Design Science

Modular oversight methodology: a framework to aid ethical alignment of algorithmic creations

Kyriakos Kyriakou[1,2] and Jahna Otterbacher[1,2]

[1] *Fairness and Ethics in AI–Human Interaction Multidisciplinary Research Group (fAIre MRG), CYENS Centre of Excellence, Nicosia, Cyprus*
[2] *Cyprus Center for Algorithmic Transparency (CyCAT), Open University of Cyprus, Nicosia, Cyprus*

Abstract

Evaluating the algorithmic behavior of interactive systems is complex and time-consuming. Developers increasingly recognize the importance of accountability for their algorithmic creations' unanticipated behavior and resulting implications. To mitigate this phenomenon, developers not only need to concentrate on the observable inaccuracies that can be measured quantitatively but also the more subjective outcomes that can perpetuate social bias, which are challenging to identify. We require a new approach that involves humans in scrutinizing algorithmic behavior. It leverages a combination of quantitative and qualitative methods to support an ethical, value-aligned design and a system's lifecycle, informed by users' perception and values. To date, the literature lacks an agreed-upon framework for such an approach. Consequently, we propose an oversight framework, *Modular Oversight Methodology (MOM)*, which aids developers in assessing the behavior of their systems by involving a carefully crowdsourced society-in-the-loop. The framework facilitates the development and execution of an oversight process and can be tweaked according to the domain and application of use. Through such an oversight process, developers can assess the human perception of the algorithmic behavior under inspection, and extract valuable insights that will aid in assessing its implications. We present the MOM framework, as a first step toward tailoring more robust, domain-specific solutions to exercise human oversight over algorithms, as a means for software developers to keep the generated output of their solutions fair and trustworthy.

Keywords: Black-box systems, Human oversight, Algorithmic auditing, Software engineering, Quality assurance

Received 02 September 2023
Revised 04 June 2024
Accepted 07 June 2024

Corresponding author
Kyriakos Kyriakou
k.kyriakou@cyens.org.cy

Des. Sci., vol. 10, e32
journals.cambridge.org/dsj
DOI: 10.1017/dsj.2024.23

1. Introduction

Implications arising from the unpredictable behavior of algorithmic systems are a complex burden for software developers (hereon: developers). This is particularly true of systems based on artificial intelligence (AI), the current industry trend. Both the scientific community and the general public have reported numerous issues of inappropriate algorithmic behavior, where its influence can cause harm to groups or individuals (Danks & London 2017; Chen *et al.* 2017; Buolamwini & Gebru 2018; Köchling & Wehner 2020; Kyriakou *et al.* 2020, 2019; Imana *et al.* 2021;

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

Li 2023; Sun *et al.* 2023).[1,2,3,4] Although on some occasions these problematic behaviors can simply be considered inaccurate, other times they have been shown to perpetuate bias and discrimination. Consequently, problematic machine behavior can amplify known issues (e.g., rampant social stereotyping) or even result in new phenomena impacting society at large.

To minimize harm, developers have been called to take responsibility for their algorithmic creations, by testing not only their intended functionality but also by assessing potential unexpected behaviors that might bear inappropriate outcomes. The GDPR directive[5], for instance, requires developers to provide privacy-related mechanisms to give users control of their personal data and the manner in which they are processed. In a similar vein, the European Commission (EC) released its Ethics Guidelines for Trustworthy AI,[6] highlighting key principles for practitioners, which must be respected in the development, deployment and use of AI systems. For example, the Guidelines hold that developers must:

- Develop, deploy and use AI systems in a way that adheres to the ethical principles of respect for **human autonomy, prevention of harm, fairness and explicability**. Acknowledge and address the potential tensions between these principles.
- Pay particular attention to situations involving more **vulnerable groups** such as children, persons with disabilities and others that have historically been disadvantaged or are at risk of exclusion, and to **situations which are characterized by asymmetries of power or information**, such as between employers and workers, or between businesses and consumers.
- Acknowledge that, while bringing substantial benefits to individuals and society, **AI systems also pose certain risks and may have a negative impact**, including impacts which may be difficult to anticipate, identify or measure (e.g. on democracy, the rule of law and distributive justice, or on the human mind itself.) Adopt adequate measures to mitigate these risks when appropriate, and proportionately to the magnitude of the risk.

Developers should follow responsible software engineering (SE) approaches to establish fair and nondiscriminatory systems. As Schieferdecker (2020) suggested, we must develop responsible SE programs within the university syllabus and in training industry practitioners. The authors reviewed the literature to understand the ethical principles in SE and concluded that responsible SE constituents include:

- *Sustainability by design* by people in power (i.e. decision-makers).
- *Technosocial responsibility* by the software community, based on agreed-upon societal principles.
- *Responsible technology development* by the society, based on societal and sustainable development goals.

[1] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[2] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] https://www.technologyreview.com/2023/06/13/1074551/an-algorithm-intended-to-reduce-poverty-in-jordan-disqualifies-people-in-need/

[4] https://www.theverge.com/2020/8/17/21372045/uk-a-level-results-algorithm-biased-coronavirus-covid-19-pandemic-university-applications

[5] https://eur-lex.europa.eu/eli/reg/2016/679/oj

[6] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

- *State-of-the-art SE* within every software project, based on societal responsibilities.
- *Weizenbaumian oath*[7] (Weizenbaum 1972) (via professional ethics).

Although quality assurance (QA) in SE is an extensively explored domain, and testing of algorithmic systems consists of a wide collection of tools, methods and frameworks for developers to assess their creations (e.g. AI Fairness 360 (Bellamy *et al.* 2019), Aequitas (Saleiro *et al.* 2018)), it is critical to examine their applicability to intelligent systems based on AI. Marijan *et al.* (2019) necessitated the adaptation of these approaches to the context of new technologies, such as machine learning (ML). There are many open issues and challenges in the application of QA practices in traditional systems, as compared to modern intelligent systems (Marijan *et al.* 2019; Al Alamin & Uddin 2021; Felderer & Ramler 2021; Côté *et al.* 2024). Furthermore, the literature lacks a unified framework that assesses the algorithmic behavior of sociotechnical systems. Developers are left to deal with and be responsible for the implications and harms caused by the behavior of their systems. Thus, we need a new approach to enable the assessment of black-box systems' behaviors.

We envision a *modular oversight framework (MOM)* to assess algorithmic behavior, which can be tailored to the context and application of use. As will be elaborated, this framework can facilitate the use of microtask crowdsourcing for simulating a society-in-the-loop (SITL), consisting of people with diverse perspectives and values. This provides developers a new human-centric approach to assess their systems, identifying possible biases or discrimination perpetuated, while also mitigating their propagation and amplification phenomena from/to third-party systems, especially in critical applications.

2. Literature review

We review previous work, focusing mainly on the SE and developers' worldview. We then draw a connection to microtask crowdsourcing, providing basic notions, concepts and other supporting literature, to motivate the MOM framework.

2.1. Open issues and challenges in assessing the behavior of black-box intelligent systems

Because opaque systems often consist of proprietary algorithms or software where their inner workings cannot be shared or explained, monitoring and assessing system behaviors is challenging. In addition, as third-party AI-based components or services can be also integrated into these systems, they turn into a collection of black-box algorithmic components and services. This situation has magnified the complexity of assessing the behavior of such systems.

Furthermore, the literature calls attention to countless concerns around system transparency, explainability and trust. Pedreschi *et al.* (2019) provided a framework that focuses on constructing meaningful explanations of opaque AI/ML systems. They argue that explanations are vital in helping *"[software] companies for creating safer, more trustable products, and better managing any possible liability they may have"*. Similarly, Asatiani *et al.* (2020) proposed a six-dimensional

[7] The Weizenbaumian Oath was introduced by Joseph Weizenbaum (1923–2008). It focuses on the responsible use of technology, in which the tech community could commit to a set of general principles on the development, application, and usage of software systems (Schieferdecker 2020).

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

framework, along with a set of recommendations, for addressing challenges when explaining black-box behavior. Their framework's dimensions focus on (a) the model, (b) the goals [of the system], (c) the training data, (d) the input data, (e) the output data and the (f) environment [in which the AI system operates]. In another line of work, von Eschenbach (2021) presented a philosophical analysis that places transparency as the necessary condition for trust in systems. The author claims that *"[we need to] acknowledge that AI is situated within a socio-technical system that mediates trust, and by increasing the trustworthiness of these systems, we thereby increase trust in AI"*. In addition, the author emphasizes the applicability limitations of eXplainable AI (XAI).

Previous work focused on developing novel frameworks for automated black-box testing solutions. Aggarwal *et al.* (2019) proposed a methodology for auto-generating test inputs to detect individual discrimination issues in black-box behavior. Others (e.g., Viglianisi *et al.* (2020), Martin-Lopez *et al.* (2020)) created frameworks for executing an automated black-box testing via REpresentational State Transfer (RESTful) web APIs. Viglianisi *et al.* (2020) introduced RestTestGen, an approach that automatically generates test cases for REST APIs, based on their interface definition, by leveraging Swagger's information to compute an *operation dependency graph*. Martin-Lopez *et al.* (2020) presented RESTest, which uses a combination of the RESTful API parameters and novel test oracles to assess the system. Their solution followed the *constraint-based testing* technique proposed by Gotlieb (2015) that enables better coverage of the system under test (SUT) via systematically generating a combination of valid or invalid inputs, along with using novel output assertions (i.e., test oracles) (Martin-Lopez *et al.* 2020). The concept of test oracles assesses the correctness of the output of an SUT (Weyuker 1982; Marijan *et al.* 2019).

Applying test oracles over black-box systems is a challenging task on its own. Marijan *et al.* (2019) commented on the important distinction when testing *traditional systems* compared to *intelligent systems*. By *traditional systems*, we refer to systems that consist of algorithmic processes that are not based on any AI (e.g., ML, deep learning) technology or component. On the contrary, *intelligent systems* are based on AI technologies and components. Test oracles were primarily introduced for testing traditional systems (Weyuker 1982). Some of them might be *testable*, while others can be seen as *nontestable* (Weyuker 1982), something that the software testing community defined as the *oracle problem*. Pseudo-oracles was a solution proposed by Davis & Weyuker (1981), to test nontestable systems. Their technique focused on differential testing where, by using the same inputs, they query multiple software following the same specification to the SUT and observe their outputs. In addition, to assess nontestable intelligent systems, Chen (2015) adopted another kind of pseudo-oracle methodology he called *metamorphic testing (MT)*, which as Cheverda *et al.* (2022) described, can take advantage of outcomes and lead to novel testing techniques, and might aid in overcoming the *oracle problem*. As Chen states, *"MT is based on the simple intuition that although we may not be able to know the correctness of the computed output for any particular input, we may know the relation between relevant inputs and their outputs."* We argue that when developing a new methodology, it is beneficial to reflect on these thoughts, especially on MT and the relation between the provided input and generated algorithmic output, in a way that human(s)-in-the-loop (HITL) can aid in scrutinizing systems' behavior based on the human acumen (see Figure 1). Another

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

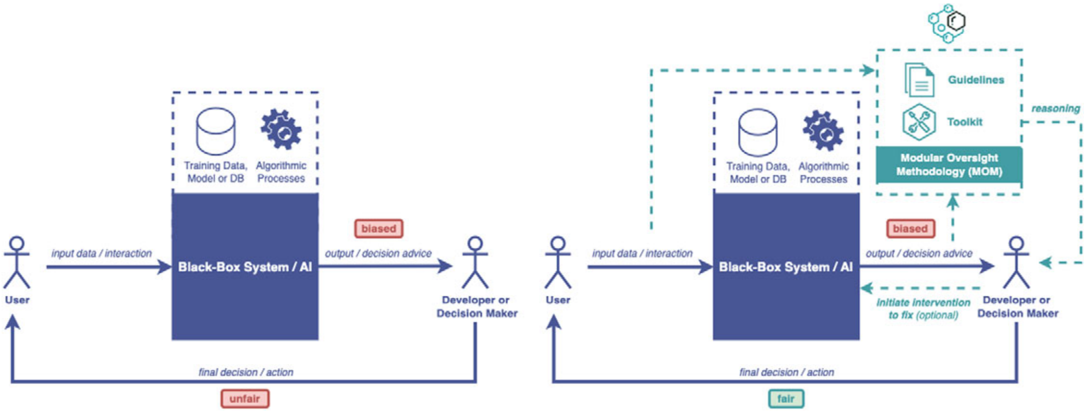

Figure 1. An example of a black-box system when applying (right) and without applying (left) MOM.

