperation between, the diverse elements of society and consequently, result in "good" (in the widest possible sense of that term) for the society rather than the individual. A "good" society has been defined [85] as one in which the members willingly cooperate for the ultimate benefit of all. It is in the context of establishing a good society, that trust is typically defined in moral philosophy.

Where the many views of moral philosophers diverge when it comes to trust, is in their assumptions regarding what values, behaviors, systems, and paradigms result in a desirable form of “good.” They therefore may differ in their perspectives regarding what mechanisms are most essential, but we posit that using a system like *MultiVerse* is an example of enabling infrastructure that does not force a choice of perspectives, but instead, supports those that can be aided by greater tracking of contributions and intentions of different individuals involved in the translation and transmission of information.

4.2. Pragmatic Approach on Trust

In contrast to purely philosophical and ethical perspectives on trust, a pragmatic approach abandons the question of the purest definition of the term, and deals directly with the question of what kinds of trust can be offered in a system. In this approach, one addresses questions such as: What does it mean to trust and validate data when deepfake phenomena are evermore on the rise on social-media-enabled platforms/forums? What does it mean to certify authenticity of data when the sources of data are neither available nor traceable? How do we trust the custodianship of data when the custodians of data are prone to economic benefits based on data collection and dissemination? Why does one trust a distributed ledger certifying a bitcoin transaction rather than any online transaction? Or why does one trust an authenticated website rather than a simple web server? What does it take to make a trusted digital content in a digital system?

We identify three ways, among others, that answer why some systems can be trusted. They are: trustable holder of the data, trustable minimum quorum of members, and trustable incentive mechanism.

1. Trusting because of the holder of the data, for example the holder is viewed as not only trustworthy in themselves, but capable of vouching for the trustworthiness of the integrity of data they hold.

2. Trusting because a quorum of members agree on the data. In this case no individual member has the authority to vouch for the data’s integrity, but trust is gleaned from a faith in the unlikelihood of enough members erring that a quorum cannot agree on the canonical form of the data.

3. Trusting because an incentive mechanism required to corrupt the large number of members that use the system to alter the reality becomes untenable. For example, in the case of bitcoin, the proof of work model - a decentralized consensus mechanism - requires members of a network to expend effort solving a mathematical puzzle to prevent anybody from gaming the system.

MultiVerse is an example of a mechanism that incentivizes honesty and trustworthiness in the actors that use it. It discourages individual participants from violating data, or provenance, integrity through an architecture that allows for dependence on not just a quorum of honest parties, but by ensuring that the change of any piece of the record invalidates the entire store's integrity to date. This kind of immutable store is therefore both a quorum based system for maintaining the integrity of the MultiVerse data store as a whole, and an internal structure that renders individual, undetected, edits to sub-parts of the store infeasible.

5. Trusting the Cyber-Archivist – *MultiVerse*

MultiVerse is designed as a digital data infrastructure that preserves multiple perspectives, and thereby allows better support for multicultural digital content. We contend that in order to better support transparency, intercultural ethics, and more ethical digital media curation across cultures, such an infrastructure is needed. So, what is *MultiVerse*? *MultiVerse* is a digital data representation infrastructure intended to track provenance of multi-varied translations of scholarly texts and their derivatives. Provenance can be defined as the recording of the history of user activities that create and transform data. The *MultiVerse* infrastructure allows users to remix/combine existing translations and/or add one's own personal translations at will and add annotations to it. Annotations can be made regarding the scope, context, or other relevant metadata. Provenance refers to the recording of the history of user activities that create and transform data. *MultiVerse* is primarily concerned with the metadata needed to store such provenance alongside the data to which it refers. In this project, provenance tracking is done by capturing all translations (users' activities) without any preferences, prejudices, and prizes (value judgements/correctness), at the time of their composition.

To realize this concept, we have used the well known 13th century Italian poet Dante Aligheri's the *Divine Comedy*, and some of its many English translations [86]. We have combined these into a single repository that allows the remixing and composition of new translations, while offering detailed tracking of the origins and transformations of such texts. A user has the option of either collating different versions of verses or adding in his/her versions of verses from/to this repository to compose his/her unique version of translation of the *Divine Comedy*. Moreover, the user can tag richer semantic metadata like context, intent, scope, or tone/sentiment to his/her composition. Multiple versions of the *Divine Comedy* are thereby stored in a single repository with rich version histories. A high level architectural overview of *MultiVerse* are depicted in Figure 1.

