AI & SOCIETY (2024) 39:1975–1994
https://doi.org/10.1007/s00146-023-01653-w

OPEN FORUM

No such thing as one-size-fits-all in AI ethics frameworks: a comparative case study

Vivian Qiang[1] · **Jimin Rhim**[2] · **AJung Moon**[2]

Received: 15 June 2022 / Accepted: 29 March 2023 / Published online: 6 May 2023

Abstract
Despite the bombardment of AI ethics frameworks (AIEFs) published in the last decade, it is unclear which of the many have been adopted in the industry. What is more, the sheer volume of AIEFs without a clear demonstration of their effectiveness makes it difficult for businesses to select which framework they should adopt. As a first step toward addressing this problem, we employed four different existing frameworks to assess AI ethics concerns of a real-world AI system. We compared the experience of applying the AIEFs from the perspective of (a) a third-party auditor conducting an AI ethics risk assessment for the company, and (b) the company receiving the audit outcomes. Our results suggest that the feel-good factor of doing an assessment is common across the AIEFs that can take anywhere between 1.5 and 20 h to complete. However, each framework provides different benefits (e.g., issue discovery vs. issue monitoring) and is likely best used in conjunction with one another at different stages of an AI development process. As such, we call on the AI ethics community to better specify the suitability and expected benefits of existing frameworks to enable better adoption of AI ethics practice in the industry.

Keywords AI ethics · AI ethics framework · AI toolkits · Audit · Healthcare AI · Startup

1 Introduction

In the last decade, there has been a flurry of activities calling for more ethical practices in the AI industry. Early attempts primarily focused on establishing a set of high-level principles and codes of ethics to guide responsible design and deployment of AI systems (Amershi et al. 2019; Jobin et al. 2019); for example, Microsoft AI Principles state that "AI systems should treat all people fairly" (Microsoft 2022). Soon, over 200 ethics frameworks—in the form of harm/risk/impact assessments, toolkits, checklists, and government processes—have been published to translate these abstract, high-level principles into practice (Ayling and Chapman 2021). However, despite the availability and large variety of AI ethics frameworks (AIEFs) today, it is unclear what impact they have on an AI business, product, or service. In other words, we do not know whether the adoption of existing AIEFs delivers any specific and observable benefit to the AI industry today. We posit that long-term, industry-wide adoption of the frameworks will not come to fruition if the AI ethics community fails to demonstrate a concrete benefit in adopting AI ethics practices, and articulate how to unlock such benefits.

As a means to fill this broader knowledge gap, this paper presents our efforts to address the following research question: how effective are the various types of AIEFs in helping businesses assess AI ethics risks? We present a real-world comparative case study that involves a startup—herein referred to as the "client." The startup's main product is an AI-powered recommender system that clinicians can use to find promising treatment options for patients with clinical depression. We employed a qualitative case-study methodology (Baxter and Jack 2008), wherein AIEFs acted as an "intervention" practiced in the context of a healthcare AI startup in Canada.

With a brief overview of existing work on AIEFs (Sect. 2), we present a landscape analysis. This analysis informed our selection of the widely different styles of frameworks applied to the case study (Sect. 3). The researchers took on the role of a third-party auditor assessing the

✉ AJung Moon
ajung.moon@mcgill.ca

[1] Department of Epidemiology, Biostatistics, and Occupational Health, School of Population and Global Health, McGill University, Montreal, QC, Canada

[2] Department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada

ethics risk of the pre-deployment AI system, and applied all four frameworks to the case study (Sect. 4). We compared the frameworks from two different perspectives: (a) the experience of a third-party auditor performing AI ethics risk assessments for the client company, and (b) the main recipient of the audit (i.e., the client).

Our results suggest that different types of AIEFs are better suited for different stages of an AI product design process, although this is rarely specified in the frameworks. We also find that domain-specific frameworks (e.g., AI in healthcare) are needed to help avoid or resolve contradictions between the output of an AI ethics assessment and existing regulatory requirements. We discuss the implications and conclusions from our investigation in Sects. 5 and 6.

The findings from this study contribute to the urgent task of investigating the efficacy of AIEFs in real-world applications. Further, we conjecture the need to treat AIEFs as an intervention mechanism to instill *a culture of responsible design and deployment* within the industry, rather than as a one-time measure to produce more thoughtfully designed end products. Finally, our comparative case-study approach led us to identify how an AI ethics checklist that takes 1.5 h to complete should serve a different function from process-oriented guides that take much more time and necessitate multi-stakeholder engagement in the process. We call upon the AI ethics community to help develop scientific, iterative means with which improvements to existing frameworks can be made, and to better specify the suitability and expected benefits of existing frameworks for better adoption of AI ethics practices in the industry.

2 Background

In this section, we provide a brief overview of the recent development in AIEFs (Sect. 2.1) and known hindrances to industry adoption of the frameworks (Sect. 2.2). Given the diversity of frameworks, we distinguish the terms *frameworks* and *toolkits*. In this paper, we use the term AI ethics *framework* (AIEF) to denote a body of work that generally aim to promote AI ethics, while AI ethics *toolkit*—a subcategory of AIEFs—refers specifically to tangible, step-by-step processes or a set of technical tools that support ethics risk discovery and management. AI ethics toolkits are further distinguished from tools, which are technical libraries (e.g., local interpretable model-agnostic explanations package (Ribeiro 2022) as an AI explainability tool (Ribeiro et al. 2016)) and a singular process (e.g., Datasheet for Datasets, a documentation tool). Given the fact that our work involves a real-world product in development (a healthcare AI system developed in Canada), we briefly contextualize the state of regulation and AIEFs relevant to the case study (Sect. 2.3).

2.1 AI ethics frameworks

A recent synthesis of six prominent ethical guidelines by Floridi et al. found that there is a significant degree of overlap between principles, which leads to the following five shared principles: beneficence, non-maleficence, autonomy, justice, and explicability (Floridi et al. 2018). However, these AI ethics principles have often been criticized for being too broad and high level to be practical, and for employing ambiguous terms that oversimplify conceptual nuances and potential differences in interpretation (Whittlestone et al. 2019; Jobin et al. 2019; Hagendorff 2020). The need for clarity surrounding ethical principles, as well as practical mechanisms for prioritizing these normative values remains.

In recognizing this gap between theoretical principles and practical implementation, several AIEFs have been published. AIEFs generally focus on shifting the "what" of AI ethics principles toward "how" one can apply these ethical principles. Often aligned with the most common AI ethics principles (Jobin et al. 2019), AIEFs offer instructions on how to identify and mitigate the ethical risks posed by AI systems across key normative values such as fairness, transparency, and accountability (IBM 2020). A corporation may employ AIEFs as an internal auditing tool that provides a set criterion for complying with globally recognized AI ethics principles. For example, International Business Machines (IBM) Corporation defined five foundational properties for AI ethics, which serve as the company's guiding values as they develop and adopt AI applications (IBM 2022). Other prominent organizations, such as BSA The Software Alliance, have published their own frameworks for assessing the impacts and risks of AI bias throughout a system's life cycle (BSA 2021).

One of the most comprehensive reviews of existing AIEFs reviewed 106 responsible AI tools and frameworks (Morley et al. 2020). They found a diversity of different frameworks from high-level conceptual guidelines to technical tools for identifying data bias. The study also revealed a distinct lack of applicability; although AIEFs offer potential mechanisms for stakeholders to design and deploy ethical systems, a vast majority of these processes are not actionable, as they offer minimal guidelines for how to implement them in practical settings (Vakkuri and Kemell 2019). Unlike hard governance mechanisms such as the rule of law, ethics does not have enforcement mechanisms beyond voluntary and non-binding cooperation between stakeholders in AI (Hagendorff 2020). As such, in an industry wherein corporate stakeholders prioritize profitability, one must recognize the pivotal challenges inherent to the subjective nature of AIEFs; companies may choose to employ frameworks to bolster public

image over maintaining ethical integrity—consequently impressing corporate biases within AI ethics evaluations.

2.2 Hindrance to industry adoption

While a common intention behind developing AIEFs has been to promote ethical development/deployment of AI systems, scholars have identified a number of practical obstacles to bringing the intention to reality. First, the ethics implications of an AI system and the appropriate mitigation strategies can differ significantly from one use case to another. The AI explainability community has expressed this more explicitly and articulated that not all AI explainability technique solves all AI explainability needs and problems (Arya et al. 2019). Likewise, it has been suggested that AIEFs should be tailored to specific industries, such that ethical risk assessment and mitigation is built into existing operations and standards (Blackman 2020). Second, the recent inundation of guidelines has also made the application-specific framework selection process difficult. Some state that stakeholders are already beginning to feel overwhelmed by an overabundance of AIEFs and may experience difficulties with selecting, comparing, and assessing the utility of the different tools (Schiff et al. 2021). The sheer abundance of AIEFs then may lead them to be no longer considered an effective means of instilling ethics considerations in AI. Moreover, an AI system can undergo multiple phases of the design and deployment process that require different levels of ethics considerations (Peters et al. 2020). However, readily available AIEFs or ethics toolkits tend to not distinguish the characteristics of different steps in a design process. Consequently, without selecting the proper toolkit or frameworks suitable for the specific purposes or stages of AI deployment, validating the effectiveness of the ethical dimension becomes difficult. Therefore, we posit that examining the effectiveness of existing AIEFs and iteratively improving upon them—rather than producing more AIEFs—may lead to higher industry adoption. Toward this end, our work provides a comparative analysis of existing frameworks using a real-life case study.