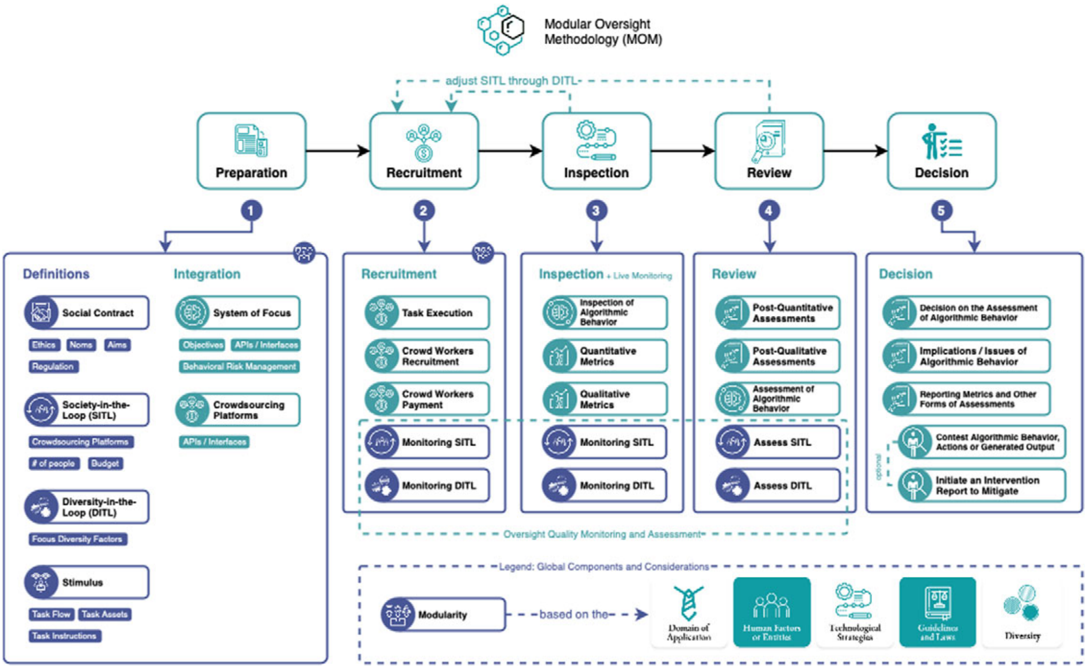
Modular Oversight Methodology (MOM)
adjust SITL through DITL
Preparation
Recruitment
Inspection
Review
Decision
1
2
3
4
5
Definitions
Social Contract
Ethics
Noms
Aims
Regulation
Society-in-the-Loop (SITL)
Crowdsourcing Platforms
# of people
Budget
Diversity-in-the-Loop (DITL)
Focus Diversity Factors
Stimulus
Task Flow
Task Assets
Task Instructions
Integration
System of Focus
Objectives
APIs / Interfaces
Behavioral Risk Management
Crowdsourcing Platforms
APIs / Interfaces
Recruitment
Task Execution
Crowd Workers Recruitment
Crowd Workers Payment
Monitoring SITL
Monitoring DITL
Inspection + Live Monitoring
Inspection of Algorithmic Behavior
Quantitative Metrics
Qualitative Metrics
Monitoring SITL
Monitoring DITL
Review
Post-Quantitative Assessments
Post-Qualitative Assessments
Assessment of Algorithmic Behavior
Assess SITL
Assess DITL
Decision
Decision on the Assessment of Algorithmic Behavior
Implications / Issues of Algorithmic Behavior
Reporting Metrics and Other Forms of Assessments
optional
Contest Algorithmic Behavior, Actions or Generated Output
Initiate an Intervention Report to Mitigate
Oversight Quality Monitoring and Assessment
Legend: Global Components and Considerations
Modularity
based on the
Domain of Application
Human Factors or Entities
Technological Strategies
Guidelines and Laws
Diversity

Figure 2. A blueprint of the MOM framework to advise software developers during a human oversight process to scrutinize algorithmic behavior. The five phases are explained in a synoptic visual representation.

reason to consider these thoughts is that MT has been established by the literature as the go-to approach for testing the validity of such systems (Segura *et al.* 2016; Felderer *et al.* 2019).

To date, many have tried to shed light on the challenges, obstacles and issues concerning the assessment of black-box systems' behavior. Cheverda *et al.* (2022) conducted a review on the state of the art of taxonomies for QA of intelligent systems to identify existing approaches, key measurable attributes for AI, statistical

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

or ML models that are commonly used and their effectiveness. Al Alamin and Uddin (2021) provided a snapshot of the existing QA issues in Machine Learning Software Applications (MLSA) by mapping various ML adoption challenges across different phases of the software development life cycle (SDLC). They found 31 challenges in total, grouping them into three main categories such as challenges in (a) data (i.e., data collection and cleaning), (b) practice (i.e., issues faced during the SE practice) and (c) standard[s] (i.e., issues arising due to the lack of standard specification or guidelines). In addition, they discussed the need for further research across various disciplines to handle the challenges of testing MLSA.

In a similar vein, Felderer *et al.* (2019; Felderer and Ramler 2021) provided an overview of the challenges in *Data-Intensive Software Systems (DISS).* Although DISS can also be described as a type of *intelligent system*, their main characteristic is the processing of big data and the conclusions they derive from them. To facilitate testing for the peculiar nature of DISS, the same authors suggest four new additions to the testing dimensions (test objective, test level and execution level). Hummel *et al.* (2018) identified eight challenges for the QA of data-intensive systems, such as the *(a) challenging visualization and explainability of results* in finding a way to balance the explanations to the user so they can better provide support and, thus, ensure trustworthiness and understandability which is difficult; *(b) nonintuitive notion of consistency* due to the large volume of data that needs to be weakened in order to avoid becoming confusing to the users; *(c) complex data processing and different notions of correctness* are difficult to define; *(d) high hardware requirements for testing* because of the big data; *(e) difficult generation of adequate, high-quality data* for testing; *(f) lack of debugging, logging, and error-tracing methods* because of the distributed nature of DISS; *(g) state explosion in verification* in which to process its requests due to the distributed nature results into an exponential number of states, and *(h) ensuring data quality* in big data, which is hard. To identify quality issues in machine learning software systems (MLSS), Côté *et al.* (2024) surveyed software development practitioners. From the interviews, they extracted 18 recurring quality issues and proposed 24 distinct strategies to mitigate them. Finally, Rosen (2020) provided an overview of the current SE environment regarding the QA of intelligent systems. The author points out that it is quite important to see QA as a matter of commitment between all the involved developers, stakeholders and their work environment, instead of simply being seen as a matter of compliance.

2.2. Responsible AI

Responsible AI is a set of recommendations for designing, developing, implementing and monitoring AI-based solutions following ethical and legal principles, frameworks and good practices (Sambasivan & Holbrook 2018; Peters *et al.* 2020; Shneiderman 2021; Lu, Zhu, Xu, Whittle, Douglas & Sanderson 2022; Matsui & Goya 2022a; Soklaski *et al.* 2022). Through responsible AI, developers can ensure the genesis of AI solutions that focus on benefiting society by adhering to its values while minimizing possible negative impacts or harms.

In the context of black-box intelligent systems, the literature demonstrates urgency in designing and following approaches that enable and promote Responsible AI. Islam (2021) aimed to develop SE methods for responsible AI by which ethical considerations can be addressed throughout the systems' SDLC. His

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

framework aimed to promote ethical, legal, social, economic and cultural values by converting them into functional specifications to make the system's objectives transparent. This is in line with the work of Schieferdecker (2020), who also argued that we need to consider extending software quality with considerations for societal impact, transparency, fairness and trustworthiness. Thus, there is a need not only for new approaches and tools but also for relevant updates in processes and regulations, to achieve such changes.

Many have tried to provide a snapshot of the current practices and needs around responsible AI. For instance, Lu, Zhu, Xu, Whittle, Douglas and Sanderson (2022) conducted an empirical study by interviewing 21 scientists and engineers to understand the practitioners' views on AI ethics and their implementation. The authors crafted a preliminary list of 17 operationalized responsible AI assurance process/design patterns. Among others, some of them are the *extensible adaptive, dynamic risk assessment*, which requires continuous adaptation in assessing systems in distinct contexts, the existence of *standardized documents* compliant with standards and accessible by stakeholders, and a *decision mode switcher* with which the system can switch to fully automatic or semi-automatic with the aid of HITL (kill switch, override, fallback) along with a *human-centered explainable interface*, a set of *ethical acceptance tests*, predefined metric(s) to *audit the black-box* in a continuous run-time validation *(continuous validator)*. Following up, Lu *et al.* (2023), expanded this work and elaborated further on the explanation of their pattern collection for Responsible AI. In parallel, Lu, Zhu, Xu, Whittle and Xing (2022) provided a roadmap for Responsible AI in SE that focuses on: *"(a) establishing multi-level governance mechanisms for Responsible AI systems, (b) setting up the development processes incorporating process-oriented practices for responsible AI systems, and (c) building Responsible AI-by-design into AI systems through system-level architectural style, patterns and techniques"*.

2.3. From DevOps to MLOps

Most SE companies have adopted development operations (DevOps) practices. DevOps is a "set of practices and tools focused on software and systems engineering"[8] (Ebert *et al.* 2016; Sharma 2017; Symeonidis *et al.* 2022), to facilitate adequate communication and collaboration between developers (development and QA teams) and operation teams to improve the quality of service (Farroha & Farroha 2014; Fitzgerald & Stol 2017; John *et al.* 2021; Gift & Deza 2021). DevOps reduces the time for developing and deploying a software change while keeping high-quality delivery standards (Zhu *et al.* 2016). In DevOps, there are two main principles: *continuous integration* and *continuous delivery*. *Continuous integration* is the practice in which code is integrated at frequent intervals after thorough testing and application of required improvements (Raj 2021; Symeonidis *et al.* 2022). *Continuous delivery* is the practice by which a new software version is constantly ready to be tested, evaluated and then released in production (Karamitsos *et al.* 2020; Symeonidis *et al.* 2022).

Although the benefits of practicing DevOps in traditional systems are many, previous works criticize its inability to support the development of AI-based intelligent systems. While some emphasize the need to apply the same principles

[8] https://devops.com

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

that govern DevOps in ML models (Alla & Adari 2021), many suggest a new approach they call *machine learning operations* (MLOps) that basically fuses ML and DevOps practices (Alla & Adari 2021). This trend created a paradigm shift for MLOps, in an effort to improve the development cycle of ML applications.

Recent work aims to define MLOps, mapping its basic concepts. Testi *et al.* (2022) sketched out a taxonomy of MLOps accompanied by a standardized methodology for its application. Specifically, the authors describe a set of MLOps methodologies revolving around the *business problem understanding, data acquisition, ML methodology, ML training and testing, continuous integration, delivery, training and monitoring, XAI* and *sustainability (of the carbon footprint).* Symeonidis *et al.* (2022) presented an overview of the MLOps area by listing definitions, tools and challenges derived from the literature. John *et al.* (2021) provided a framework for the adoption activities involved in MLOps and a model that classifies companies' maturity levels with respect to their MLOps practices. Similarly, Matsui and Goya (2022a,b) identified five steps to guide the understanding and adoption of MLOps in the context of responsible AI.