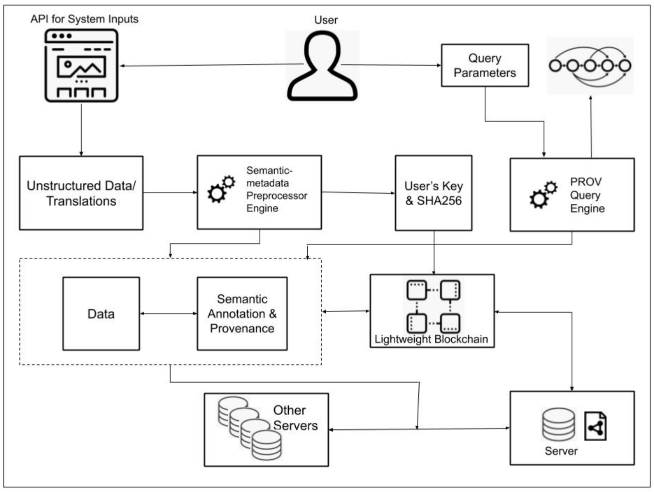

Figure 1. *MultiVerse*'s Architecture Overview

The primary purpose of this project is to demonstrate the importance of a robust data storage technology, in the context of human-in-the-loop system, that captures and represents pluralistic views cutting across individuals' cultural, ethnic, religious, gender, social, political, economic, emotional, etc., stances/ viewpoints. At its very beginning, a key design principle of *MultiVerse* is to enhance technology to represent pluralistic multicultural perspectives of all users, rather than after-the-fact. This is achieved by designing *MultiVerse* which enables users to record not only their views irrespective of their correctness but also accommodate their contexts and intents.

We might ask, "what are the benefits of this technology design principle in the first place?" Without arguments, it can be stated that all voices (decisions/judgements) are preserved. Single versions can be presented on demand. But the history and identity of those who selected the individual versions and the provenance of the documents can be permanently stored using blockchain technology [87] and can not be tampered with in any way. By virtue of its immutability, *MultiVerse* becomes a means to establish the source of any loss of nuances, and makes arguments (by allowing future archaeology on such repositories) about the correct form moot. More precisely, while it does not eliminate contention over the ideal translation, it does not force that debate to be fought over the preserved version. There need be no permanent winner, and past mistakes can be corrected in future revisions. But this leads us to consider the broader ethical implications of such multicultural pluralistic digital infrastructures.

In the context of AI, it helps to record the decisions of users and machines, and to preserve them for as long as they might be needed. Such logs are useful in case we need to revisit them, whether to better understand past behavior or to further enhance future decisions. As the underlying data storage repository in *MultiVerse* preserves all versions of decisions in an immutable manner, any additions, deletions, and modifications may be made as annotations without corrupting original logs, or conflicting with their subsequent versions. It thereby helps by protecting their lineage/provenance.

5.1. Using *MultiVerse*

A user creates a resource (multi-varied translated texts in our initial example, *MultiVerse*), either by copying an existing resource, or by newly translating a resource. For this new resource, or for an existing resource we wish to add to the system, *MultiVerse* generates a few standard properties, also known as the structural metadata for that resource. This metadata describes the resource, such as its type, size, creation date, and creator, and adds this information to a provenance log. The *Semantic Metadata Module* extends this mechanism to allow generation of user-specified descriptive properties such as context, scope, and sentiments. These additional properties are a concrete example of what we mean by "richer semantic metadata." These properties will be based on the uploaded data as well as newly derived sources. Consequently it is possible to register new translations for existing resources and/or generate a new resource. This new resource can be described as a source, with its own location, i.e., context (which would be, for example, specified through a URL). It could, for example, be generated from an existing resource via a copy operation (where that existing resource would be the source for this copy). To help track a copied resource's origin, *Semantic Metadata Module* adds a source property, which becomes part of the provenance of the resource. This source property is added to the new copy, which links it to the original URL.

Once a user integrates a translated version of the data into his/her work space, the user can proceed to the next task in the plan. In the next task, if a user chooses to make his/her own translation, the *Semantic Metadata Module* generates a *hasTranslation* property and enables a user to tag information about the user-as-the-translator, its creation time, context, and scope of the translation. Using the provenance log, the *Semantic Annotation-Provenance Module* will help document the data's provenance into annotated provenance documents that contain both structural as well as user-specified descriptive metadata.