At present, there are only a few published use cases that illustrate AIEFs in action (Vakkuri and Kemell 2019; de Swarte et al. 2019). For instance, the Foresight into AI Ethics (FAIE) Toolkit published by the Open Roboethics Institute (Open Roboethics Institute 2019) stems from an AI ethics audit conducted for Technical Safety BC—a safety regulator—in Canada with the full audit report available publicly (Generation R Consulting 2018). O'Neil Risk Consulting & Algorithmic Auditing, a US consultancy, published an audit report of HireVue, a company known for automating hiring and admissions processes (Zuloaga 2021). Apart from a few exceptions, recent findings by Vakkuri et al., suggest that industry professionals largely rely on ad-hoc methods to approach and address ethical risks (Vakkuri and Kemell 2019). This is despite the findings from the same authors that developers do care about the ethical aspects of technologies (Vakkuri et al. 2020a, b). Therefore, some consider AIEFs as a promising means for enabling accountability in AI development processes, both internally or by an independent body. The Algorithmic Impact Assessment (AIA) created by the Treasury Board of Canada is intended as a means to assess and manage the risks of deploying automated decision systems by various departments and agencies within the Canadian government (Treasury Board of Canada 2021).

In this work, we sought to delve deeper into the hindrances to industry adoption of AIEFs by applying four widely different styles of existing AIEFs to the same case study. Our results build on the previous work by providing concrete pros and cons of each form of AIEFs from both auditor and AI-business perspectives.

2.3 Contextualizing the case study

Understanding the geopolitical and domain-specific context is important in interpreting the outcome of any case study. Hence, we briefly provide the context for the startup and its product here.

Our case study involves a recommender system developed by an AI startup based in Montréal, Canada. The client's AI product involves recommending patient-specific treatment options to clinicians. As such, the company's business and design decisions are influenced by at least the following: (a) their status as a startup, (b) the local AI ecosystem in the city of Montréal and Canada more broadly, and (c) medical devices regulation in Canada and the province of Québec.

As one of the birthplaces of deep learning, Canada has an active AI ecosystem heavily supported by the federal and provincial governments (e.g., over $1 billion CAD of public funds have been invested in the city of Montréal's AI ecosystem alone) and hosts over 800 AI startups. While its AI market is small compared to the global AI giants—namely China and the United States—Canada has shown an emphasized interest in responsible AI practices, actively leading initiatives such as the Declaration of the International Panel on Artificial Intelligence. The city of Montréal is also known to be home to the Montréal Declaration for a Responsible Development of Artificial Intelligence (Montréal Declaration Responsible AI 2017), and academic and non-profit AI ethics organizations. As such, an AI startup in Canada is likely to be in a metropolitan environment surrounded by a high density of academic and industry AI practitioners with at least passing knowledge of AI ethics issues.

Over the past decade, we have seen an explosive increase in AI startups worldwide (Tricot 2021). Despite this, only a few studies explored AI ethics activities in smaller companies

(Vakkuri et al. 2020a). In a study involving 249 respondents from 211 software companies, approximately half of the respondents reported not having any fallback plans for handling unexpected use cases of their system and felt that their system could not be misused (Vakkuri et al. 2020a). Among startups, in particular, software developers felt that their sense of responsibility lies primarily with the technical aspects of their products, such that their role in upholding ethical standards remained unclear to them (Vakkuri et al. 2020b).

Although AI companies must adhere to certain regulations, such as data protection and privacy law, the scope and efficacy of current policies are variable and often insufficient for upholding ethics (Wachter and Mittelstadt 2019). Currently, existing policies and ethical guidelines lag behind the rapid adoption of AI in healthcare (Health Canada 2021). Many scholars have highlighted how today's health agencies are unequipped to address ethical risks arising from the adoption of AI (Gerke et al. 2020); a majority of current regulations, such as Health Canada's, evaluate AI technologies as general medical devices on a case-by-case basis (Health Canada 2021). The term "medical devices" refers to a broad range of technologies, defined as a "wide range of health or medical instruments used in the treatment, mitigation, diagnosis, or prevention of a disease or abnormal physical condition." (Health Canada 2004) Other regulatory agencies like the FDA have stringent assessment processes that may be unduly tedious and time consuming for AI developers (Benjamens et al. 2020).

In response to a growing need for standardized assessment processes, new regulatory activities in health-related AI applications are actively taking place. For example, the U.S. Food and Drug Administration (FDA)'s recent regulatory framework, "Artificial Intelligence and Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan" (FDA 2021), proposes a promising assessment process specifically for AI/ML-driven medical devices. However, it has yet to be widely implemented. The World Health Organization also recently published a framework for AI governance in health that articulates definitions, key ethical principles, and policy recommendations (WHO 2021). Although the report briefly discusses how industry stakeholders can implement ethical practices, it primarily focuses on recommendations for government action such as those to be taken by ministries of health. Given the relative novelty of our objective, the study followed an exploratory paradigm which allows for a focus on observation and analysis without a clear, single set of expected outcomes (Yin 2009).

3 AI ethics framework selection

Due to the abundance of existing AIEFs and the lack of concrete evaluation mechanisms in AI ethics, conducting a controlled experiment on a real-life case study comparing every single framework is neither feasible nor effective. Thus, we conducted a landscape analysis as a means to select four AIEFs best suited for our case study. To observe maximum contrast in the outcomes from the AIEFs, we were motivated to find a set of AIEFs that are widely different from one another while serving the same AI ethics function for the case study.

To be systematic in our approach, we completed a narrative review of existing AIEFs guided by Morley et al. (2020)'s comprehensive typology. Morley et al.'s typology includes a list of 106 AIEFs, which is a result of reviewing 425 sources from multiple databases. From the 106 AIEFs, we applied the following inclusion and exclusion criteria:

- *Format* As this study's objectives focus on evaluating the practical application of AI ethics, we specifically aimed to evaluate frameworks which present clear, implementable instructions or guidelines, and thus excluded frameworks that discussed higher-level ethical values
- *Third-party accessibility* Due to the researchers' positionality as external reviewers, we selected frameworks which only require data that third-party assessors can practically attain. As such, AIEFs which incorporate group tutorials and extensive company-wide workshops were excluded.
- *Development stage applicability* The AI product examined in this case study is currently being developed and undergoing clinical trials, and therefore we selected frameworks that were designed to be applied at this stage of development; we excluded frameworks that are meant to be implemented exclusively during early design, late deployment, and post-launch analysis stages.
- *Author and format diversity* To provide an accurate representation and analysis of the current landscape of existing AIEFs, we selected a variety of different assessment methods (e.g., checklist, guideline, questionnaire) to compare their distinctive benefits and drawbacks. We also aimed to reflect the interdisciplinary nature of the AI ethics domain, and thus aimed to choose frameworks generated by authors from different sectors.
- *Grey literature* Historically, systematic reviews have rarely included grey literature, as they are often excluded in well-curated academic databases (Adams et al. 2017). However, excluding grey literature may result in excluding sources from government or business sectors. As such, in an effort to review and analyze material that may not be peer-reviewed, we exclusively included grey literature—which a large majority of AIEFs belong to.

Given that the application of the criteria involves the interpretation of AIEFs that often do not outline the above information, two of the authors independently applied the inclusion/exclusion criteria. We found moderate

agreement between the researchers [inter-rater reliability with Cohen's kappa ($\kappa = 0.545$)]. This process resulted in a total of 38 AIEFs in our corpus. We further classified the resulting AIEFs based on their format, sector of development, required time investment, stage of product development for application, and target user/audience. The complete catalog of the AIEFs can be found at this link: https://tinyurl.com/mr8mv2v4.

This process revealed two broad forms of AIEFs: checklists and step-by-step processes. While checklists seem to structure ethical principles as clear and measurable standards (e.g., a series of close-ended questions), step-by-step processes invite open-ended exploration of ethical issues. In addition, there was a wide range of intensiveness—i.e., amount of time investment and thoroughness or detailed nature of the analysis—involved in applying a given AIEF. We selected the following four AIEFs: Foresight into AI Ethics Toolkit (Open Roboethics Institute 2019), Guidelines for Trustworthy AI (High-Level Expert Group on Artificial Intelligence 2019), Ethics and Algorithms Toolkit (Anderson et al. 2018), and Algorithmic Impact Assessment (Treasury Board of Canada 2021). The Foresight into AI Ethics Toolkit (FAIE) took the longest time to apply at approximately 20 h, while the Guidelines for Trustworthy AI (GTA) framework took about 9 h to apply. The duration of applying the Ethics and Algorithms Toolkit (EAT) and Algorithmic Impact Assessment (AIA) was relatively shorter, at approximately 4 h and 90 min, respectively.