In the MLOps pipeline, *continuous monitoring* of the data and model becomes a key aspect of this practice (Matsui & Goya 2022a,b; Symeonidis *et al.* 2022). Research has demonstrated that the accuracy of an ML model may decay over time because of the nonrepresentative training data compared to the new data in production (Ruf *et al.* 2021; Matsui & Goya 2022b). Other problematic phenomena, such as *model degradation* (Treveil *et al.* 2020), or *data* or *concept drift* (Treveil *et al.* 2020; Hussain *et al.* 2021) have also been described. *Continuous monitoring* is a practice that enables "the identification of risks and maintenance of the model in production, aligned with the business metrics" (Schlossnagle 2018; Matsui & Goya 2022a). In fact, Matsui and Goya (2022b) argued that "creating metrics and tracking changes in data [and the model] makes it possible to identify the root cause of deviation."[9]

At present, MLOps is still in its infancy (van den Heuvel & Tamburri 2020; Testi *et al.* 2022; Matsui & Goya 2022a; Symeonidis *et al.* 2022). With time, it is expected that useful practices will be established, to extend the MLOps toolbox, while helping developers overcome its present challenges.

2.4. The need for human oversight over black-box systems

Despite the technological growth in algorithmic and AI systems, we are still in need of robust methods to identify the possible risks and harms of those systems. This is why many have called for exercising *human oversight* over these algorithmic processes and scrutinizing them, both in terms of their intended and unexpected behaviors.

In its Ethics Guidelines for Trustworthy AI,[10] the European Commission (EC) refers to *human agency*, highlighting the protection of individual [users'] autonomy, which must be central to the system's functionality. Per the EC, the key to this is a human's right not to be subjected to a decision based solely on automated processing when this results in significant consequences. Similarly,

[9] https://learn.microsoft.com/en-us/azure/machine-learning/how-to-monitor-datasets?view=azureml-api-1&tabs=python

[10] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

human oversight helps ensure that an AI system does not undermine human autonomy or cause other adverse effects. Oversight can be achieved using various governance mechanisms that involve HITL.

As previously described, the nature of *traditional systems* compared to *intelligent systems (AI-based)* is quite different (Marijan *et al.* 2019). As *traditional systems* are developed based on rule-based and pre-programmed algorithmic procedures, we can describe them as deterministic systems, in which an input always results in the same outcome. However, *intelligent systems* are developed to generate reasoning in a probabilistic manner, exhibiting nondeterministic behavior, in which an input does not always result in the same outcome. This makes the QA process for this last category of systems even more complex, especially in critical and/or high-risk applications such as healthcare, security, self-driving cars, financial institutions, governmental services, etc. (Al Alamin & Uddin 2021).

Many have noted that we require a new approach to inspecting those systems and assessing their behavior (Marijan *et al.* 2019; Schieferdecker 2020). Similarly, we argue that another kind of framework is needed to approach the problem using a human-centric lens, in which ethics, regulation, human values and perception are placed at the center. More specifically, we envision a modular approach that can be tailored and extended to focus on specific domains and applications of use. We do not characterize this methodology as a generic auditing approach, but rather, as the first step for constructing this framework as the foundation for expanding into other domains and applications of use later.

Our framework is fundamentally inspired by the concept of SITL proposed by Rahwan (2018). Rahwan described SITL as the pact between various stakeholders – including those affected by the system – that might bear competing interests and values, mediated by machines. Basically, it involves HITL monitoring the compliance with the agreement based on an agreed-upon *algorithmic social contract*. Rahwan described the essence of SITL through a simple equation: *SITL = HITL + algorithmic social contract*. He derives the notion of *social contract* from the domain of political philosophy to articulate the *algorithmic social contract*. In this kind of social contract, society must agree on two important aspects:

1. Society must resolve trade-offs between the different values that AI can strive toward (e.g., different notions of security, privacy or fairness).
2. Society must agree on which stakeholders will reap which benefits and pay which costs (e.g., what is acceptable or even which degree of collateral damage is acceptable).

To date, SITL remains a theoretical concept consisting of high-level principles. As such, it is difficult to implement practices, tools and methods in a real-world setting. SITL can be seen as a noble objective but true SITL is nonetheless unrealistic. For example, one of the biggest challenges for SITL is finding a way to balance the competing interests of different stakeholders, including the interests of those who govern through algorithms (Rahwan 2018). Through our framework, we approximate the notion of SITL by managing it through diversity-in-the-loop (DITL), providing a workaround for these challenges. By DITL, we refer to the set of attributes and characteristics of a recruited “crowd” to scrutinize an algorithmic behavior that can be denoted into objective [hereon: objectual] (i.e., race and gender) and subjective [hereon: functional] (i.e., human perception) (Giunchiglia *et al.* 2021)

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

diversity factors. Such factors might be influential to the public's perception of the observed algorithmic behavior. Depending on the context and application of the intended system, the analyst should consider the relevant diversity factors. We further elaborate on the diversity considerations later in the article.

To simulate an SITL, we exploit the capabilities of microtask crowdsourcing, to form a microsociety of crowd workers [hereon: workers] and leverage the "wisdom of the crowd". Previous work in crowdsourcing attempted to use microtask crowdsourcing techniques to audit or monitor various systems for faulty behavior. For instance, Nushi *et al.* (2018) proposed a set of hybrid human–machine methods and tools for describing and explaining system failures, by leveraging both system-generated and human observations gathered from micro-tasks. Bansal *et al.* (2019) simulated an abstract version of an AI-advised human decision-making system and used crowdsourcing to study the role of mental models in team performance in such environments. Other studies focused on user trust (Honeycutt *et al.* 2020) and the accuracy of post-hoc interpretations (Shen & Huang 2020).

It is crucial to consider the multi-disciplinary nature of the revolving topics around the matter of *human oversight* of algorithms. It is important to propose a foundational framework, grounded by core essential components, deriving from each of the relevant domains (e.g., computer science, sociology, and law) to obtain the maximum benefit from such an oversight process.

3. The modular oversight methodology

We dive deep into the specifics of exercising human oversight over algorithmic behavior. We describe the modular oversight methodology (MOM), which aids software developers in carrying out an oversight process to scrutinize the behavior of their algorithmic creations. First, we provide a definition for MOM. In addition, we elaborate further on its aims and purpose, describing how MOM differs from other approaches. Next, we present the core components required to consider when applying MOM. Finally, we discuss the basic phases of MOM by providing further information on the actions needed in each of those phases.

3.1. Definition

MOM is a framework that facilitates *human oversight* over black-box algorithmic processes, by simulating an SITL process. We approximate the SITL process via the use of DITL mechanisms, leveraging microtask crowdsourcing techniques to gather the perception and "wisdom of the crowd," to assess algorithmic behavior based on an agreed-upon *algorithmic social contract* (including ethics and regulation) and contest its outcome. The framework is modular at the core, which means it can be adapted depending on the context and application of use. Utilizing it consists of five distinct phases: (1) preparation, (2) recruitment, (3) inspection, (4) review and (5) decision.

3.2. Purpose

This framework is created predominantly for – but is not limited to – software developers and companies who build opaque, proprietary systems, often integrating owned or third-party black-box components and services that might be

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

AI-based. We envision this foundational framework in the developers' gamut of tools for assessing the behavior of their systems in a human-centric way, an important aspect currently neglected by most QA approaches. Thus, the main aim of MOM is to facilitate the process of human oversight on target algorithmic processes when their behavior might bear societal risks for impacting – either directly or indirectly – people's lives.

More specifically, there is a multifold set of objectives that the MOM framework can support, as listed below:

- Facilitating *human oversight* of algorithms by ensuring that *human agency*, and as a result human autonomy, is in place.
- Providing a plug-and-play modular framework for scrutinizing algorithmic processes that can be tweaked depending on the context (e.g., in financial institutions for granting loans to individuals) and applications of use (e.g., computer vision systems – in the context of automated security).
- Recruiting a crowd-in-the-loop as an SITL, with diverse perspectives on a targeted matter to contest the algorithmically generated advice and, as a result, its overall behavior from a multidimensional worldview.
- Providing an accessible crowd to simulate an SITL, by monitoring and managing DITL to maximize the benefits of such an oversight process.
- Extracting and reporting inappropriate algorithmic behavior phenomena such as perpetuated biases and discrimination from a human-centric lens.
- Contemplating human norms, ethical values and national or international regulations that apply per context and application of use.
- Simulating the production environment and interaction of the end-users.
- Enabling applicability during the design phase of the system (predevelopment), the development, and after the deployment of the system (postdevelopment) to monitor and assess its operation.
- Enabling compatibility with existing SE practices such as DevOps, MLOps and QA.
- Enabling compatibility with agile SE approaches as it minds agile practices by design.
- Supporting and providing a reliable and accountable human-centric approach for examining algorithmic behavior that increases developers' trust in their own systems or other candidate third-party components and services to be integrated.
- Supporting compliance with the relevant regulations affecting intelligent systems and the way they affect groups of people or individuals.

The MOM framework is not limited to the objectives above; rather, those objectives form the foundation of the framework toward enabling the composition of domain-specific oversight solutions. For example, future works could expand in exercising human oversight of applications focused on the healthcare domain, by tailoring the modules of this framework according to the goals of the algorithmic system and the organization that developed or operates the system of focus. Moreover, the actors are not limited to software developers and companies but can be any other interested parties that want to contest an algorithmically mediated decision. For instance, nonsoftware companies and organizations, or governmental institutes that utilize these systems to automate decisions, can also use this framework by asking a developer to provide a tailored solution based on their needs to ensure the equal and fair use of such systems. However, in this work, we

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

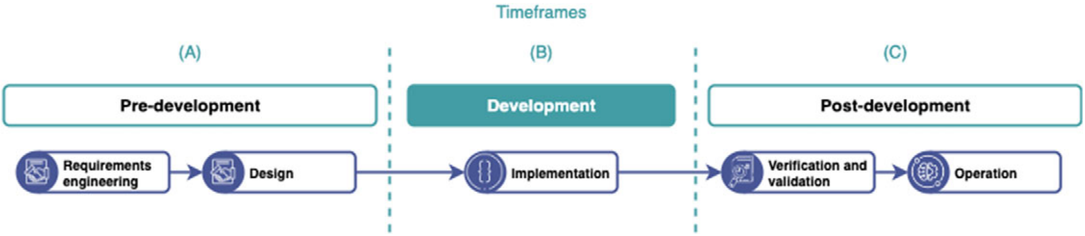

Figure 3. The applicable timeframes of MOM framework, based on the summarized SDLC practices in SE by Lu, Zhu, Xu, Whittle and Xing (2022).

intentionally narrow our scope to software developers and companies to provide a thorough overview and definition for our proposed framework by limiting any additional noise other applications might bear.

We envision the MOM framework as providing the groundwork for a new kind of *testing oracle* similar to *differential* and *adversarial testing* (Marijan *et al.* 2019), simulating the real-world input from a diverse "society," recruited through micro-task crowdsourcing. In this *testing socio-oracle*, SITL contributes to assessing the behavior of target systems by feeding a number of inputs to challenge the algorithmic outcome, based on individuals' worldviews and perceptions, which are usually characterized by contradicting views and values. Likewise, this testing technique relates to the idea of *metamorphic testing* where we need to consider formalizing and studying the relation of input–output behavior to assess black-box systems (Felderer *et al.* 2019). To sum up, the *testing socio-oracle* will in fact place the system behavior assessment into the sphere of societal values, ethics and norms by operating as in the real-world setting. Our proposed framework aims to follow this kind of *testing socio-oracle* concept to facilitate *human agency and oversight* in algorithmic systems.