Given the final derived product's URL, anyone granted access to *MultiVerse* can trace backward following the links in *hasSource* and *hasTranslation* properties to discover the input data and relevant user-specified metadata entries. This kind of query would not be possible without the added metadata (i.e., the semantically-enrichable provenance framework we have proposed in *MultiVerse*). Adding this metadata would increase the storage demands of the system as a whole, but these would be increases in capacity demands (simply the volume of data stored, as opposed to the needed storage system performance), which is arguably a cheaper resource than the time, energy, and temporary storage demands of having to reconstitute such information at a later point in time. In other words, assuming that it is possible to reconstruct the varied versions of our data at a later date (which is not necessarily possible), then there is a tradeoff between efficient space utilization today, and the cost of future computation and data retrieval demands tomorrow. The decision to store such metadata thereby holds efficiency considerations, in addition to the added transparency it could provide.

MultiVerse is not just a repository of multivariate data, but a means of ensuring the preservation of those versions against malicious action attempting to rewrite history, hence the immutability requirement is incorporated into *MultiVerse*. To keep such a repository consistent, it is structured as an immutable data store, allowing the addition of new content and amendments, but disallowing any modification or deletion of data that has been committed to this store. The immutable aspect of *MultiVerse* is achieved by adapting a basic model of blockchain technology [**?**]. The technical details of blockchain technology is beyond the scope of this paper. The interactive aspects of *MultiVerse* are enabled by offering a user application programming interface (API) to annotate semantic analysis decisions and allow access to the repository in a secured manner. We discuss the ethical implications of the *MultiVerse* framework in the next subsection.

5.2. Ethical Considerations

A moral question that arises on *MultiVerse* is : How does *MultiVerse* change the ethical debate around allowing an algorithm to judge/annotate and provide an actionable opinion? Our approach, illustrated through the *MultiVerse* example, shows that it is possible to construct systems whose impacts are more easily reviewed and evaluated against each other (since multiple versions are readily accessible for comparison), or that allow decisions taken by an automated algorithm to be less permanent in their effect (since alternative results that have been preserved, can be retroactively embraced). In other words, by allowing for one of three outcomes:

1. The decisions can be undone by preserving results of the prior decision and superseding it by adopting an alternate decision at a later date;

2. If undoing is not possible, then perhaps it allows us to defer making the decision at all, if we delay the aggregation or selection amongst alternative annotations (judgements) until the latest possible point in time, we would have guaranteed the adoption of the best and fairest technology available for that decision; or finally,

3. Assuming that decisions can neither be undone nor delayed, it is still beneficial to have on hand the results of competing models, if only to aid the more rapid analysis and evaluation of new and improved models, and to improve and accelerate our understanding of where and how defunct models may have failed.

On the contrary, leaning too heavily on an ability to defer or delay decisions, or a false sense of immunity to bad decisions, can lead to more reckless human adoption of algorithmic decision-making technologies.

However, one view of what distinguishes human intelligence from AI in decision-making is our ability to make connections in ways that are not formalizable (through unconscious processes, for example, or by involving emotions). When seen from that perspective, an AI algorithm would be a tool enacting what is ultimately a human will. That human will may be inexplicable, but the algorithms can and should be transparent and open to revision, making it easier to adopt in an informed manner. The use of an infrastructure like *MultiVerse* may aid in documenting such open algorithms, or may host the results of more opaque algorithms. It does not dictate taking one approach or the other.

The moral considerations of *MultiVerse* are slightly different than the moral considerations of using AIs for sentiment analysis. Harm is mitigated by potentially making sure that no decision is necessarily permanent, or that bad decisions can be attributed to specific sources (allowing for greater accountability), but this still leaves concerns. It is possible to confuse the mitigation of harm with the elimination of the possibility of harm, which of course is not the case here. A decision can be revised if enough provenance data is available to retroactively consider alternatives, but the effects of decisions might not be reversible (e.g., we can learn to improve a sentencing algorithm, but cannot expect any data storage system to restore a single day of unjustly lost freedom. While it may be possible to retroactively determine what a sentence algorithm could have recommended, it is definitely not possible to undo a sentence that has already been served). A potential harm that could be introduced arises if users of *MultiVerse* are lulled into a sense of complacency, such that human errors that would result in poor decisions might be made more often. *MultiVerse* provides the ability to mitigate harms and add greater accountability, but it is still up to individual deployments of systems to actually monitor the performance of their "cyber-librarians" and to temper their decisions when there is doubt about the quality of their outputs.