They were chosen due to the fact that they all provide a means to assess AI ethics risks—a function applicable for a pre-deployment AI system, as is the case for our case study—and suggested to be implemented during the development stage of an AI system. They also represent the two different forms of AIEFs across the opposite ends of the intensiveness spectrum. We briefly describe each framework below. Table 1 provides a snapshot of the four AIEFs.

3.1 The Foresight into AI Ethics Toolkit (FAIE)

FAIE is developed by a non-profit think-tank, the Open Roboethics Institute, which presents a step-by-step guideline with the intention of anticipating, managing, and mitigating AI ethics risks. FAIE introduces ethical principles using accessible language and illustrates each step of the assessment process with an example use case and is, therefore, applicable for all stakeholders. In comparison to the other selected frameworks, FAIE focuses more on stakeholder perspectives and values, posing open-ended questions which allow stakeholders to relate stakeholders' personal and organizational goals with societal values.

3.2 Guidelines for Trustworthy AI (GTA)

GTA is developed by the European Commission's High-Level Expert Group on Artificial Intelligence. It aims to guide assessors in building "trustworthy AI" systems. It provides a list of questions that primarily require "yes" or "no" answers. Its emphasis is on the details of the technology design, and therefore requires the most comprehensive and technical knowledge about the AI system and related data management process being assessed.

3.3 Ethics and Algorithms Toolkit (EAT)

Created by the County of San Francisco Data Science Team, EAT presents a set of step-by-step instructions with open-ended questions and a fillable worksheet. The worksheet includes a guideline for interpreting and recording the results of the assessors' ethics evaluations across a risk spectrum (e.g., high or low risk). The EAT is similar to FAIE in that they both provide clear guidelines for open-ended exploration of AI systems' ethics risks. However, the process of recording and designating specific risk levels leads the outcome of EAT to be parametric.

3.4 Algorithmic Impact Assessment (AIA)

Developed by the Government of Canada, AIA aligns with the federal government's *Directive on Automated Decision-Making*. It is to be used to determine a technology's impact levels according to federal policy. The AIA is delivered as an online questionnaire and primarily requires "yes" or "no" answers to the provided questions. The questions consider potential ethics risks as well as risk mitigation strategies implemented by stakeholders. The questionnaire does not survey the entire scope of established ethical principles, as the other frameworks aim to do. Rather, it focuses on the process, data, and system design decisions relevant to the *Directive*. Each question is associated with different points which cumulate to a "Current Score", "Raw Impact Score", and "Mitigation Score". These scores correspond to "impact levels" which dictate the policy requirements that the technology must adhere to.

Additionally, Fig. 1 presents a snapshot of the different characteristics of the frameworks. This includes the format, phases, instructions, and results of how each framework is applied in AI system evaluations. (1a) "Persona profiles" were developed for each stakeholder group, highlighting their goals, values, and unique challenges. The information was gathered via one-on-one interviews conducted by the researchers. (1b) "Value questions" aimed to identify potential risks and action items related to the AI system's input, model, and output. The questions focused on addressing societal values, including transparency, trust, accountability,

Table 1 Framework selection and characteristics summary

Framework names			
Foresight into AI Ethics Toolkit (FAIE)	Guidelines for Trustworthy AI (GTA)	Ethics and Algorithms Toolkit, Part 1: assess algorithm risk (EAT)	Algorithmic Impact Assessment (AIA)
Author(s)			
Non-profit organization	Intergovernmental agency	Local government/Academia	Federal government
Intended purpose			
A step-by-step guided process for evaluating how the technology's stakeholders relate to each other and the product itself, what values are important for each stakeholder group and society as a whole, and how the technology relates to and impacts those values	An assessment checklist used to operationalize and evaluate progress toward "trustworthy AI"	A worksheet with indicators which guides assessors in evaluating their algorithm's level of ethical risk	An online questionnaire based on Canada's *Directive on Automated Decision-Making* and assigns scores according to the technology's impact level and implemented mitigation techniques
Intensiveness			
25-page roadmap for discovering, evaluating, and managing ethics risks. FAIE involves three phases: 1. Identify your use case and stakeholders 2. Discover ethics risks 3. Create a roadmap and implement it A total of 11 steps are completed, including interviews with multiple stakeholder groups	41 pages and includes definitions of ethical principles and explanations for how stakeholders can achieve "Trustworthy AI". The main application portion of the framework is a 6-page checklist comprising 60 questions	A series of questions that aims to guide assessors in understanding the ethical risks posed by an algorithm. The EAT presents a 13-page set of guidelines and questions and a 4-page worksheet with risk assessment matrices	An online questionnaire that is stated to take 35 min to complete. The AIA includes 48 risk identification and 33 risk mitigation questions
Intended audience			
All stakeholders	Unspecified	AI for the public sector	AI for the public sector

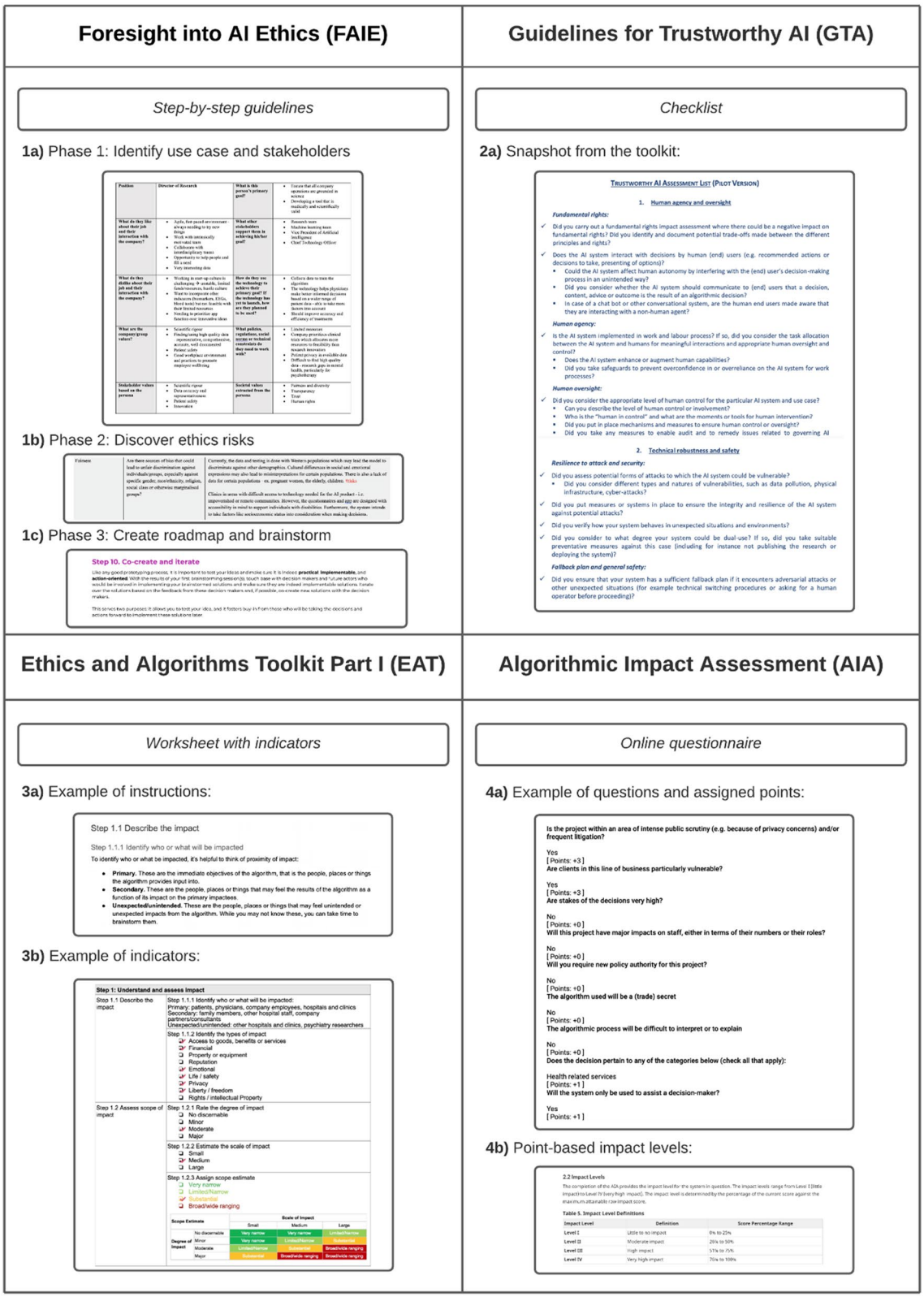

Fig. 1 Snapshots of the selected frameworks applied to our case study. All four AIEFs aim to help assess AI ethics risks and produce vastly different styles of output

human rights, autonomy/consent, and fairness. (1c) Step-by-step guidelines were provided for developing a customized "roadmap" to manage and address identified ethical risks. (2a) An assessment list used to operationalize and evaluate progress toward "trustworthy AI" based on seven key requirements: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination, and fairness, (6) environmental and societal well-being, and (7) accountability. The assessment list operates generally as a checklist but also prompts critical reflection beyond "yes" or "no" answers. (3a) A PDF document outlining the steps and questions stakeholders should consider when evaluating the algorithm and its data. (3b) The instructions correspond with a fillable worksheet consisting of clear benchmarks or criteria for recording the outcomes of stakeholders' ethics evaluations. (4a) A questionnaire assessing the potential risks of an automated decision system and the organization's risk mitigation strategies. Points are assigned for each question. (4b) The questionnaire assigns a score for identified potential risks and another score for implemented mitigation techniques, and these scores correspond to an impact level determined by the *Directive on Automated Decision-Making.*

3.5 Discussion

In this section, we provide a discussion of the patterns found in the landscape analysis and a deeper analysis of the four selected frameworks.