It is important to specify the application timeframe of the MOM framework. As MOM is adaptable to different contexts and applications of use, we argue that it can also be applicable to different timeframes during the SDLC or lifetime of the system. Again, we illustrate this through the developer's worldview, placing the *development phase* at the epicenter of this timeframe. As a result, we divide this timeframe following the development process practices in SE summarized by Lu, Zhu, Xu, Whittle and Xing (2022).[11] To be exact, we observe three major timeframes: the *predevelopment, [during] development* and *postdevelopment* periods as shown in Figure 3. Mapping the SE processes and practices will help us understand the different purposes and practices needed to extend MOM's application according to the respective timeframe.

Consequently, we map the *predevelopment* timeframe, consisting of the *requirements engineering* and the *design* stages. In this timeframe, developers and involved stakeholders alike would be able to collaborate to determine the end-user requirements and design the system. When black-box systems are involved in any of the two stages, developers can apply the MOM framework to

[11] For the sake of simplicity, we have followed the conceptualization of Lu, Zhu, Xu, Whittle and Xing (2022). We recognize that in modern software engineering the conceptualization of the process might include other phases, i.e., considering pre- and post-deployment. In any case, MOM can easily be mapped to the appropriate phases and contexts of use.

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

assess the behavior of the analogous opaque components before they finalize their decision on the system's architecture. For example, when a commercial third-party computer vision service (AI-based) – that is attributing predicted short descriptions on photos depicting people – is under consideration for potential integration within the inner workings of the system under design, and in parallel, there are risks on the perpetuated behavior of the service, developers can apply the MOM framework to separately assess the behavior of such components and then conclude to their use or abandonment. If, for example, the service can be found to discriminate against a certain group of individuals (e.g., systematically attributes stereotypical descriptions on photos of darker-skinned individuals), the developer can decide whether a workaround should be developed to monitor/mitigate the problem and integrate it into the system (a design decision), or completely abandon the respective service and move on in finding other candidate alternatives. This practice will decrease potential societal issues derived by those third-party components to be perpetuated into the new system that will amplify or propagate the problem to other functions and contexts.

Correspondingly, in the *development* timeframe, which consists of the *implementation* stage, developers can apply the aforementioned approach during the implementation of the system. Appending to the previous example, developers would be able to assess, in an agile approach, whether the developed system functions could operate in a fair and nondiscriminatory manner.

Finally, is the *postdevelopment* timeframe, which consists of the *verification and validation* and *operation* stages. During this timeframe, developers, in collaboration with other relevant stakeholders, can verify and validate the algorithmic behavior of a complete version of the developed system that is close to deployment by applying the MOM framework while simulating the real-world setting. Finally, the last stage of the postdevelopment timeframe, the operation (or after deployment) stage, can be seen as an ongoing lifetime monitoring of the system, analogous to the discussion on continuous monitoring in MLOps (Schlossnagle 2018; Tamburri 2020; John *et al.* 2021; Matsui & Goya 2022a; Testi *et al.* 2022) to overcome issues revolving around the evolution of a system based on newly introduced data and user interaction over time (i.e., model drift) (Treveil *et al.* 2020; Hussain *et al.* 2021). In this way, issues of inappropriate algorithmic behavior would be surfaced and developers would be able to mitigate them before or at the time they occur. We summarize the above example in Figure 4.

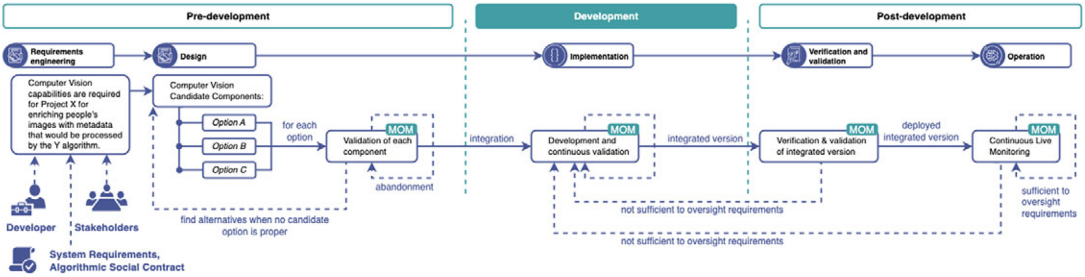
Pre-development
Development
Post-development
Requirements engineering
Design
Implementation
Verification and validation
Operation
Computer Vision capabilities are required for Project X for enriching people's images with metadata that would be processed by the Y algorithm.
Computer Vision Candidate Components:
Option A
Option B
Option C
for each option
MOM
Validation of each component
abandonment
find alternatives when no candidate option is proper
Developer
Stakeholders
System Requirements, Algorithmic Social Contract
integration
MOM
Development and continuous validation
integrated version
MOM
Verification & validation of integrated version
deployed integrated version
MOM
Continuous Live Monitoring
not sufficient to oversight requirements
sufficient to oversight requirements
not sufficient to oversight requirements

Figure 4. An example of a third-party computer vision component to be integrated into a system for choosing among job candidates. We illustrate the oversight involvement during each timeframe as presented in Figure 3.

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

3.3. Core components

We further elaborate on the core components the MOM framework builds upon and describe their essential elements.

3.3.1 Society-in-the-Loop

SITL (Rahwan 2018) is a key concept upon which the MOM framework builds. We view SITL as the backbone of the framework for infusing the human-centric aspect into it. As we focus on evaluating – often subjective – unpredictable algorithmic behavior, we argue that human perception is the key to inspecting the effects of its generated outcomes while minimizing the possible risks of such systems and the harmful impact they might have on society. In addition, with SITL, we might identify at least some of the *unknown unknowns* we cannot determine prior to the investigation for biased and discriminatory behavior. Thus, individuals' contradicting views and values are crucial.

Algorithmic social contract. Rahwan (2018) advocated for an *algorithmic social contract* derived from the notion of *social contract.* In the political philosophy domain, *social contract* is largely perceived as a contract by which the society, government or moral principles rely on agreements between voluntary agents for their existence (Seabright *et al.* 2021). Influential political philosophers such as Hobbes (1651), Rousseau (2003, 1964) and Locke (2013) have heavily contributed to the fundamental notions of the *social contract theory.* Although they have various disagreements, they also have their commonalities (Seabright *et al.* 2021). Based on Seabright *et al.* (2021), some important commonalities are the quality of being human (referred to as *the state of nature*, revolving different qualities and fundamental rights [e.g., being equal]), the existence of a contract by which humans agree with each other forming a united force to decide in a way that is being considered representative to "the will of all" (referred to as *the contract of association*), and the humans' voluntary surrendering of some individual liberty and the promise to obey the government (referred to as *the government contract*).

While many researchers and theorists claim that diverse values, ideals and settings would be beneficial in the context of the social contract, others debate this from the lens of its practical implications where their adoption is *"unlikely to reflect the direct participation and consent of all concerned"* (Jos 2006). Thus, the existing literature on the social contract is still fuzzy due to the magnitude of disagreements within the community (Campos 2019). Furthermore, Campos (2019) realized that such a contract *"does not merely determine which acts are right and wrong but it also establishes what reasons and forms of reasoning are justifiable,"* which is a fundamental point, especially in the context of weaving a dedicated social contract for exercising human oversight of algorithms.

Therefore, Rahwan (2018) described an *algorithmic social contract* as the contract by which an SITL can scrutinize algorithmic behavior by embedding the *general will.* As already stated, SITL consists of involved key stakeholders who share – sometimes conflicting – values and ideals to agree upon the tradeoffs between the different values that system behavior is expected to perpetuate. As a result, the algorithmic social contract consists of the aforementioned expectations along with the particular benefits the stakeholders should be focused on achieving with the use of the system and the corresponding costs they should pay for, in the face of the implications emanated by incidents of unexpected algorithmic behavior.

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

Based on Rahwan's suggestion to take into consideration the types of behaviors people expect from these systems, we argue that there is a need to examine these depending on the context and application of use. In addition, Rahwan (2018) emphasized the urgency of developing new mediums such as methods, metrics and tools to enable the public to articulate any expectations based on goals, ethics, norms and the social contract itself into these systems and, by the necessary programming, debugging and monitoring, enforce the *algorithmic social contract* between humans and algorithms.

Further to this, we envision an *algorithmic social contract* in which some core elements would be considered vital for its composition in our SITL-inspired framework. Although we elaborate further on those core elements in the next sections of this work, we sketch out the basic structure of the respective contract. To begin with, depending on the context and application of use, we need to be sure that diversity is in place, to mediate the fair and equal representation of a crowdsourced society of workers via DITL. In this sense, we not only need to consider national and international ethics and laws but also basic societal values, ideals and norms targeting the actions and effects of the algorithmic behavior. Developers must shift towards a human-centric worldview consisting of the aforementioned social artifacts. As Weizenbaum (1972) suggested, *"the [algorithmic] revolution need not and ought not to call [hu]man's dignity and autonomy into question that is a kind of pathology that moves [humans] to wring from it unwarranted, enormously damaging interpretations."*

Microtask crowdsourcing

There is a growing research community that strives to provide solutions to real-world problems by recruiting people (a "crowd") to contribute to defined tasks with the aid of technology. Estellés-Arolas and de Guevara (2012) systematically analyzed the scholarly literature for different interpretations of crowdsourcing and concluded that:

> *Crowdsourcing is a participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage what the user has brought to the venture, whose form will depend on the type of activity undertaken.*

Adding to that, Brabham lists the three key ingredients of crowdsourcing:

1. *An organization that has a task it needs [to be] performed,*
2. *A community (crowd) that is willing to perform the task voluntarily,*
3. *An online environment that allows the work to take place and the community to interact with the organization, and*
4. *Mutual benefit for the organization and the community.*

Moreover, in a review of the key design elements of a crowdsourcing initiative, Karachiwalla and Pinkow (2021) identified four dimensions, namely, the *(a) task,*

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

(b) crowd, (c) platform and *(d) crowdsourcer*, which is aligned with Brabham's list of the key ingredients of crowdsourcing (Brabham 2013). Karachiwalla & Pinkow's main contribution is to deliver a blueprint for practitioners to utilize when designing and executing crowdsourcing projects.

Crowdsourcing falls into two main categories: *microtask* and *macrotask* crowdsourcing. *Microtask* crowdsourcing is a helpful instrument for combining human and machine intelligence (i.e., for improving the accuracy of an ML model). *"Microtasks are relatively quick, simple, and repeatable activities that can be – and often are – completed by volunteers in parallel, without the need for specific training or specialist knowledge"* (Ibáñez *et al.* 2020). On the other hand, *macrotask* crowdsourcing is a more complex procedure that requires much more time (even hours) to complete by individuals who are specialized in the corresponding domain and context of the task (Haas *et al.* 2015; Ibáñez *et al.* 2020). To support our framework, we focus on *microtask* crowdsourcing, involving the public in the scrutinization of algorithms. In the following sections, we argue that recruiting individuals with specific characteristics (e.g., current occupation) might be helpful for the context and application of the use of the corresponding task. This maintains a more sophisticated crowd during a *microtask* crowdsourcing oversight process without eliminating the nature of *microtask* crowdsourcing per se (i.e., the speed of completing a task).

Numerous works on *microtask* crowdsourcing have contributed to this domain from various perspectives. The work of Gadiraju *et al.* (2015) focuses on distinguishing between trustworthy and untrustworthy workers' behavior. Others focused on the psychological and sociological reasons behind workers' participation in crowdsourcing (Deng & Joshi 2016). From an SE perspective, Zhen *et al.* (2021) investigated the effects of microtask crowdsourcing to enhance or support SE. In another systematic review, Zulfiqar *et al.* (2022) found a number of microtask activities from the workers' perspective.