A significant portion of the potential harm of automated systems can arise as a result of those systems shifting the focus of responsibility away from humans. In other words, when we lose accountability, harm caused by acting on AI-provided data would not necessarily be blamed on those who should have maintained human oversight of how we got there. A mechanism that can improve the accountability of such systems, improving tracking of problems to failures of algorithm selection or oversight, would therefore have the potential to encourage both system builders and system adopters, to be more conscientious and ethical (thanks to an awareness of provenance tracking), but may also be helped in their oversight tasks thanks to the long term evaluation and auditing of the performance of different algorithms. The different choices regarding whether we defer to the algorithms, when and how often we defer to the algorithms, or when and how often we defer to the algorithms that are deployed for a specific problem is a question related to best practices around auditing and system improvements.

Finally, one might perceive *MultiVerse* as a system that is deliberately designed to record too much metadata, thereby creating an unnecessary information overload; or as a scientific apparatus to dissect the intellectual work of others; or as a blockchain mechanism to prevent the ability to edit what is stored. This leads to the issue of (data) privacy in the context of immutability of stored information about persons interacting with *MultiVerse*. To prevent these undesired consequences, there is a choice, by design, for users either to opt out from recording all their creative activities or to opt in to reveal as much as it is needed or to choose documenting the synthesizing process of a digital product. Such decisions regarding opting in or out would affect what data is recorded by the system, but it's important to recognize that when it comes to the question of an individual's right to be forgotten, such a question is not simply decided by the presence or absence of data, but is a question of the retrievability of such data. A data store can be immutable, and hold data that is never completely removed, and yet can still honor an individual's right to be forgotten within such a system, for example, adapting users' data access and retrieval rights and policies as appropriate.

To return to the use a library analogy, we go beyond prior efforts by focusing less on making librarians less flawed, but instead highlighting how an improved library could perhaps lessen the risk of harm posed by less-than-perfect librarians, and help all who support and benefit from librarians to better support and improve the library.

6. Concluding Remarks and Further Applications

To demonstrate how rethinking underlying technical infrastructure can reshape the questions we face with AI, we illustrated an example of one such "rethought" realization of a data storage system. By combining elements of version control systems, trusted immutable stores, and provenance technologies, *MultiVerse* shows that we can defer and revise decisions between human and automated analysis.

Such an infrastructure functions as an example of how to critically rethink the either/or decision regarding the applicability of AI. In fact, this infrastructure is useful for any AI domain that involves NLP and text processing/classification of texts, etc. While we've used the analogy of a librarian, to emphasize that our focus is on systems that automate the processing and tagging of textual information, our arguments should hold for any data processing task that could involve AI. It, therefore, would have applications beyond scholarly articles and references, including domains like managing fake news, social media, synthetically generated media, legal and governmental processes, materials in the broader arts and sciences (beyond simple workflow management), and can encompass more than purely textual media and materials.

7. Copyright statement

7.1. Copyright

Acknowledgement. Special thanks are to the Department of Computer Science and Engineering of Santa Clara University, Santa Clara, USA., for making available its computing research lab facilities.

References

[1] Garfinkel, P. (2016) A linguist who cracks the code in names to predict ethnicity. *New York Times* .

[2] Angwin, J., Parris Jr, T. and Mattu, S. (2016), Breaking the black box: When algorithms decide what you pay. propublica.

[3] Kharif, O. (2016) No credit history? no problem. lenders are looking at your phone data. *Bloomberg.com* .

[4] Katwala, A. (2020), An algorithm determined uk students' grades.

[5] O'Neil, C. (2016) *Weapons of math destruction: How big data increases inequality and threatens democracy* (Broadway Books).

[6] Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2019) Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. 2016. *URL https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing* .

[7] Wexler, R. (2017) How companies hide software flaws that impact who goes to prison and who gets out. *Washington Monthly* .

[8] Wisser, L. (2019) Pandora's algorithmic black box: The challenges of using algorithmic risk assessments in sentencing. *Am. Crim. L. Rev.* **56**: 1811.