3.5.1 Gap between measurement and implications of analyzing AIEFs

Of the 38 frameworks included in the analysis, few integrate concrete evaluation outcomes with defined risk levels (e.g., the *Ethics and Algorithms Toolkit* developed by the County of San Francisco Data Science Team and the *Algorithmic Impact Assessment* developed by the Government of Canada), while most other AIEFs offer more open-ended results [e.g., the *EthicalOS Toolkit* (Institute for the Future), the *Foresight into AI Ethics Toolkit* (Open Roboethics Institute 2019), and *Ethics in Practice: A Toolkit* (Santa Clara University)]. While flexible outcomes may allow for nuanced exploration and adaptation to different contexts, setting explicit measures may lead to more actionable and enforceable standards (Theodorou and Dignum 2020). Furthermore, even among the selected AIEFs with measurable results, there is a need for clearer and more comprehensive descriptions of each risk level's definitions, implications, and subsequent steps. For instance, the *Algorithmic Impact Assessment* produces a numerical "Impact Score" and "Mitigation Score" which correspond to national policy requirements for autonomous systems. However, there remains some ambiguity surrounding how these values are calculated: Why were certain questions weighed more heavily than others? Accordingly, are all questions, and therefore ethical principles, equally weighted in this assessment system? Furthermore, while the AIA subsequently presents assessors with the "Impact Level" of their AI technology, the description of such "levels" is quite broad and inoperable. For instance, stating that technology has "moderate impacts on the rights of individuals or communities" does not adequately provide stakeholders with actionable steps to mitigate those impacts. The lack of explainability of AI contradicts ensuring transparent AI, which is one of the core AI ethics values (Adadi and Berrada 2018). Therefore, improving the clarity and explanations for AIEFs outcomes, particularly with explicit benchmarks, will offer better mechanisms for regulatory bodies to enforce ethical standards and increase accountability among industry stakeholders. Furthermore, as there is currently no standard outcome among AIEFs, it is even more difficult for stakeholders to determine which method to use. This research, thus, aims to examine a variety of different outputs to accurately represent the current landscape of existing frameworks.

3.5.2 Need for opportunities to engage stakeholders

Many researchers emphasize the need to include perspectives of multi-stakeholders in AI ethics (Bogina et al. 2021). However, AIEFs currently present insufficient opportunities for participatory design (Madaio et al. 2020), thereby leading to potentially inaccessible and biased ethics assessment processes. Among the four AIEFs we analyzed in this project, only the *Foresight into AI Ethics Toolkit* (FAIE) incorporates concrete steps to include the perspectives of diverse groups of stakeholders. For example, the AIA can theoretically be completed by a single individual without stakeholder engagement, although multi-stakeholder engagement is encouraged. However, several frameworks (i.e., the GTA and AIA) involve in-depth technical assessments, which necessitate expert knowledge to implement, thus leaving minimal opportunities for non-technical stakeholder involvement. Moreover, given that algorithmic discrimination exists as a product of societal inequity (Noble 2018), AI ethics assessments must consider the historical, political, and institutional contexts within which their systems operate. Accordingly, research methods that consult historically marginalized or underrepresented groups within data science can guide equitable design decisions and also clarify the consequences of such decisions (Katell et al. 2020). By centering the perspectives of those most affected by unethical technologies, AI developers can begin to learn about and address biases that the tech industry too often fails to notice (Harrison et al. 2020). Engaging in participatory action research, Katell et al. (2020) developed the *Algorithmic*

Equity Toolkit by consulting with local community groups, advocacy campaigns, and policy stakeholders, with the intention of the framework being used by the very same stakeholder groups. We pose that this inclusive paradigm is implemented in assessing real-world technologies—as steps *within* AIEFs (rather than for *designing* frameworks)—such that impacted stakeholders may directly provide insights to improve AI systems.

3.5.3 Need for context-specific industry standards

While the current landscape of AIEFs largely addresses established AI ethics principles, there remains a gap surrounding how one should address legal and ethical considerations within specific industry domains. In practice, AIEFs operate differently depending on existing industry practices, standards, and cultures. What value does AI ethics add to a field with widely recognized and deeply embedded ethical structures? Floridi and Cowls (2022) argue that AI ethics principles converge at the four commonly recognized bioethics principles—beneficence, non-maleficence, autonomy, and justice—and proceed to add a new principle specific to AI ethics. The additional value of explicability, which encompasses both intelligibility (how does the technology work?) and accountability (who is responsible for how the technology works?), addresses the AI specific that these technologies are often not understandable to the general public and even sometimes to the developers themselves (Floridi et al. 2018). The need for a new principle, therefore, demonstrates that introducing AI in the healthcare domain requires additional ethical considerations beyond bioethics.

In this study, we realized the potential for tensions to arise between AI ethics principles and existing standards in domain-specific cases. For instance, the healthcare domain presents an interesting case study because of its well-established ethics and stringent regulations. More specifically, stakeholders in digital health technology development must comply with clinical and medical ethics principles, as well as medico-legal codes such as the Health Insurance Portability and Accountability Act (HIPAA) (CDC 2019) and the Personal Information Protection and Electronic Document Act (PIPEDA) (Office of the Privacy Commissioner of Canada 2018). Although these regulatory systems share numerous principles with AI ethics, including privacy and data protection, they may not adequately cover concerns relating to AI-driven health technologies. For example, HIPAA relies on outdated privacy strategies such as allowing de-identified health information to be shared freely for research and commercial purposes (Cohen and Mello 2018). With the emergence of "Big Data", the possibility of data triangulation—re-identifying data through amalgamating multiple datasets—threatens HIPAA's capacity to protect patient privacy (Gerke et al. 2020). Stakeholders may, therefore, feel the need to exclusively follow or prioritize medico-legal codes such as HIPAA, rather than spending additional time and resources to further assess AI-specific ethics risks. As such, there may be challenges in complying with industry standards due to regulatory agencies being unfamiliar with AI technologies; existing processes are often unclear, unduly tedious, and insufficient in addressing AI-specific concerns.

4 Comparative case study

The case-study approach is effective at answering "how" questions (Yin 2003). To answer our research question—"How effective are the various types of AIEFs in helping businesses assess AI ethics risks?", we selected the qualitative case-study method. To assess the benefits of existing AIEFs, we put the selected four frameworks (FAIE, GTA, EAT, and AIA) as interventions to the real-world AI system. More specifically, we adopted a comparative case study to explore similarities and differences between each framework (Knight 2001). As independent researchers, we assumed the role of a "third-party auditor" conducting ethics assessments for the startup (client) using the frameworks. Before the data collection phase of our work, all company stakeholders were made aware that the audit was being conducted by external researchers.

The client company was founded within the last 5 years, and has approximately 25 employees. The company's main AI product aids physicians in selecting treatments for patients with clinically diagnosed major depressive disorder. The product is currently in development and undergoing clinical trials.

In this section, we first describe the client's main product as the subject of our case study and contextualize our analysis. Subsequently, we outline our method for applying the frameworks and analyzing their respective outputs. In presenting our results, we report on the experience of using the AIEFs from an auditor's perspective, provide a qualitative comparison of the framework output, and perceived benefits of each framework as expressed by the client in a post-mortem discussion. Furthermore, as the scope of this study is not to provide technical solutions or feedback to software-centered developments, we did not assess the company's system, dataset, or algorithms. Figure 2 provides an overview of the comparative case-study process. This study was approved by the University research ethics board (REB File # 20-06-022).