During the past years, there is an increased interest in crowdsourcing platforms (CPs) (Brabham 2013; Liu 2020; Zhen *et al.* 2021) as many commercial solutions have emerged. Some of the most common ones – that are also highly cited and used by the research community – are Amazon Mechanical Turk (MTurk),[12] Prolific,[13] Clickworker[14] and Appen's Crowdsourcing Solutions.[15] These platforms offer a set of features that can aid in various tasks, ranging from simple customer discovery to labeling datasets and training ML models. As mentioned, previous work in *microtask* crowdsourcing combined HITL approaches to assess, monitor and improve systems' performance. The benefits of HITL are reported in the crowdsourcing literature, especially when using microtasks, and particularly when ethical concerns present challenges from a technological perspective that cannot easily be addressed by automated solutions. The controversial work of Awad *et al.* (2020) is a representative example of crowdsourcing moral judgments on machines via HITL. The authors created a web platform called "The Moral Machine (MM)" to gather data on human perception of the moral acceptability of decisions made by automated vehicles faced with choosing which humans to harm and which to

[12] https://www.mturk.com
[13] https://www.prolific.co
[14] https://www.clickworker.com
[15] https://appen.com

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

save. Moreover, they emphasized their belief that *"social scientists and computational social scientists have a pivotal role to play as intermediaries between engineers and humanities scholars in order to help them articulate the ethical principles and priorities that society wishes to embed into intelligent machines"* (Awad *et al.* 2020). In a similar line of work, Nakao (2022) used MTurk to study workers' perceptions of fairness in a hiring process scenario where an HR department uses an AI tool found to discriminate against female job applicants. Nakao's work underscores the need to embed HITL AI fairness perception and metrics when AI is used in diverse decision-making processes.

Microtask crowdsourcing is a vital ingredient for the MOM framework. By utilizing it, we can simulate SITL by convening a diverse crowd for exercising oversight in algorithms and monitoring its diversity in various aspects through DITL. Crowdsourcing could bridge the gap in time and cost when recruiting a diverse crowd, according to the needs of the oversight process. The ease of crowd pool availability and the capabilities and task types that can be facilitated make the MOM framework an important control tool for developers to assess their systems. Bearing in mind that each CP provides a different crowd community (i.e., based on region, ethnicity, age, occupation and more), the idea of combining those communities into a unified solution – under the MOM framework – will further empower developers' toolbox.

Diversity

Another core component of the MOM framework is diversity. We should observe diversity from various perspectives, concerning different levels and aspects of it.

Although *diversity* can be depicted differently depending on the discipline and there is little consensus on the precise terminology, we likewise focus on the socio-technical notion of diversity derived by Drosou *et al.* (2017). Drosou *et al.* (2017) defined diversity as a concept able to characterize the *quality* of a collection of items and consists of various interpretations depending on the context and application of use. Another crucial aspect we need to consider depending on the context of use is the notion of *novelty*, which as the same authors (Drosou *et al.* 2017) mention is usually being used in specific applications to reduce redundancy. As a result, *novelty-based diversity* is defined with respect to what has been observed in the past, up to the current point in time.

An essential element of using the MOM framework is the *diversity of workers*, and as a result, the diversity of the crowdsourced SITL. Particularly, we are concerned with how personal characteristics correlate to or affect workers' perceptions of algorithmic behaviors and their responses to the given tasks during the execution of the MOM framework when exercising human oversight. Therefore, we adopt the two types of diversity described in Giunchiglia *et al.* (2021), Giunchiglia and Fumagalli (2017) and, Schelenz *et al.* (2021). Namely, we distinguish those two types into *objectual* and *functional diversity.*

Objectual diversity – referred to as "observable diversity" or "surface diversity" (Giunchiglia *et al.* 2021) – is the facet of diversity that applies to "observable" characteristics such as sex, race, ethnicity, national origin, age, membership in formal organizations (e.g., religious, political) or physical features. It can be seen as the set of common observable attributes that describe an individual actor at a superficial level. On the contrary, *functional diversity* (Giunchiglia *et al.* 2021)

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

applies to less observable characteristics. Such characteristics include one's technical abilities, role in an organization, socioeconomic and cultural background, personality traits, cognitive abilities and values. As these less observable attributes tend to characterize an individual actor at a deeper level of understanding, this type of diversity takes longer to recognize in others. Despite that it is usually what we exploit when considering online contexts – where prolonged interactions are not always supported – identifying functional diversity becomes even more challenging and complex. Considering both types of diversity is crucial for navigating through different views and values during the crowdsourced SITL oversight process.

The benefits of considering diversity during an oversight task have been reflected in previous research. For instance, in their crowdsourcing study for investigating with the inclusion of a number of predictors in algorithmic decision-making in the context of recidivism, van Berkel *et al.* (2019) found evidence that more diverse groups tend to more closely align with the majority agreement. In fact, some researchers (van Berkel *et al.* 2019; Suresh *et al.* 2021; Grgić-Hlača *et al.* 2022) suggested that it is crucial to include workers with diverse interests and perspectives, as to consider the range of social consequences across domains of algorithmic processes, while others (Nakao 2022) argued that using crowdsourcing techniques can aid in recruiting diverse stakeholders and aggregating their views (e.g., about fairness).

Unfortunately, the literature lacks methods, metrics and tools that can aid in determining the *diversity factors* in such oversight tasks, especially those that might influence the workers' perception of the observed algorithmic behavior. In this context, the scientific community has to prioritize understanding these *diversity factors* that could be crucial in fairly and adequately assessing algorithmic behavior in a beneficial manner that promotes the maintenance, co-existence and interaction with algorithmic systems in a societal environment.

Diversity factors. When we refer to *diversity factors* from the worker's perspective, we refer to a set of attributes that characterize an individual in an objective (e.g., physical characteristics) or subjective (e.g., fairness perception) manner that can be distinguished into two main types, *objectual* and *functional* diversities. We argue that these factors are able to influence the workers' perception during an oversight process in various ways. For instance, Grgić-Hlača *et al.* (2022) found that individuals with personal experiences that are closely related to the decision-making setting, assess algorithmic fairness differently compared to those who did not have such experiences. More specifically, the authors provide the example of having attended a bail hearing, which negatively correlates to the perceived fairness when using defendants' juvenile criminal history information for making bail decisions. In this case, *personal experiences* might be a diversity factor worth considering and investigating. Of course, further work is needed to define a comprehensive set of *diversity factors.*

It is evident the *diversity factors* can have different effects and influences depending on the context and application of use. This is why, when designing an oversight task using the MOM framework, we suggest carefully choosing the factors that could be important for the respective oversight application, and the aims of the developer or organization that exercise oversight. We believe that their careful usage and monitoring would bring only benefits, allowing the parties to successfully and effectively assess algorithmic behavior. That said, the use of

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

diversity factors is not necessarily easy; sometimes it can be complex and other times even controversial. Thus, these factors should be seen as a set of tools that are equipped in the MOM framework to sufficiently facilitate oversight of algorithms, rather than a "golden solution" that can aid in all circumstances.

Finally, we have to point out the interrelated nature of *diversity factors.* They should not be interpreted as individual elements that are mutually exclusive, but rather as a set of factors that might influence each other. This is another challenge the literature must address when it comes to the actual effects of the set of *diversity factors* in this domain of knowledge. While our current focus is on describing the high-level oversight framework (i.e., an investigation of the specific diversity factors that could be involved in MOM is out of our current scope), future work must address the identification of the key diversity factors for a given oversight context (e.g., which diversity factors should be considered when crafting an oversight process to be used for assessing the behaviors of algorithmic decision-making in the banking context, etc.)

DITL. True SITL is a noble but impractical goal; we will never be able to involve an entire society in oversight. Rather, we aim to approximate SITL through diversity management, a process we call *DITL.* Utilizing DITL, we monitor the *diversity factors* of interest to manage the composition of the crowdsourced oversight force during the entire time of the oversight process; from the design *(preparation phase)* and execution of the process *(recruitment and inspection phase)* to the *review* of the results and the final *decision* (advise the phases of the MOM framework depicted in Figure 2). DITL can be a crucial tool for managing diversity by which we can maximize the utilization of its benefits in the oversight context and leverage a diverse set of views and values derived from the *"wisdom of the crowd."* In other words, DITL is a tool approaching the notion of SITL, by using a predefined set of *diversity factors* of interest to monitor the composition of the crowdsourced society during a human oversight process, depending on the context and application of use.

Ethics

Because of the human-centric nature of the MOM framework, ethics are an integral part of it. Numerous discussions reflect the need for considering human values and ethical principles in an oversight process, especially during the past few years. From an EU setting, the EC's proposed ethics guidelines on Trustworthy AI[6] reflect four basic ethical principles, such as (a) respect for human autonomy, (b) prevention of harm, (c) fairness and (d) explicability. In another EU initiative called AI4People, Floridi *et al.* (2021) synthesized five ethical principles in the context of algorithmic systems that have to be considered during the SDLC. More specifically, the authors based their recommendations on the existing principles in bioethics to redefine a similar set of principles such as *beneficence, nonmaleficence, autonomy, justice* and *explicability*. Others (Mittelstadt *et al.* 2016) reviewed the discussion around algorithmic ethics, providing a perspective map to organize the debate and assess the literature to identify areas for improvement toward the development of algorithmic ethics. Social norms are another aspect worth considering when dealing with ethics. For example, when parallelizing the algorithmic with human behavior, we should ask a few key questions: How do people act in X situations?

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

What are the norms behind driving this act? Which of those norms should we consider applying in or assessing by, the current algorithmic solution?

Of course, ethical considerations in this core component are not limited to the aforementioned. Further work is needed, to provide a comprehensive set of ethical principles that are commonly agreed upon at least by the scientific literature, key organizations (i.e., EU) and initiatives. Moreover, we should be aware that as technology evolves, it is quite possible to deal with upcoming and emerging ethical concerns that are yet to come. This is why we need a modular framework such as MOM, to operate and facilitate human oversight and assessment of algorithmic behavior in a dynamic manner.

Guidelines and legal framework

Depending on the context of use – for instance, not only the country where the MOM framework is being used but also, where the developed system will be operating – regulations might apply differently. Many such regulations and guidelines aim to protect individuals' freedom and core values, especially in the context of algorithmic governance. Some consider human values and ethical principles that are important to the scope of use. Last but not least, both national and international laws might apply depending on the occasion.

We mention once more the practices in the EU setting as a paradigm. The EC drafted the first regulatory framework for AI, the EU AI Act, where different rules apply for different risk levels of the AI application.[16] The initial plan was to reach an agreement on the final form of the law by the end of 2023. Despite the concerns about the complexity of the recent introduction of generative AI, the member states came into an agreement on December 2023[17] that will gradually be put into force during the next couple of years.[18] Thus, software developers and, more broadly, organizations that develop or use software that could pose risks will be enforced to comply with this law, as appropriate per the risk level of their systems. Particularly, the EC defines two classes of risks: *unacceptable risk* and *high-risk.*[16] AI systems categorized under the *unacceptable risk* are considered to be a threat to people and will be banned. On the other hand, AI systems that could negatively impact safety or fundamental rights will be considered *high risk.* Nevertheless, there is a fine line between these two classes, where a gray area exists, which encompasses systems that cannot be classified explicitly and transparently. Further work is required to provide methods for determining the risk type of those systems accordingly.

Despite the evolution speed of AI and technology in general, the legal framework is still ill-prepared with what currently exists and how is applied to the real world. The works of Floridi *et al.* (2021), Mittelstadt *et al.* (2016) and Kaur *et al.* (2021) mirror the gap of retaining the necessary mechanisms, instruments, regulation and other schemes (i.e., government financial incentivization) that at the same time will both support and enforce organizations comply with the relevant regulations for algorithmic systems.