[9] Dos Santos, C. and Gatti, M. (2014) Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*: 69–78.

[10] Redhu, S., Srivastava, S., Bansal, B. and Gupta, G. (2018) Sentiment analysis using text mining: a review. *International Journal on Data Science and Technology* **4**(2): 49–53.

[11] Al Asaad, B. and Erascu, M. (2018) A tool for fake news detection. In *2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)* (IEEE): 379–386.

[12] Monti, F., Frasca, F., Eynard, D., Mannion, D. and Bronstein, M.M. (2019) Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673* .

[13] Zinovyeva, E., Härdle, W.K. and Lessmann, S. (2020) Antisocial online behavior detection using deep learning. *Decision Support Systems* : 113362.

[14] Rafiq, R.I., Hosseinmardi, H., Han, R., Lv, Q. and Mishra, S. (2018) Scalable and timely detection of cyberbullying in online social networks. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*: 1738–1747.

[15] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) From data mining to knowledge discovery in databases. *AI magazine* **17**(3): 37–37.

[16] Piateski, G. and Frawley, W. (1991) *Knowledge discovery in databases* (MIT press).

[17] Grove, W.M. and Meehl, P.E. (1996) Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, public policy, and law* **2**(2): 293.

[18] Johnson, C. and Taylor, J. (2016) Rejecting technology: A normative defense of fallible officiating. *Sport, Ethics and Philosophy* **10**(2): 148–160.

[19] Lehner, P.E., Mullin, T.M. and Cohen, M.S. (1990) A probability analysis of the usefulness of decision aids. In *Machine Intelligence and Pattern Recognition* (Elsevier), **10**, 427–436.

[20] Mateos-Garcia, J. (2017) To err is algorithm: Algorithmic fallibility and economic organisation .

[21] Taylor, T.B. (2018) *Judgment Day: Big Data as the Big Decider*. Ph.D. thesis, Wake Forest University.

[22] Carr, B. and Ellis, G. (2008) Universe or multiverse? *Astronomy & Geophysics* **49**(2): 2–29.

[23] Deutsch, D. (2002) The structure of the multiverse. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* **458**(2028): 2911–2923.

[24] Education, A., (AERI), R.I. and the Archival Curriculum Group (PACG), P. (2011) Educating for the archival multiverse. *The American Archivist* : 69–101.

[25] Gilliland, A.J. and Willer, M. (2014) Metadata for the information multiverse. *iConference 2014 Proceedings* .

[26] Drozdal, J., Weisz, J., Wang, D., Dass, G., Yao, B., Zhao, C., Muller, M. *et al.* (2020) Trust in automl: exploring information needs for establishing trust in automated machine learning systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*: 297–307.

[27] Martin, K. (2019) Ethical implications and accountability of algorithms. *Journal of Business Ethics* **160**(4): 835–850.

[28] Madrigal, A. (2018) Inside facebook's fast-growing content-moderation effort. *The Atlantic* .

[29] Dridi, A., Atzeni, M. and Recupero, D.R. (2019) Finenews: fine-grained semantic sentiment analysis on financial microblogs and news. *International Journal of Machine Learning and Cybernetics* **10**(8): 2199–2207.

[30] Nakov, P. (2017) Semantic sentiment analysis of twitter data. *arXiv preprint arXiv:1710.01492* .

[31] Saif, H., He, Y. and Alani, H. (2012) Semantic sentiment analysis of twitter. In *International semantic web conference* (Springer): 508–524.

[32] Saif, H., He, Y., Fernandez, M. and Alani, H. (2016) Contextual semantics for sentiment analysis of twitter. *Information Processing & Management* **52**(1): 5–19.

[33] El Alaoui, I., Gahi, Y., Messoussi, R., Chaabi, Y., Todoskoff, A. and Kobi, A. (2018) A novel adaptable approach for sentiment analysis on big social data. *Journal of Big Data* **5**(1): 12.

[34] Rajput, A. (2020) Natural language processing, sentiment analysis, and clinical analytics. In *Innovation in Health Informatics* (Elsevier), 79–97.

[35] Alowaidi, S., Saleh, M. and Abulnaja, O. (2017) Semantic sentiment analysis of arabic texts. *International Journal of Advanced Computer Science and Applications* **8**(2): 256–262.