4.1 AI system description

The client's product collects user data through clinically validated questionnaires about the patient's daily feelings, behaviors, and activities. The patient, or an approved

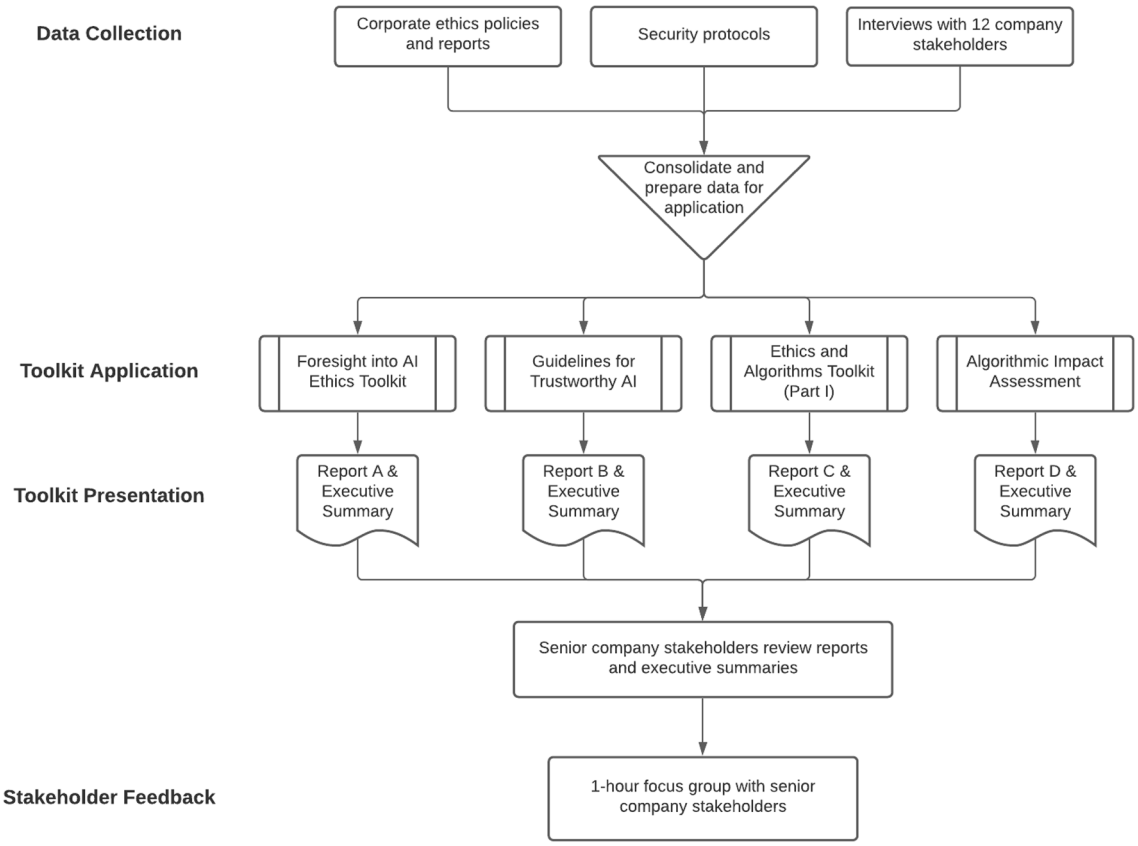

Fig. 2 The research process of the comparative case study

representative (i.e., a family member or close friend), routinely answers these questionnaires on an app through their mobile device. Using an AI model to analyze the users' answers, the technology presents physicians with the best medication options for each individual patient based on their clinical profiles. These treatment recommendations correspond to the *probability of recovery using the treatment*, which is a value calculated based on patient data, and a list of clinical and other factors considered in the system. The physician then examines this information to ultimately make the final decision about which treatment the patient receives. The physician continuously monitors the patient's well-being and any potential side effects with the selected medication through the app to ensure that the medication remains suitable. Figure 3 illustrates a visual schematic of the system.

4.2 Methods

The four AIEFs require varying amounts and types of information to complete. We collected the maximum of the data required by the four frameworks through (a) interviews with the client and potential stakeholders of the client's product (Sect. 4.2.1), and (b) conducting secondary research. The secondary research involved examining publicly and internally available documentation related to the client's main product, including security protocols (e.g., data breach plans) and corporate policies (e.g., ethics policy). The same collection of data, as derived from the interviews and secondary research, was used to implement the FAIE, GTA, EAT, and AIA—each according to their distinct processes (Sect. 4.2.2). In this process, we documented our experience in applying the frameworks from an auditor's perspective. Subsequently, the output resulting from each framework, as well as an executive summary were presented to the client. A focus group with the client was conducted to gather the client's feedback (Sect. 4.2.3).

4.2.1 Stakeholder interviews

Interview protocol A total of 12 participants were invited to a semi-structured teleconference interview with a researcher. As the purpose of the interviews was to identify and collect information on perspectives held by representative stakeholder groups of the AI-health startup, we used a purposive sampling method (Etikan 2016). We used either Zoom or Google Hangout depending on the participant's preference. Each interview took 30 to 60 min and was audio recorded

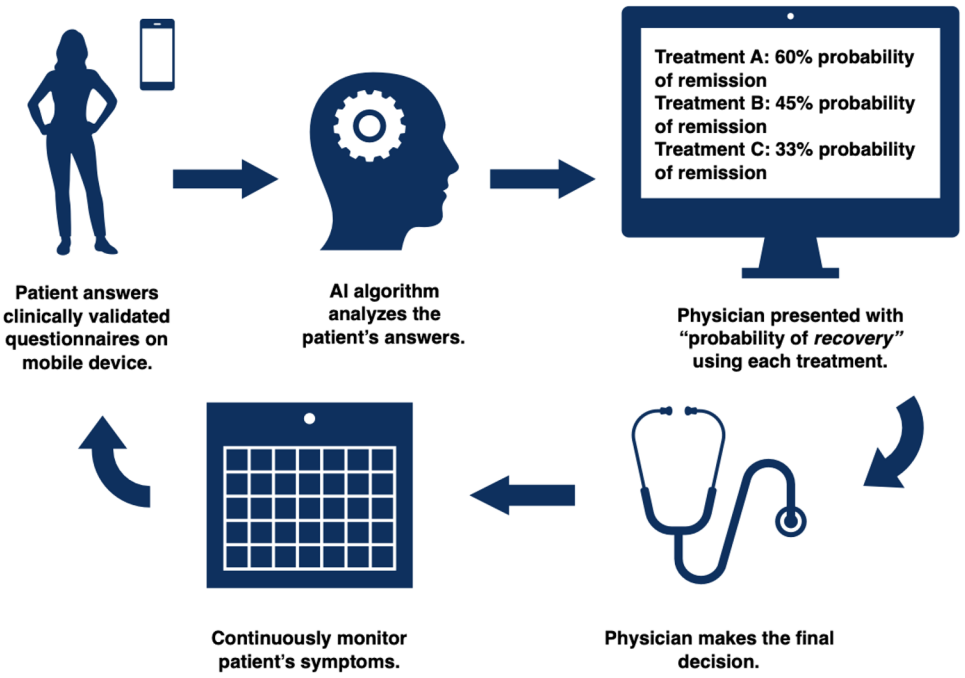

Fig. 3 A schematic of the client's AI technology. This schematic was used to explain the AI system to potential patient users at the beginning of the semi-structured interviews

and transcribed. All participants provided written and verbal consent prior to the interview.

Internal stakeholder interviews Company (internal) stakeholders were recruited through written instructions via email and the company Slack channel. We interviewed six internal stakeholders who hold the following roles: Chief Science Officer, Chief Technology Officer, Vice President of AI, Director of Research, Clinical Trial Coordinator, and a physician collaborator. Participants were asked questions pertaining to their personal thoughts about the technology. Following the ethical themes presented by the frameworks, participants also discussed societal values—such as transparency, trust, fairness and diversity, accountability, and human rights—as they relate to the use case. These stakeholders were asked questions such as, "What do you hope to achieve through your position at the company?", as well as more technical questions relevant to their positions like, "How did you test the algorithm before you put it to use?".

Potential patient-user interviews We also recruited six stakeholders external to the company, who represented "potential patient-users" through social media. These participants did not have any previous or expected future interaction with the technology and were only provided with a simple description of the technology (see Fig. 2). Demographic information was not collected on any of the participants due to the lack of its necessity for the research question. During the interview, potential users answered questions about their personal opinions on AI and mental healthcare, such as "How much do you know about AI and its use in mental healthcare?" and "What are your thoughts on Artificial Intelligence (AI) in mental healthcare?". Additionally, these external stakeholders provided insights into questions specific to the technology, including "If you were a patient-user, what would you hope to gain from using this technology?" and "How important would being able to understand the AI algorithm be to you?".

4.2.2 Framework application and presentation

Upon consolidating the data collected, we followed the instructions outlined in the AIEFs as written. One member of the research team with a limited technical background in AI applied all four AIEFs while guided by another member with AI and AI ethics expertise.