[16] https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[17] https://www.reuters.com/technology/france-now-backing-eu-ai-rules-eu-source-says-ahead-bloc-endorsement-2024-02-02/

[18] https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)698792

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

To sum up, the respective ethical guidelines and legal framework are core components of the MOM framework. They should be integrated into the process, in the way that MOM itself and the respective target algorithmic system of focus, stay compliant with any national or international laws that apply.

Modularity

To ensure a framework that can be dynamically adapted to different contexts and applications of use, it needs to be modular at the core. Consequently, we refer to *modularity* of the framework having three main characteristics in mind: (a) the framework consists of distinct important modules that facilitate its utilization (i.e., the core components), (b) a framework that is expandable towards new required modules in the future and (c) the ability of the framework to be altered and adapted into different contexts and applications of use. This enables the framework to operate in different situations and types of applications.

Modularity is vital to the MOM framework, as it is involved in various key aspects. Below we expand on five main critical aspects:

- **Domain of application:** The framework has to consider different application contexts and uses. For example, it should be applicable in a diverse set of domains (e.g., not only in financial systems). The type of the target system under scrutiny is also essential and the framework should have the ability to facilitate human oversight for various types of applications ranging from simple algorithmic implementations to complex computer vision and natural language processing components. Finally, it should be able to be appended with new modules of upcoming emerging domains, leaving space for future improvement.
- **Human factors or entities:** Different oversight executions might require a different pool of stakeholders depending on the context and application of use. Therefore, the framework should provide the capability to assemble the crowdsourced simulated society accordingly. This is where microtask crowdsourcing would be also valuable for recruiting the necessary people in the loop. Looking further into the crowdsourced society, some domains might require several additional stakeholders as experts in the fields (e.g., doctors in the health domain, when exercising oversight over a medical system). These expert stakeholders will be teaming with the crowdsourced society, both as part of a SITL to scrutinize the target algorithmic behavior.
- **Technological strategies:** Various technological strategies for integration with the target systems to operate or assess them would be required for applying the MOM framework to exercise oversight. The framework itself should be flexible in terms of the way it can be integrated (e.g., using RESTFul services and library APIs). Also, there are various auditing strategies (Sandvig *et al.* 2014) that the developer who uses the MOM framework might want to apply.
- **Guidelines and laws:** As both the MOM framework and the system of focus would operate in distinct contexts that might be affected by different legal perspectives, the proposed framework should be able to facilitate this kind of compliance. National and international guidelines and laws that apply should be considered for inclusion in the framework.
- **Diversity:** The notion of diversity has a modular nature by definition. As described, modularity in diversity is a crucial factor for monitoring SITL through DITL, by recruiting a crowdsourced society of workers who bear diverse views

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

and values on the target matter during the oversight of algorithmic behavior. The framework has to facilitate a DITL approach by providing insights during the execution of the oversight process and the means for altering the crowdsourced society based on the aforementioned *diversity factors* chosen, to maximize the benefits of such process.

3.4. The five phases

Now, we describe the five main phases of the application of the MOM framework. We provide further details on the essential elements during each phase along with important and crucial points for its successful adoption. Figure 2 presents a summary of the main activities involved during each phase of the framework.

Phase 1: Preparation

This preparation phase is the most important step for executing a comprehensive human oversight process using the MOM framework. During this step, all the necessary preparation decisions and actions should be made to tweak the modules of the framework depending on the context and application of use. More specifically, this step can be distinguished into two main consequent sub-categories: (a) the *Definitions* and then (b) *Integration.* Following, we elaborate further on these two subcategories.

Definitions. The developers, along with any other crucial involved stakeholders [hereon: the overseers], have to come to a consensus on the definitions of four main aspects, depending on the context and application of use. First, they should conclude with the ethics and norms involved, the aims of the target system, and the regulations that apply. These would structure an *algorithmic social contract* where both the overseers and the SITL would follow to assess the target algorithmic behavior. Second, they must agree on choosing the CPs they will exploit based on their capabilities, and as a result, the pool of workers each platform provides. In addition, they have to decide the number of workers in the crowdsourced society that would be formed and the corresponding budget they will be compensated for their contributions, depending on the complexity and duration of the task. Third, specific *diversity factors* of focus should be determined prior to the execution along with a set of quantitative and qualitative metrics. These will aid in the monitoring of DITL throughout the oversight process and continuously improve aspects of the crowdsourced society to provide observations of higher quality and scrutiny. Fourth, designing the stimulus is vital (i.e., the crowdsourcing task). More specifically, they have to define clear task instructions, the task flow, and any assets that would be involved considering also the ones that have to be prepared prior to the oversight task (e.g., a predefined dataset of diverse people images). We suggest designing micro-tasks – rather than macro-tasks – because of their simple, quick and efficient nature that enables them to be repeatable and parallel when necessary.

Integration. After the definitions are formed, the developers should begin the necessary actions to integrate both the system of focus and the selected CPs in the MOM framework. Particularly, the system of focus would be integrated via APIs (e.g., RESTful APIs) or other interfaces needed. In addition, the developer should be aware of the clear objectives of the system and follow a behavioral risk management plan for the components that are susceptible to generating unexpected behavior that might cause harm to groups of people and individuals. Finally,

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

the developer has to integrate and configure the chosen CPs either through their provided interfaces or using their offered features to incorporate the flow of the task into the MOM framework for human oversight execution.

Phase 2: Recruitment

When the necessary preparation is done, the developer can begin proceeding in this recruitment step to start forming the simulated SITL via micro-task crowdsourcing by utilizing the chosen CPs. In this phase, the crowdsourcing task is executed and the recruitment of the crowdsourced society begins, always based on the characteristics derived from the chosen diversity factors. Also, the workers' compensation begins. Finally, it is important to point out that this is the beginning of a continuous monitoring process of SITL through DITL that will be active also during the next two phases of *inspection* and *review*. This would enable an agile approach to improving the oversight as facilitated by the MOM framework through the adjustment of SITL through DITL. When the crowdsourced observations are not sufficient or representative of the aims of the task, focus diversity factors can be altered accordingly to enrich or make more specific the pool of workers collected so far. This phase acts as the anchor for the following ones to step back and repeat the necessary actions to improve the quality of the oversight process.

Phase 3: Inspection

The next phase is the inspection phase, where live monitoring and oversight are active during the inspection of algorithmic behavior by SITL and the overseers (i.e., the developer and involved stakeholders). During this phase, a set of predefined quantitative and qualitative metrics (from the preparation phase) are generated based on the SITL observations. The SITL/DITL monitoring continues to play an active role in this phase. The overseers can modify SITL through DITL by altering the diversity factors of focus at any given time to re-recruit and re-execute the crowdsourced oversight task by going back to the second phase when observations are not sufficient.

Phase 4: Review

When the execution of the crowdsourcing part is done, the overseers should review the observations gathered by SITL. More specifically, they should conduct post-quantitative and postqualitative assessments to examine the target algorithmic behavior. Again, when observations are insufficient or nonrepresentative of the aims of the system or the oversight task, the overseers can roll back to the second phase where they could adjust SITL through DITL and re-execute the crowdsourcing part to improve the results. This is the last phase where the monitoring of SITL/DITL is still active and ends by proceeding to the next phase.

Phase 5: Decision

The final phase is focused on decision-making after the necessary inspection and review of the target system is done. To be exact, the overseers come to a decision on the assessment of the target algorithmic behavior and the implications or issues perpetuated. At the same time, a set of reports is generated based on the reporting metrics and other forms of assessments that have been initially agreed upon (in the preparation phase). Optionally, this process can aid in initiating a developer

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

intervention to fix or mitigate the issue if possible. Otherwise, the user of the system can contest the decision of the generated output and override it to mitigate any harmful effects on groups of people or individuals.

4. Discussion

In this work, we conceptualize and propose a framework for exercising human oversight over algorithmic behavior, which we call MOM. The framework enables primarily developers or software companies to exercise oversight over their algorithmic systems, and secondarily, other organizations using them to exercise oversight to ensure their algorithmically generated outcomes do not have a negative impact on society.

We presented the basic components of this framework, which are involved during a human oversight process. As a result, we further elaborated on the notions of SITL, microtask crowdsourcing, Diversity, ethics, guidelines and legal framework, and modularity as defined in the context of MOM. In addition, we provided further details on the five phases of the MOM framework by explaining its application and modification depending on the context and application of use.

As we explained, although SITL is a noble and ideal goal, its actual application cannot be realized. Consequently, we propose managing DITL based on a set of diversity factors to provide an implementation of a simulated SITL while managing it through DITL correspondingly. In addition, we argue that microtask crowdsourcing is the way to go for bringing a more controlled and efficient notion of SITL while also considering various diverse aspects of the crowdsourced society.

It is crucial to uncover socially harmful algorithmic implications that might have a negative impact on people's lives. As the EC proposed in its Ethic Guidelines on Trustworthy AI,[6] we need to consider human oversight for examining the behavior of such systems and override their decisions or influence when needed. Individual developers and software companies alike bear a substantial amount of responsibility for keeping the influence of those systems fair to society at large. The MOM framework can provide the groundwork for the development of future domain-specific oversight solutions that aid in exercising human oversight in a more specific, systematic and controlled way.

5. Challenges and future work

As previously described, the MOM framework is not a "golden solution" for any given situation, but rather, it is an umbrella modular methodology that can be adjusted to different future domain-specific applications. It can be the first step towards exercising human oversight over target algorithmic behavior according to a set of considerations such as relevant ethical principles, regulations and social norms. Future domain-specific solutions could contribute to examining systems of various types, operated in distinct contexts.

When recruiting a simulated SITL, it is important to acknowledge two major limitations. To begin with, it is impossible to create an exhaustive set of diversity factors that consider every possible characteristic of individual workers in the crowdsourced society. In fact, further work is needed to determine at least a

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

basic set of these factors that commonly influence the way workers observe and assess algorithmic behavior. Future investigations might also lead to domain-specific diversity factors to be considered during an oversight process. While we expect that key factors will vary by context, we also believe that a core set of common factors – and appropriate metrics to measure and manage them – can be derived.

Furthermore, another limitation of SITL recruitment – in the MOM framework – is the crowd pool and feature limitations of the chosen CPs (Ross *et al.* 2009; Vakharia & Lease 2015; Garcia-Molina *et al.* 2016). For instance, some platforms have a limited crowd pool in specific categories (e.g., per region) compared to others (Chandler *et al.* 2014; Miller *et al.* 2017). Another example is that some CPs do not have adequate worker pools that reflect equal gender or age representation in specific characteristics. In addition, some of these CPs might not offer the features needed for an oversight approach (e.g., providing workers' socioeconomic status), which raises the bar in the complexity of recruiting a diverse society.

Another crucial point concerns the scalability and effectiveness of the MOM framework in ensuring ethical alignment and mitigating biases in algorithmic systems. In the current work, we have not investigated this kind of scalability and/or effectiveness, but rather, we present a conceptual framework. Future research on investigating these aspects by considering different viewpoints – using both quantitative and qualitative methods – would be useful to optimize and/or maximize the benefits of such a framework.

Also, there is an important concern when it comes to the adoption of the framework by companies. There are numerous challenges in the corporate sector, such as an organization's potential (a) apprehensiveness about disclosing any data that might often be perceived as part of its intellectual property (IP), (b) tendency to prioritize compliance only to a certain extent (usually to the extent required while maximizing the benefits of their development efforts or business goals) and (c) willingness to adopt a higher level of rigor in their algorithmic systems. Consequently, finding ways to encourage organizations – by also raising their awareness and providing motivation – might be highly correlated with the *feasibility, adaptability* and *trust* of the framework in real-world contexts. We need compelling strategies to approach organizations and communicate the important aspects, benefits and motivations behind such oversight approaches. So, this is again an area worth investigating in future work.