[36] Molina-González, M.D., Martínez-Cámara, E., Martín-Valdivia, M.T. and Perea-Ortega, J.M. (2013) Semantic orientation for polarity classification in spanish reviews. *Expert Systems with Applications* **40**(18): 7250–7257.

[37] Mukku, S.S., Choudhary, N. and Mamidi, R. (2016) Enhanced sentiment classification of telugu text using ml techniques. *SAAIP at IJCAI* **2016**: 29–34.

[38] Athar, A. and Teufel, S. (2012) Context-enhanced citation sentiment detection. In *Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*: 597–601.

[39] Cambria, E., Olsher, D. and Rajagopal, D. (2014) Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In *Proceedings of the twenty-eighth AAAI conference on artificial intelligence*: 1515–1521.

[40] Yousif, A., Niu, Z., Tarus, J.K. and Ahmad, A. (2019) A survey on sentiment analysis of scientific citations. *Artificial Intelligence Review* **52**(3): 1805–1838.

[41] Amershi, S., Cakmak, M., Knox, W.B. and Kulesza, T. (2014) Power to the people: The role of humans in interactive machine learning. *Ai Magazine* **35**(4): 105–120.

[42] Gil, Y., Honaker, J., Gupta, S., Ma, Y., D'Orazio, V., Garijo, D., Gadewar, S. *et al.* (2019) Towards human-guided machine learning. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*: 614–624.

[43] Ackerman, M.S. (2000) The intellectual challenge of cscw: the gap between social requirements and technical feasibility. *Human–Computer Interaction* **15**(2-3): 179–203.

[44] Seering, J., Wang, T., Yoon, J. and Kaufman, G. (2019) Moderator engagement and community development in the age of algorithms. *New Media & Society* **21**(7): 1417–1443.

[45] Stecklow, S. (2018) Why facebook is losing the war on hate speech in myanmar. *URL: https://www. reuters. com/investigates/special-report/myanmar-facebook-hate* .

[46] Wah, C., Van Horn, G., Branson, S., Maji, S., Perona, P. and Belongie, S. (2014) Similarity comparisons for interactive fine-grained categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*: 859–866.

[47] Russakovsky, O., Li, L.J. and Fei-Fei, L. (2015) Best of both worlds: human-machine collaboration for object annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*: 2121–2131.

[48] Vondrick, C., Patterson, D. and Ramanan, D. (2013) Efficiently scaling up crowd sourced video annotation. *International journal of computer vision* **101**(1): 184–204.

[49] Peng, J., Mit, C., Liu, Q., Uci, I., Ihler, A. and Berger, B. (2013) Crowdsourcing for structured labeling with applications to protein folding .

[50] Vijayanarasimhan, S. and Grauman, K. (2009) What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE): 2262–2269.

[51] Gao, H., Barbier, G. and Goolsby, R. (2011) Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems* **26**(3): 10–14.

[52] Lintott, C.J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M.J. *et al.* (2008) Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society* **389**(3): 1179–1189.

[53] Jhaver, S., Birman, I., Gilbert, E. and Bruckman, A. (2019) Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM*

Transactions on Computer-Human Interaction (TOCHI) **26**(5): 1–35.

[54] Licklider, J.C. (1960) Man-computer symbiosis. *IRE transactions on human factors in electronics* (1): 4–11.

[55] Altintas, I., Barney, O. and Jaeger-Frank, E. (2006) Provenance collection support in the kepler scientific workflow system. In *International Provenance and Annotation Workshop* (Springer): 118–132.

[56] Bavoil, L., Callahan, S.P., Crossno, P.J., Freire, J., Scheidegger, C.E., Silva, C.T. and Vo, H.T. (2005) Vistrails: Enabling interactive multiple-view visualizations. In *VIS 05. IEEE Visualization, 2005.* (IEEE): 135–142.

[57] Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T. *et al.* (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **20**(17): 3045–3054.

[58] Pilkington, M. (2016) Blockchain technology: principles and applications. In *Research handbook on digital transformations* (Edward Elgar Publishing).

[59] Dingledine, R., Freedman, M.J. and Molnar, D. (2001) The free haven project: Distributed anonymous storage service. In *Designing Privacy Enhancing Technologies* (Springer): 67–95.