Once all four frameworks were completed, output from each was prepared as individual reports to be sent to the client. To preserve the integrity of the assessment methods, we kept the intended format of the output expressed in each AIEF. The FAIE report was 21 pages long and followed the general structure of the assessment process, beginning with an exploration of the stakeholders and use cases of the technology, followed by an in-depth analysis of potential ethical risks and value tensions, and ending with suggestions for addressing identified ethical challenges. The GTA report was 15 pages long and listed all the questions posed by the guideline. Answers to questions were reported as primarily "yes" or "no", with some exceptions for questions which require nuanced considerations, such as "To what degree does the system's decision influence the organization's decision-making process?". The EAT report was composed of two documents: (1) the worksheet with benchmarks indicating risk levels and (2) paragraph- and point-form explanations

for how we assigned risk levels. In total, the EAT report was 12 pages long. Finally, the AIA report was nine pages long and presented a PDF printout of the online questionnaire. The report listed the questions posed by the framework, along with our answers and their associated points. The final scores and requirements specific to the assigned "impact level", as mandated by the *Directive on Automated Decision-Making* (0), were also described.

Moreover, we produced two-page executive summaries for each framework highlighting a brief description of the framework and the ethical risks discovered. For each framework, we provided a short overview of the theory and process, our experiences implementing the frameworks, the nature of their outputs, and key findings.

4.2.3 Client feedback

The full report of the framework outputs and their respective executive summary was sent to the client for review. We pseudonymized any identifying information about each framework in the reports to mitigate potential biases (e.g., FAIE was referred to as Framework A). The client was given 3 weeks to review the documents. Subsequently, we held a semi-structured focus group session via teleconference (Zoom) to gather their qualitative feedback on each framework.

Two members of the client—the Chief Science Officer and the Director of Research—participated in the session. The researchers began with a brief presentation of the study's methodology and the summaries of each AIEF's process and outcomes (i.e., a review of the material sent to them three weeks ago). For each framework, the participants were asked about their perceived value of each framework. They were prompted with questions relating to specific aspects of the framework, including: "How actionable do you find the outcomes of Framework A?", "Is there any preference between 'checklist' versus 'guideline' assessment processes", and "How important are measurable indicators for assessing ethical risks?". Additionally, we asked questions that aimed to explore the practicality of the frameworks, such as "How much of the report reveals previously unknown ethics issues pertaining to your product?" and "Would you consider using the frameworks yourself?" The focus group session took 1 h.

4.3 Results

In this section, we outline a summary of the findings on each framework (also see Table 2). We contrast the AIEFs from an auditor's perspective in Sect. 4.3.1, followed by the findings from the client's perspective captured in the focus group session in Sect. 4.3.2.

4.3.1 Auditor's assessment

As auditors using the AIEFs, one of the major differentiators in our auditing experience depended on the checklist-based vs. process-oriented nature of the frameworks. For instance, with relatively simple procedures and a structured list of questions, the checklist-based frameworks took overall less time and effort to complete than the process-based AIEFs. A researcher with a limited technical background in AI was able to complete the AIA, a short online questionnaire, in 90 min. The comprehensive checklist presented in GTA took the same researcher 9 h to complete. In comparison, the researcher spent 20 h applying FAIE to produce a 21-page report, which included a stakeholder analysis. Four hours were spent applying the EAT—another process-oriented AIEF—which resulted in a four-page summary of various risk levels associated with the AI technology. Categorizing the AIEFs by the degree of rigor, among the more comprehensive frameworks, the process-oriented framework (FAIE) required more time than the checklist-based framework (GTA) to complete. Likewise, in comparing the more efficient frameworks, the EAT required more time than the AIA to complete. Below, we highlight the experiential differences between the two styles of AIEFs.

4.3.1.1 Checklist-based frameworks: GTA and AIA The closed-ended nature of the questions in the checklist-based frameworks oriented us toward conclusive outcomes for each ethical issue considered. It provided a sense of preemptively checking the existing AI system design against a specific set of benchmarks (e.g., the question "Did you take measures to enhance privacy, such as via encryption, anonymization and aggregation?" in the GTA framework preempts the need for such privacy-enhancing measures to be implemented).

To that end, the GTA and AIA have a heavy emphasis on more technical aspects of the technology and require a certain degree of technical expertise and familiarity with the product to complete (e.g., "Will the system require the analysis of unstructured data to render a recommendation or a decision", AIA). Our experiences implementing the two frameworks primarily involved consulting the company's software development team and company policies, rather than speaking directly with stakeholders (e.g., potential users). Therefore, both frameworks provide a strong impression that ethics assessments should be conducted by a technical expert.

Many of the non-technical questions were also posed in a close-ended manner and gave the impression that some key ethical issues are oversimplified. For instance, a simple "yes" or "no" answer seems to be expected in the question "Did you ensure that the social impacts of the AI system are well understood?" (GTA, pg. 33). However, from the perspective of an auditor, such a simplistic answer without

Table 2 Summary of the selected frameworks' format, duration of application, outcomes, and key advantages and challenges

Foresight into AI ethics (FAIE)	**Framework format**	**Duration of application**
	Step-by-step guidelines	20 h
	Outcomes	
	"Persona profiles" outlining the values and goals of each stakeholder. (Fig. 1.1a) Identified ethics-related challenges tagged as either #risks or #needAction through "value questions" assessing the technical facets of the AI product. (Fig. 1.1b) A step-by-step guideline or "roadmap" to addressing the ethical tensions identified. (Fig. 1.1c)	
	Advantages	**Challenges**
	Of the four frameworks, FAIE presents the most accessible guidelines in terms of explaining the reasoning and theoretical principles behind each step Presents opportunities and guidelines for directly engaging with stakeholders (e.g., developing "Persona profiles" and posing interview questions)	Difficult to establish measurable, objective outcomes Relatively more time consuming to complete, as the questions required more thought and discussion
Guidelines for Trustworthy AI (GTA)	**Format**	**Duration of application**
	Checklist	9 h
	Outcomes	
	An extensive list of questions whose answers pointed to potential ethical risks. A majority of the questions inquire whether the company assessed certain ethical risks or implemented particular mitigation strategies. (Fig. 1.2a)	
	Advantages	**Challenges**
	Offers the most comprehensive report of fulfilled technical requirements Overall, simple to implement and does not require extensive reporting of results	Most questions required technical expertise to answer Some questions were unclear and tedious to answer While the checklist generally only prompts "yes" or "no" answers, certain questions necessitate more nuanced considerations. For example, questions such as "Did you assess whether there could be persons or groups who might be disproportionately affected by negative implications?" prompt critical reflection beyond simple "yes" or "no" answers
Ethics and Algorithms Toolkit (EAT) (Part I)	**Format**	**Duration of application**
	Worksheet with indicators	4 h
	Outcomes	
	Evaluations of the algorithm's level of risk are benchmarked by graded indicators in six major areas: impact, data use, accountability, third-party methodology, historic bias, and technical bias. (Fig. 1.3b)	
	Advantages	**Challenges**
	Presents a unique balance between open-ended questions and measurable outcomes with evaluation indicators The worksheet was easy to fill out based on the instructions provided (Fig. 1.3a)	The worksheet's benchmarks may be ambiguous without contextual information. Assessors should document the reasoning behind their evaluations Difficult to establish inter-rater reliability The scope of assessment is not as comprehensive as other frameworks

Table 2 (continued)

Algorithmic Impact Assessment (AIA)	**Format**	**Duration of application**
	Online questionnaire	90 min
	Outcomes	
	The questionnaire assigns a score for identified potential risks and another score for already implemented mitigation techniques (Fig. 1.4a), and these scores correspond to an assigned impact level with specific requirements mandated by the Directive on Automated Decision-Making (Fig. 1.4b)	
	Advantages	**Challenges**
	The AIA directly references existing regulations that the AI product must adhere to Questions were easier to interpret than the GTA's when assessing the technical elements of the technology	The AIA does not survey the entire scope of established principles, but rather focuses on the business processes, data, and system design decisions involved in developing an automated product; therefore, assessors may need to consider additional ethical aspects beyond what is covered by this framework. Despite the questionnaire's relative simplicity, its point system is difficult to understand, as it remains unclear how points were assigned and what they mean relative to the assigned "Impact Level"

a detailed description of the efforts made to study the social impacts misses the point of conducting an ethics assessment.

The absence of extensive analysis or multi-stakeholder discussions—an element common to process-oriented frameworks—combined with the lack of opportunities to ask or explain "why" and "how" in checklists seems to limit their ability to help evaluate the full scope and depth of ethical issues present.

4.3.1.2 Step-by-step process guides: FAIE and EAT In comparison to checklist-based frameworks, process-based frameworks such as FAIE and EAT guide stakeholders through the steps of ethical evaluations and invite open-ended answers (see Table 2). Rather than posing "if" questions, these frameworks prompt auditors to consider "how" and "why" certain ethical principles are salient to the AI system being examined. This also required us to collect perspectives from multiple stakeholders within and outside of the client's company.