Last but not least, upcoming future technologies might need a different kind of approach to exercise oversight over their behavior. For example, the recent introduction of large language models (LLMs) and associated applications (e.g., ChatGPT, which can in turn be built into third-party applications) makes this process even more dynamic and unpredictable than other AI-based implementations. This makes us consider how human oversight frameworks might have to be applied in the future. Such technological advancements, like LLMs, only highlight how crucial is to have such frameworks in place, to provide guidelines and a standardized approach, given that oversight would have to be deployed at a fast pace and into a huge range of contexts. Consequently, the situation will become even more dynamic. Oversight frameworks such as MOM, should without a doubt, share a modular, dynamic, and expandable nature to stay relevant to upcoming technological developments.

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

Acknowledgments

This project has received funding from the Cyprus Research and Innovation Foundation under grant EXCELLENCE/0421/0360 (KeepA(n)I), the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 739578 (RISE), and the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation and Digital Policy.

References

Aggarwal, A., **Lohia, P.**, **Nagar, S.**, **Dey, K.** & **Saha, D.** 2019 Black box fairness testing of machine learning models. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 625–635. ESEC/FSE 2019, Association for Computing Machinery. https://doi.org/10.1145/3338906.3338937

Al Alamin, M. A. & **Uddin, G.** 2021 Quality assurance challenges for machine learning software applications during software development life cycle phases. In *2021 IEEE International Conference on Autonomous Systems (ICAS)*, pp. 1–5.

Alla, S. & **Adari, S. K.** 2021 *What Is MLOps?*, pp. 79–124. Apress. https://doi.org/10.1007/978-1-4842-6549-93

Asatiani, A., **Malo, P.**, **Nagbol, P.**, **Penttinen, E.**, **Rinta-Kahila, T.** & **Salovaara, A.** 2020 Challenges of explaining the behavior of black-box AI systems. *MIS Quarterly Executive* **19**(4), 259–278.

Awad, E., **Dsouza, S.**, **Bonnefon, J.-F.**, **Shariff, A.** & **Rahwan, I.** 2020 Crowdsourcing moral machines. *Communications of the ACM* **63**(3), 48–55. https://doi.org/10.1145/3339904

Bansal, G., **Nushi, B.**, **Kamar, E.**, **Lasecki, W. S.**, **Weld, D. S.** & **Horvitz, E.** 2019 Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. **7**, pp. 2–11.

Bellamy, R. K. E., **Dey, K.**, **Hind, M.**, **Hoffman, S. C.**, **Houde, S.**, **Kannan, K.**, **Lohia, P.**, **Martino, J.**, **Mehta, S.**, **Mojsilović, A.**, **Nagar, S.**, **Ramamurthy, K. N.**, **Richards, J.**, **Saha, D.**, **Sattigeri, P.**, **Singh, M.**, **Varshney, K. R.** & **Zhang, Y.** 2019 Ai fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* **63**(4/5), 4:1–4:15.

Brabham, D. C. 2013 *Crowdsourcing*. Cambridge, MA: The MIT Press.

Buolamwini, J. & **Gebru, T.** 2018 Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler & C. Wilson, eds, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, Vol. **81** of *Proceedings of Machine Learning Research*, PMLR, pp. 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

Campos, A. S. 2019 'The idea of the social contract in the history of 'agreementism'', *The European Legacy* **24**(6), 579–596. https://doi.org/10.1080/10848770.2019.1608049

Chandler, J., **Mueller, P.** & **Paolacci, G.** 2014 Nonnaïveté among amazon mechanical turk workers: consequences and solutions for behavioral researchers. *Behavior Research Methods* **46**, 112–130.

Chen, D., **Li, X.** & **Lai, F.** 2017 Gender discrimination in online peer-to-peer credit lending: evidence from a lending platform in China. *Electronic Commerce Research* **17**, 553–583.

Chen, T. Y. 2015 Metamorphic testing: A simple method for alleviating the test oracle problem. In *2015 IEEE/ACM 10th International Workshop on Automation of Software Test*, pp. 53–54.

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

Cheverda, A., **Jabborov, A.**, **Kruglov, A.** & **Succi, G.** 2022 State-of-the-art review of taxonomies for quality assessment of intelligent software systems. In *2022 3rd International Informatics and Software Engineering Conference (IISEC)*, pp. 1–6.

Côté, PO., **Nikanjam, A.**, **Bouchoucha, R.** 2024 Quality issues in machine learning software systems. *Empir Software Eng 29*, **149** https://doi.org/10.1007/s10664-024-10536-7

Danks, D. & **London, A. J.** 2017 Algorithmic bias in autonomous systems. *IJCAI* **17**, 4691–4697.

Davis, M. D. & **Weyuker, E. J.** 1981 Pseudo-oracles for non-testable programs. In *Proceedings of the ACM '81 Conference*, pp. 254–257. Association for Computing Machinery. https://doi.org/10.1145/800175.809889

Deng, X. N. & **Joshi, K. D.** 2016 Why individuals participate in micro-task crowdsourcing work environment: revealing crowdworkers' perceptions. *Journal of the Association for Information Systems* **17**(10), 3.

Drosou, M., **Jagadish, H. V.**, **Pitoura, E.** & **Stoyanovich, J.** 2017 Diversity in big data: a review. *Big Data* **5**(2), 73–84.

Ebert, C., **Gallardo, G.**, **Hernantes, J.** & **Serrano, N.** 2016 Devops. *IEEE Software* **33**(3), 94–100.

Estellés-Arolas, E. & **de Guevara, F. G.-L.** 2012 Towards an integrated crowdsourcing definition. *Journal of Information Science* **38**(2), 189–200. https://doi.org/10.1177/0165551512437638

Farroha, B. & **Farroha, D.** 2014 A framework for managing mission needs, compliance, and trust in the devops environment. In *2014 IEEE Military Communications Conference*, pp. 288–293.

Felderer, M. & **Ramler, R.** 2021 Quality assurance for ai-based systems: Overview and challenges (introduction to interactive session). In D. Winkler, S. Biffl, D. Mendez, M. Wimmer & J. Bergsmann, eds, *Software Quality: Future Perspectives on Software Engineering Quality*, pp. 33–42. Springer International Publishing.

Felderer, M., **Russo, B.** & **Auer, F.** 2019 *On Testing Data-Intensive Software Systems*, pp. 129–148. Springer International Publishing. https://doi.org/10.1007/978-3-030-25312-76

Fitzgerald, B. & **Stol, K.-J.** 2017 Continuous software engineering: a roadmap and agenda. *Journal of Systems and Software* **123**, 176–189. https://www.sciencedirect.com/science/article/pii/S0164121215001430

Floridi, L., **Cowls, J.**, **Beltrametti, M.**, **Chatila, R.**, **Chazerand, P.**, **Dignum, V.**, **Luetge, C.**, **Madelin, R.**, **Pagallo, U.**, **Rossi, F.**, **Schafer, B.**, **Valcke, P.** & **Vayena, E.** 2021 *An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*, pp. 19–39. Springer International Publishing. https://doi.org/10.1007/978-3-030-81907-13

Gadiraju, U., **Demartini, G.**, **Kawase, R.** & **Dietze, S.** 2015 Human beyond the machine: challenges and opportunities of microtask crowdsourcing. *IEEE Intelligent Systems* **30** (4), 81–85.

Garcia-Molina, H., **Joglekar, M.**, **Marcus, A.**, **Parameswaran, A.** & **Verroios, V.** 2016 Challenges in data crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering* **28**(4), 901–911.

Gift, N. & **Deza, A.** 2021 *Practical MLOps*. O'Reilly Media, Inc.

Giunchiglia, F., **Bison, I.**, **Busso, M.**, **Chenu-Abente, R.**, **Rodas, M.**, **Zeni, M.**, **Gunel, C.**, **Veltri, G.**, **De Götzen, A.**, **Kun, P.**, **and Ganbold, A.** 2021 A worldwide diversity pilot on daily routines and social practices (2020). University of Trento, Technical Report. No.# DISI-2001-DS-01.

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

Giunchiglia, F. & **Fumagalli, M.** 2017 Teleologies: Objects, actions and functions. In H. C. Mayr, G. Guizzardi, H. Ma & O. Pastor, eds, *Conceptual Modeling*, pp. 520–534. Springer International Publishing.

Gotlieb, A. 2015 *Chapter two - constraint-based testing: An emerging trend in software testing. Vol. 99 of Advances in Computers, pp.* 67–101. Elsevier. *https://www.sciencedirect.com/science/article/pii/S0065245815000340*

Grgić-Hlača, N., **Lima, G.**, **Weller, A.** & **Redmiles, E. M.** 2022 Dimensions of diversity in human perceptions of algorithmic fairness. In *Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO '22.* Association for Computing Machinery. https://doi.org/10.1145/3551624.3555306

Haas, D., **Ansel, J.**, **Gu, L.** & **Marcus, A.** 2015 Argonaut: macrotask crowdsourcing for complex data processing. *Proceedings of the VLDB Endownment* **8**(12), 1642–1653. https://doi.org/10.14778/2824032.2824062

Hobbes, T. 1651 Leviathan, London: Andrew Crooke, 1651. *Project Gutenberg.*

Honeycutt, D., **Nourani, M.** & **Ragan, E.** 2020 Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* **8**(1), 63–72. https://ojs.aaai.org/index.php/HCOMP/article/view/7464

Hummel, O., **Eichelberger, H.**, **Giloj, A.**, **Werle, D.** & **Schmid, K.** 2018 A collection of software engineering challenges for big data system development. In *2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA).* pp. 362–369.

Hussain, S. S., **Hashmani, M.**, **Uddin, V.**, **Ansari, T.** & **Jameel, M.** 2021 A novel approach to detect concept drift using machine learning. In *2021 International Conference on Computer Information Sciences (ICCOINS).* pp. 136–141.

Ibáñez, L.-D., **Reeves, N.** & **Simperl, E.** 2020 *Crowdsourcing and Human-in-the-Loop for IoT*, John Wiley Sons, Ltd, chapter 8, pp. 91–105. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119545293.ch8

Imana, B., **Korolova, A.** & **Heidemann, J.** 2021 Auditing for discrimination in algorithms delivering job ads. In *Proceedings of the Web Conference 2021, WWW '21*, pp. 3767–3778. Association for Computing Machinery. https://doi.org/10.1145/3442381.3450077

Islam, Z. U. 2021 Software engineering methods for responsible artificial intelligence. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1814–1815.

John, M. M., **Olsson, H. H.** & **Bosch, J.** 2021 Towards mlops: A framework and maturity model. In *2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pp. 1–8.

Jos, P. H. 2006 Social contract theory: implications for professional ethics. *The American Review of Public Administration* **36**(2), 139–155. https://doi.org/10.1177/0275074005282860

Karachiwalla, R. & **Pinkow, F.** 2021 Understanding crowdsourcing projects: a review on the key design elements of a crowdsourcing initiative. *Creativity and Innovation Management* **30**(3), 563–584. https://onlinelibrary.wiley.com/doi/abs/10.1111/caim.12454

Karamitsos, I., **Albarhami, S.** & **Apostolopoulos, C.** 2020 Applying devops practices of continuous automation for machine learning. *Information* **11**(7). https://www.mdpi.com/2078-2489/11/7/363

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

Kaur, D., **Uslu, S.** & **Durresi, A.** 2021 Requirements for trustworthy artificial intelligence – a review. In L. Barolli, K. F. Li, T. Enokido & M. Takizawa, eds, *Advances in Networked-Based Information Systems*, pp. 105–115. Springer International Publishing.

Köchling, A. & **Wehner, M. C.** 2020 Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research* **13**(3), 795–848.