[60] Baumann, A., Peinado, M. and Hunt, G. (2015) Shielding applications from an untrusted cloud with haven. *ACM Transactions on Computer Systems (TOCS)* **33**(3): 1–26.

[61] Haeberlen, A., Mislove, A. and Druschel, P. (2005) Glacier: Highly durable, decentralized storage despite massive correlated failures. In *Proceedings of the 2nd conference on Symposium on Networked Systems Design & Implementation-Volume 2*: 143–158.

[62] Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., Kieseberg, P. *et al.* (2018) Explainable ai: the new 42? In *International cross-domain conference for machine learning and knowledge extraction* (Springer): 295–303.

[63] Holzinger, A., Kieseberg, P., Weippl, E. and Tjoa, A.M. (2018) Current advances, trends and challenges of machine learning and knowledge extraction: from machine learning to explainable ai. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction* (Springer): 1–8.

[64] Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K. and Müller, K.R. (2019) *Explainable AI: interpreting, explaining and visualizing deep learning*, **11700** (Springer Nature).

[65] Altintas, I., Barney, O. and Jaeger-Frank, E. (2006) Provenance collection support in the kepler scientific workflow system. In *International Provenance and Annotation Workshop* (Springer): 118–132.

[66] Bavoil, L., Callahan, S.P., Crossno, P.J., Freire, J., Scheidegger, C.E., Silva, C.T. and Vo, H.T. (2005) Vistrails: Enabling interactive multiple-view visualizations. In *VIS 05. IEEE Visualization, 2005.* (IEEE): 135–142.

[67] Davidson, S.B. and Freire, J. (2008) Provenance and scientific workflows: challenges and opportunities. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*: 1345–1350.

[68] Freire, J., Koop, D., Santos, E. and Silva, C.T. (2008) Provenance for computational tasks: A survey. *Computing in Science & Engineering* **10**(3): 11–21.

[69] Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T. *et al.* (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **20**(17): 3045–3054.

[70] Kurzweil, R. (2005) *The singularity is near: When humans transcend biology* (Penguin).

[71] Makridakis, S. (2017) The forthcoming artificial intelligence (ai) revolution: Its impact on society and firms. *Futures* **90**: 46–60.

[72] Joy, B. (2000) Why the future doesn't need us. *Wired magazine* **8**(4): 238–262.

[73] Cellan-Jones, R. (2014) Stephen hawking warns artificial intelligence could end mankind. *BBC news* **2**(2014): 10–10.

[74] Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press).

[75] Peckham, M. (2016) What 7 of the most world's smartest people think about artificial intelligence. *Time Magazine* (2016).

[76] Ananny, M. and Crawford, K. (2018) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society* **20**(3): 973–989.

[77] Crawford, K. (2016) Can an algorithm be agonistic? ten scenes from life in calculated publics. *Science, Technology, & Human Values* **41**(1): 77–92.

[78] Dwork, C. and Mulligan, D.K. (2013) It's not privacy, and it's not fair. *Stanford Law Review Online* **66**(35).

[79] Desai, D.R. and Kroll, J.A. (2017) Trust but verify: A guide to algorithms and the law. *Harv. JL & Tech.* **31**: 1.

[80] Ziewitz, M. (2016) Governing algorithms: Myth, mess, and methods. *Science, Technology, & Human Values* **41**(1): 3–16.

[81] Golembiewski, R.T. and McConkie, M. (1975) The centrality of interpersonal trust in group processes. *Theories of group processes* **131**: 185.

[82] Kiffin-Petersen, S. (2004) Trust: A neglected variable in team effectiveness research. *Journal of Management & Organization* **10**(1): 38–53.

[83] Hosmer, L.T. (1995) Trust: The connecting link between organizational theory and philosophical ethics. *Academy of management Review* **20**(2): 379–403.

[84] Hosmer, L.T. (1991) *The ethics of management, 2nd ed.* (Homewood, IL: Irwin).

[85] Rawls, J. (1967) *A Theory of Justice* (Cambridge Mass: Harvard University Press).

[86] (DDP), T.D.D.P. (2013) Multiple translations of comedia di dante degli allaghieri col commento di jacopo della lana bolognese, a cura di luciano scarabelli (bologna: Tipografia regia, 1866-67), as found on dante lab. *http://dantelab.dartmouth.edu* .

[87] Nakamoto, S. (2008) Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review* .