In addition, the process-oriented frameworks posed questions that led the researchers to review the company's missions and values throughout the assessment process. For example, FAIE prompted us to develop "persona profiles" which require in-depth conversations with company stakeholders to capture each group's values, goals, and challenges. These profiles were then used to identify value tensions between stakeholder groups directly drawing from stakeholders' personal experiences.

The open-ended nature of certain processes—such as brainstorming with other stakeholders, and holding consultation sessions or interviews—seems to help the frameworks be generalizable to different industries or application domains. With both FAIE and EAT, we were able to consider how the different stakeholders expressed the advantages and shortcomings of the AI product within the application context of mental healthcare in Canada. However, the open-ended nature of process-oriented frameworks also gave the impression that the outcome of the process can be unpredictable and subjective. In addition, following through with such processes require much more resource and willpower than the efforts required to complete a checklist.

Based on our experiences applying the frameworks, we found that step-by-step processes invite open-ended investigation of value trade-offs, alternative design decisions, and sociocultural contexts. While these types of frameworks help explore a wider range of impacts an AI system can have, their lack of quantifiable benchmarks leaves some of the results open to subjective interpretations.

4.3.2 Client's assessment

By the time we started our study, the client had already published a report and hours of internal discussions related

to their ethical standards as a company. Overall, the client expressed that the outcome of the AIEFs collectively reflects some of the AI ethics-related discussions they've had internally before.

As the AI product is currently undergoing clinical trials, the company recently began working on their quality management system (QMS)—a set of policies, processes, and procedures medical technology companies must adhere to. Throughout the focus group session, the client likened AIEFs to QMS and frequently drew analogies from QMS, but stated that the latter is much more laborious to complete as it requires one "to identify the likelihood of the risk, the severity of the risk, and the risk mitigation strategies" (P1). Given that one of the shared objectives of AIEFs is to identify and mitigate AI ethics risks, the client noticed the potential utility of adopting AIEFs to document ethical risks.

4.3.2.1 Contradictions exist between AI ethics and medico-legal requirements While discussing the results from FAIE, the client was reminded of an internal discussion where the AI ethics principle for transparency and existing medico-legal requirements posed tension. For example, the *GTA* suggests that developers "ensure an explanation as to why the system took a certain choice resulting in a certain outcome that all users can understand." (p. 31) The client saw this recommendation to be inappropriate for clinical use cases because providing a full set of treatment results without a physician's interpretation can lead patients to question the physician's expertise. The client described a scenario wherein a patient may disagree with their physician's treatment plan and illegally obtain another medication solely based on the AI's recommendations. The client expressed that "[we] never want to give a situation in which you can cut the human [clinician] out of the loop." Therefore, while transparency remains prioritized in educating physicians about the technology, there are exceptional clinical limitations when disclosing certain types of information to patients.

While the above tension does not relate to the transparency about the algorithm, the client also expressed that they are prevented from making details of the algorithm more transparent to potential users; during their pre-deployment user trials, communicating the workings of the algorithm to their satisfaction was seen by regulators as an interference to the requirements of conducting a double-blind clinical study. The client remarked that ultimately, "[we] have to prioritize legal requirements over situations where it's not clear what [is] ethical" because the former presents clearer standards and harsher consequences. They also expressed that they occasionally find themselves sacrificing personal values to conform to existing regulations and industry precedents that may be outdated.

4.3.2.2 Desire for AIEFs to produce familiar indicators and objective benchmarks The client also exhibited concerns regarding the subjective nature of the ethics assessments: "given [that] different people…come from different trainings, how would they respond to the same questionnaire and is it reliable between people?" (P2) This question demonstrates a critical challenge in AI ethics. Particularly for internal ethics assessments, stakeholders from varying cultural, industry, or academic backgrounds may understand ethical processes differently, and therefore evaluate risks from different perspectives (Goodwin and Darley 2010). Furthermore, depending on one's objectives, assessors using AIEFs could invest varying degrees of effort toward the ethical analyses and still "complete" the frameworks as instructed. These variables may, thus, prevent open-ended frameworks from becoming a standardized form of ethics risk assessments.

Consequently, the client preferred checklist-based frameworks as being replicable and perceived them to be less subjective. They also expressed that overall, "the more specific the outcome, the better, because sometimes vague statements are not really that helpful. Either [they]'ve already thought of that or it's just too vague to be helpful." (P2) Checklists are ultimately easier to follow and standardize across stakeholder groups and conditions. In reviewing the *EAT,* the client enthusiastically welcomed the indicator-driven style of the framework as it was familiar to them through the QMS, which is also an indicator-driven mechanism.

4.3.2.3 Different AI ethics frameworks for different product development stages At the end of the feedback session, we asked the client to compare the AIEFs and verbalize how they see the frameworks contribute to their existing operations. We asked questions such as "How might checklist- and process-based frameworks function differently?" and "What roles will AI ethics frameworks play in design-related decision-making?" The client articulated that the frameworks serve two distinct utilities: externally, AI ethics frameworks may function as a display of ethical merit—which the stakeholders likened to a "report card" or "certification"—and involves a "feel-good" factor associated with the accomplishment; internally, the frameworks may be helpful for identifying clear action items to address.

Furthermore, the stakeholders also recognized values distinct to the type of AI ethics frameworks. They saw the value of using process-oriented frameworks in earlier stages of product development. As the developers may not be familiar with moral implications pertinent to the product during these nascent phases, frameworks like FAIE and EAT were seen as useful tools to educate the team on relevant AI ethics concepts and concerns. Then checklist-based frameworks like GTA and AIA were seen as valuable mechanisms to affirm and re-evaluate one's progress in later stages of product

design and regularly "update" their designs or policies as needed: "if you're doing interim assessments…I think the checklist is better, but I do think that the guideline form is nicer for leading discussion that leads to identifications of problems which can then be addressed by checklists after." (P1). As such, the client saw the two styles of frameworks being used in conjunction: process-oriented frameworks used first in the early stages of design, followed by the regular implementation of shorter checklist-oriented frameworks to monitor the company's progress. This aligns with previous research that indicates that a one-size-fits-all solution does not apply to AI ethics (Bellamy et al. 2018).

5 Discussion

This study presents our first step to address the question "how effective are the various types of AIEFs in helping businesses assess AI ethics risks?" We applied four different AIEFs to a real-world use case as a means to better understand the value of adopting AIEFs in the AI industry. The time it took to apply the frameworks ranged from 90 min (AIA) to 20 h (FAIE). We found that the utility of adopting AIEFs as a feel-good factor or a means to present a company as a responsible innovator is common across the AIEFs. However, as we had suspected from our landscape analysis, the experience of applying the frameworks varied extensively between checklist-based frameworks (GTA and AIA) and process-based frameworks (FAIE and EAT).

Notably, the checklist-based frameworks took a shorter amount of time to complete, and mostly required technical expertise to complete, with minimum expectations for the auditor to engage a wide range of stakeholders or be thoughtful in close-ended questions such as "Did you ensure that the social impacts of the AI system are well understood?" On the other hand, stakeholder engagement was central to the discovery of ethics issues in process-oriented frameworks. The open-ended nature of these process-oriented frameworks not only requires more commitment of resources (e.g., stakeholder consultation, time to completion) but the diverse range of hard-to-benchmark issues that can surface as a result can be unsatisfying for businesses looking for concrete and measurable outcomes. Below, we discuss these differences and identified obstacles to industry adoption.

5.1 No such thing as one-size-fits-all AI ethics framework

The main benefits of adopting the time-consuming, process-oriented frameworks such as FAIE and EAT are in: (a) rich and open-ended *ethics issue discovery* and exploration, (b) fostering multi-stakeholder AI ethics-related discussions, and (c) raising awareness of possible issues that may need to be addressed. They foster an explicit articulation of priority values and potential ethics risks associated with a product at the early stages of the development cycle. Adoption of these types of AIEFs provides value by effectively creating a roadmap/guideline that internal stakeholders can refer to make specific design and policy decisions. As the process of following these AIEFs involves multi-stakeholder consultations, it naturally serves to educate and engage stakeholders on AI ethics concerns relevant to the product being developed. However, we expect the cost of the process-oriented frameworks to overpower their benefits if they were to be conducted multiple times for the same product. Furthermore, the subjective nature of these frameworks, and the fact that many of the identified issues are likely unaccompanied by clear benchmarks to highlight their severity makes the frameworks ill-suited for *continuous* monitoring and discovery of AI ethics risks.

In contrast, checklist-based frameworks could be applied multiple times on a regular interval of time and effectively help detect AI ethics risks that creep up over incremental design changes. As such, they have the potential to serve the function of *ethics issue monitoring*. They can help incorporate AI ethics accountability to a company's operation as it enables measurement of progress toward AI ethics goals using specific indicators. Note that the version of GTA and AIA we employed in our study aims to serve the *ethics issue discovery* function—rather than *ethics issue monitoring*—which it does in highly limited ways in comparison to FAIE and EAT. Many of the checklist-based frameworks are not designed with specific technology, application domain, or regulatory contexts in mind. These AIEFs do not invite the auditors to explore the diversity and severity of AI ethics risks that are specific to the system being assessed. The majority technical nature of the question also limits the amount of multi-stakeholder perspectives that can be incorporated into the ethics issue discovery process.