Kyriakou, K., **Barlas, P.**, **Kleanthous, S.** & **Otterbacher, J.** 2019 Fairness in proprietary image tagging algorithms: a cross-platform audit on people images. *Proceedings of the International AAAI Conference on Web and Social Media* **13**(01), 313–322. https://ojs.aaai.org/index.php/ICWSM/article/view/3232

Kyriakou, K., **Kleanthous, S.**, **Otterbacher, J.** & **Papadopoulos, G. A.** 2020 Emotion-based stereotypes in image analysis services. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pp. 252–259. UMAP '20 Adjunct, Association for Computing Machinery. https://doi.org/10.1145/3386392.3399567

Li, E. 2023 'The Qianke System in China: Disorganisation, discrimination and dispersion', *Criminology &amp; Criminal Justice*, **23**(4), pp. 568–587. https://doi.org/10.1177/17488958231161436

Liu, T. 2020 Human-in-the-loop learning from crowdsourcing and social media, PhD thesis. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated 2023-06-21. https://www.proquest.com/dissertations-theses/human-loop-learningcrowdsourcing-social-media/docview/2438898542/se-2

Locke, J. 2013 Two treatises of government, 1689. *The anthropology of citizenship: A reader*, pp. 43–46.

Lu, Q., **Zhu, L.**, **Xu, X.** & **Whittle, J.** 2023 Responsible-ai-by-design: a pattern collection for designing responsible artificial intelligence systems. *IEEE Software* **40**(3), 63–71.

Lu, Q., **Zhu, L.**, **Xu, X.**, **Whittle, J.**, **Douglas, D.** & **Sanderson, C.** 2022 Software engineering for responsible ai: An empirical study and operationalised patterns. In *Proceedings of the 44th International Conference on Software Engineering: Software Engineering in Practice, ICSE-SEIP '22*, pp. 241–242. Association for Computing Machinery. https://doi.org/10.1145/3510457.3513063

Lu, Q., **Zhu, L.**, **Xu, X.**, **Whittle, J.** & **Xing, Z.** 2022 Towards a roadmap on software engineering for responsible AI. In *Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI, CAIN '22*, pp. 101–112. Association for Computing Machinery. https://doi.org/10.1145/3522664.3528607

Marijan, D., **Gotlieb, A.** & **Kumar Ahuja, M.** 2019 Challenges of testing machine learning based systems. In *2019 IEEE International Conference On Artificial Intelligence Testing (AITest)*, pp. 101–102.

Martin-Lopez, A., **Segura, S.** & **Ruiz-Cortés, A.** 2020 Restest: black-box constraint-based testing of restful web apis. In E. Kafeza, B. Benatallah, F. Martinelli, H. Hacid, A. Bouguettaya & H. Motahari, eds, '*Service-Oriented Computing*, pp. 459–475. Springer International Publishing.

Matsui, B. M. A. & **Goya, D. H.** 2022a Mlops: A guide to its adoption in the context of responsible ai. In *2022 IEEE/ACM 1st International Workshop on Software Engineering for Responsible Artificial Intelligence (SE4RAI)*, pp. 45–49.

Matsui, B. M. A. & **Goya, D. H.** 2022b Mlops: Five steps to guide its effective implementation. In *2022 IEEE/ACM 1st International Conference on AI Engineering – Software Engineering for AI (CAIN)*, pp. 33–34.

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

Miller, J. D., **Crowe, M.**, **Weiss, B.**, **Maples-Keller, J. L.** & **Lynam, D. R.** 2017 Using online, crowdsourcing platforms for data collection in personality disorder research: the example of Amazon's mechanical turk. *Personality Disorders: Theory, Research, and Treatment* **8**(1), 26.

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

Nakao, Y. 2022 Toward human-in-the-loop AI fairness with crowdsourcing: effects of crowdworkers' characteristics and fairness metrics on ai fairness perception.

Nushi, B., Kamar, E. & Horvitz, E. 2018 Towards accountable AI: hybrid human-machine analyses for characterizing system failure. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* **6**(1), 126–135. https://ojs.aaai.org/index.php/HCOMP/article/view/13337

Pedreschi, D., **Giannotti, F.**, **Guidotti, R.**, **Monreale, A.**, **Ruggieri, S.** & **Turini, F.** 2019 Meaningful explanations of black box AI decision systems. *Proceedings of the AAAI Conference on Artificial Intelligence* **33**(01), 9780–9784. https://ojs.aaai.org/index.php/AAAI/article/view/5050

Peters, D., **Vold, K.**, **Robinson, D.** & **Calvo, R. A.** 2020 Responsible AI—two frameworks for ethical design practice. *IEEE Transactions on Technology and Society* **1**(1), 34–47.

Rahwan, I. 2018 Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology* **20**(1), 5–14. https://doi.org/10.1007/s10676-017-9430-8

Raj, E. 2021 *Engineering MLOps: Rapidly Build, Test, and Manage Production-Ready Machine Learning Life Cycles at Scale*, Packt Publishing Ltd.

Rosen, C. 2020 *Software Systems Quality Assurance and Evaluation*, pp. 101–123. Springer International Publishing. https://doi.org/10.1007/978-3-030-39730-26

Ross, J., **Zaldivar, A.**, **Irani, L.** & **Tomlinson, B.** 2009 Who are the turkers? Worker demographics in amazon mechanical turk. Department of Informatics, University of California, Irvine, USA, Tech. Rep. 49.

Rousseau, J.-J. (2003) *A discourse on equality*. London: Penguin.

Rousseau, J.-J. 1964 'The social contract (1762). *Londres.*

Ruf, P., **Madan, M.**, **Reich, C.** & **Ould-Abdeslam, D.** 2021 Demystifying mlops and presenting a recipe for the selection of open-source tools. *Applied Sciences* **11**(19). https://www.mdpi.com/2076-3417/11/19/8861

Saleiro, P., **Kuester, B.**, **Hinkson, L.**, **London, J.**, **Stevens, A.**, **Anisfeld, A.**, **Rodolfa, K.T.**, **and Ghani, R.** 2018 Aequitas: A bias and fairness audit toolkit. arXiv preprint arXiv:1811.05577.

Sambasivan, N. & **Holbrook, J.** 2018 Toward responsible ai for the next billion users. *Interactions* **26**(1), 68–71. https://doi.org/10.1145/3298735

Sandvig, C., **Hamilton, K.**, **Karahalios, K.** & **Langbort, C.** 2014 Auditing algorithms: research methods for detecting discrimination on internet platforms. . *Data and Discrimination: Converting Critical Concerns into Productive Inquiry* **22**(2014), 4349–4357.

Schelenz, L., **Bison, I.**, **Busso, M.**, **de G¨otzen, A.**, **Gatica-Perez, D.**, **Giunchiglia, F.**, **Meegahapola, L.** & **Ruiz-Correa, S.** 2021 The theory, practice, and ethical challenges of designing a diversity-aware platform for social relations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES '21*. Association for Computing Machinery, pp. 905–915. https://doi.org/10.1145/3461702.3462595

Schieferdecker, I. 2020 Responsible software engineering. *The Future of Software Quality Assurance*, pp. 137–146.

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

Schlossnagle, T. 2018 Monitoring in a devops world. *Communications of the ACM* **61**(3), 58–61.

Seabright, P., **Stieglitz, J.** & **Van der Straeten, K.** 2021 Evaluating social contract theory in the light of evolutionary social science. *Evolutionary Human Sciences* **3**, e20.

Segura, S., **Fraser, G.**, **Sanchez, A. B.** & **Ruiz-Cortés, A.** 2016 A survey on metamorphic testing. *IEEE Transactions on Software Engineering* **42**(9), 805–824.

Sharma, S. 2017 *The DevOps Adoption Playbook*. Wiley.

Shen, H. & **Huang, T.-H.** 2020 How useful are the machine-generated interpretations to general users? A human evaluation on guessing the incorrectly predicted labels. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. **8**, pp. 168–172.

Shneiderman, B. 2021. Responsible AI: bridging from ethics to practice. *Communications of the ACM* **64**(8), 32–35.

Soklaski, R., **Goodwin, J.**, **Brown, O.**, **Yee, M.** & **Matterer, J.** 2022. Tools and practices for responsible ai engineering. *arXiv preprint arXiv:2201.05647*.

Sun, L., **Wei, M.**, **Sun, Y.**, **Suh, Y. J.**, **Shen, L.** & **Yang, S.** 2023 Smiling women pitching down: auditing representational and presentational gender biases in image generative ai. *arXiv preprint arXiv:2305.10566*.

Suresh, H., **Gomez, S. R.**, **Nam, K. K.** & **Satyanarayan, A.** 2021 Beyond expertise and roles: a framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21*. Association for Computing Machinery. https://doi.org/10.1145/3411764.3445088

Symeonidis, G., **Nerantzis, E.**, **Kazakis, A.** & **Papakostas, G. A.** 2022 Mlops – definitions, tools and challenges. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 453–460.

Tamburri, D. A. 2020 Sustainable mlops: trends and challenges. In *2020 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pp. 17–23.

Testi, M., **Ballabio, M.**, **Frontoni, E.**, **Iannello, G.**, **Moccia, S.**, **Soda, P.** & **Vessio, G.** 2022 MLOPs: a taxonomy and a methodology. *IEEE Access* **10**, 63606–63618.

Treveil, M., **Omont, N.**, **Stenac, C.**, **Lefevre, K.**, **Phan, D.**, **Zentici, J.**, **Lavoillotte, A.**, **Miyazaki, M.** & **Heidmann, L.** 2020 *Introducing MLOps*. O'Reilly Media.

Vakharia, D. & **Lease, M.** 2015 Beyond mechanical turk: An analysis of paid crowd work platforms. In *Proceedings of the iConference*, pp. 1–17.

Van Berkel, N., Goncalves, J., Hettiachchi, D., Wijenayake, S., Kelly, R.M., and Kostakos, V. 2019 Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), pp.1–21.

van den Heuvel, W.-J. & **Tamburri, D. A.** 2020 Model-driven ml-ops for intelligent enterprise applications: Vision, approaches and challenges. In B. Shishkov, ed. *Business Modeling and Software Design*, pp. 169–181. Springer International Publishing.

Viglianisi, E., **Dallago, M.** & **Ceccato, M.** 2020 Resttestgen: Automated black-box testing of restful apis. In *2020 IEEE 13th International Conference on Software Testing, Validation and Verification (ICST)*, pp. 142–152.

von Eschenbach, W. J. 2021 Transparency and the black box problem: why we do not trust AI. . *Philosophy & Technology* **34**(4), 1607–1622.

Weizenbaum, J. 1972 On the impact of the computer on society. *Science* **176**(4035), 609–614. https://www.science.org/doi/abs/10.1126/science.176.4035.609

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press

Weyuker, E. J. 1982 On testing non-testable programs. *The Computer Journal* **25**(4), 465–470. https://doi.org/10.1093/comjnl/25.4.465

Zhen, Y., **Khan, A.**, **Nazir, S.**, **Huiqi, Z.**, **Alharbi, A.** & **Khan, S.** 2021 Crowdsourcing usage, task assignment methods, and crowdsourcing platforms: a systematic literature review. *Journal of Software: Evolution and Process* **33**(8), e2368. https://onlinelibrary.wiley.com/doi/abs/10.1002/smr.2368

Zhu, L., **Bass, L.** & **Champlin-Scharff, G.** 2016 Devops and its practices. *IEEE Software* **33**(3), 32–34.

Zulfiqar, M., **Malik, M. N.** & **Khan, H. H.** 2022 Microtasking activities in crowdsourced software development: a systematic literature review. *IEEE Access* **10**, 24721–24737.

https://doi.org/10.1017/dsj.2024.23 Published online by Cambridge University Press