In broad terms, we echo the remarks by our client that the two styles of AIEFs should be used for different stages of a product's development cycle; rather than being used as a stand-alone, one-time intervention, checklist-based frameworks should follow process-oriented framework to help monitor and integrate AI ethics accountability. This also implies that there is more to be gained by the AI ethics community in revising existing AIEFs and better guiding their users about the known limitations of the frameworks. In particular, process-oriented frameworks could be revised to produce checklist-like mechanisms. Checklist-based frameworks could, in turn, be redesigned to be flexible to incorporate issues discovered as a result of applying process-oriented frameworks.

Our results also highlight that no one AIEF should be adopted as a one-time AI ethics assessment mechanism. Adopting an iterative approach to assessing ethical risks

allows companies to account for unexpected changes, produces feedback loops for continual improvement, and encourages collaboration with other stakeholders and end-users. Particularly in the fast-paced developmental environments of startups, the impacts and ethical considerations of AI technologies differ significantly across the developmental stages. The AIA recognizes this need to perform phase-specific ethics evaluations and explicitly states that the process should be conducted twice—once in the beginning stages of the algorithm's design, and again closer to the system's deployment. To better incentivize the resource-strapped startups, in particular, more AIEFs need to serve the *ethics issue monitoring* function in nimble ways that enable businesses to implement AI ethics into their operations.

This complicates the matters for startups who wish to operationalize AI ethics practices in the operation and culture of the company. Not only is a comprehensive discovery of AI ethics issues using process-oriented frameworks expensive, but doing both the issue discovery and monitoring throughout their product development pipeline can be costly. Given practical limitations typical of startups—namely, funding, personnel and time—startup incubators, venture capitalists, and innovation hubs have the opportunity to step-up and support startups by providing or funding AI ethics assessment activities. However, industry-wide adoption of such initiatives will likely be slow unless the business benefits of investing in AI ethics practices are made observable and articulated clearly.

In addition, our landscape analysis indicated the lack of and the need for industry or application-specific AIEFs. Our case study echoed this need in concrete terms within the Canadian mental health application context, where the results of AIEFs contradict the industry-specific regulatory requirements the client needs to meet. This, in turn, can prevent AIEF adoption as the certainty of legal ramifications overshadows ethics considerations. AIEFs then can serve as a tool for the AI industry to identify where these application-specific contradictions lie and help improve existing industry standards with characteristics of AI systems in mind.

5.2 Obstacles to adoption

Given the different values different styles of AIEFs can offer, it should be welcome news that there are so many published AIEFs in the world today. However, we found a number of key obstacles hindering the industry's adoption of AIEFs.

First, many AIEFs do not specify *who should be applying the frameworks*—or allude them to being appropriate for use by anyone. Our work reveals that not all AIEFs should be used by anyone. Rather, AIEFs that have a heavy emphasis on technical dimensions of AI cannot be applied by anyone outside of the AI system developers. Therefore, these AIEFs should make this limitation explicit. Over-emphasis on technical considerations can also signal that technical experts of the product should take ownership of the majority of the ethics assessment process. It can also incentivize the AI ethics community to create more technical solutions to AI ethics issues and overshadow the need for sociotechnical solutions. On the other hand, frameworks that refer to broad ethics standards may be difficult to implement for individuals without some background in ethics. For example, *The Ethics Canvas* (ADAPT Centre for Digital Content Technology) invites users to brainstorm ethical concerns through questions such as "What are the potential negative impacts of your product or service failing to operate or to be used as intended?" Answering such questions requires moral imagination beyond those taught in university engineering ethics classes.

Second, process-based frameworks such as FAIE are limited by how much resource/capacity the auditor has in engaging a diverse set of stakeholders. Our experiences in applying the frameworks suggest that most stakeholders could lead AI ethics assessments using process-based AIEFs as long as they are in a position to engage other stakeholders in the process. For instance, investigating risks stemming from the technical elements of the product naturally led us to seek input from the client's developer team, whereas concerns related to end-user experiences could only be addressed through engagement with potential users (physicians and patients in this use case). We found ourselves serving as facilitators between the stakeholders and AIEFs, gaining information from a diversity of sources, and translating them into language accessible to multiple stakeholder groups. However, such engagement and facilitation is a resource-intensive process that many startups cannot afford.

Lastly, despite the need and desire the industry might have to create consistent AI ethics assessment processes, existing AIEFs that invite deep reflections and moral imagination cannot be expected to yield consistent results across auditors. While the lack of consistency may steer companies away from adopting an AIEF from an audit perspective, it would be prudent for the AI ethics research community to identify clear business values of adopting AIEFs irrespective of their function as a possible auditing tool.

5.3 Limitations

As this work is a first attempt at conducting a comparative analysis of AIEFs through a case-study approach, our work involves a number of limitations. First, given that it is not feasible to apply all existing AIEFs to the same case study, our categorization of AIEFs into checklist-based and process-based frameworks is inherently limited to the four AIEFs we employed. While we feel confident that some characteristics of checklists and process-based frameworks can be generalized, these dichotomies likely do not represent

the diversity of published AIEFs and the unique benefits they provide to the industry.

Moreover, as ethical implications are case specific (Musschenga 2005), elements that were effective in this study may not be applicable to other application contexts. Researcher positionality was also unavoidable in our work. As we, the researchers, do not have clinical expertise, we chose not to interview the actual target demographic of the use case (i.e., patients with a diagnosed depressive disorder) upon consultation with the university research ethics board. Therefore, the data we collected from the potential patient-user interviews involved healthy adults who likely do not adequately express nuanced issues that the real potential users of the AI system may have been able to express.

Third, we are aware that the comparative analysis method we employed is limited to the subjective experience of our role as third-party auditors as well as the verbalized feedback the client was able to express. These were unavoidable practical limitations such as client time budget as well as the lack of established methods to compare AIEFs. While the former may vary from one client to another, we hope the latter can be solved through consolidated efforts by the AI ethics community in the near future.

6 Conclusion

In this work, we conducted a landscape analysis of the wide range of AI ethics frameworks to better understand the different types of frameworks that are publicly available. Assuming the role of a third-party auditor, we applied four different frameworks (two checklist-based, two process-based) to conduct an AI ethics risk assessment for a Canadian startup currently developing a mental health AI product.

While the different frameworks require a significantly different resource commitment in terms of time (1.5 vs. 20 h) and stakeholder engagement activities involved, they all provide a degree of feel-good factor helping the company brand itself as a responsible innovator. However, our results suggest that there is no "one size fits all" solution to adequately help discover and continue to monitor AI ethics issues: the close-ended, checklist-based frameworks (EU Guidelines for Trustworthy AI, the Government of Canada's Algorithmic Impact Assessment) are problematic if they are used to discover a comprehensive set of AI ethics issues; process-based frameworks (Foresight into AI Ethics, Ethics and Algorithms Toolkit) are too costly if they are used to regularly monitor ethics issues throughout an AI product development process. In addition, given the fact that no one AIEF outperforms all others, we call on the AI ethics community to better specify the suitability and expected benefits of existing frameworks for improved adoption of AI ethics assessments in the industry.

Acknowledgements We acknowledge the financial support of NSERC [Grant no. G13031], McGill University, and Arts Research Internship Awards (ARIA) by the Arts Internship Office of McGill University to conduct this study. The authors are grateful to all the participants who participated in the study.

Data availability statement The datasets generated during and/or analysed during the current study are not publicly available due to the condition of confidentiality with which the human participant data (e.g., interviews) was collected. However, the reports resulting from the AI ethics assessments we conducted are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Adadi A, Berrada M (2018) Peeking inside the Black-Box: a survey on Explainable Artificial Intelligence (XAI). IEEE Access 6:52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Adams RJ, Smart P, Huff AS (2017) Shades of grey: guidelines for working with the grey literature in systematic reviews for management and organizational studies: shades of grey. Int J Manag Rev 19:432–454. https://doi.org/10.1111/ijmr.12102

ADAPT Centre for Digital Content Technology The Ethics Canvas. https://www.ethicscanvas.org/canvas/index.php. Accessed 17 Mar 2022

Amershi S, Weld D, Vorvoreanu M et al (2019) Guidelines for human–AI interaction. In: Proceedings of the 2019 CHI conference on human factors in computing systems. Association for Computing Machinery, New York, NY, USA, pp 1–13

Anderson D, Bonaguro J, McKinney M et al (2018) Ethics & Algorithms Toolkit (beta). https://ethicstoolkit.ai/. Accessed 17 Mar 2022

Arya V, Bellamy RKE, Chen P-Y, et al (2019) One explanation does not fit all: a toolkit and taxonomy of ai explainability techniques. ArXiv190903012 Cs Stat

Ayling J, Chapman A (2021) Putting AI ethics to work: are the tools fit for purpose? AI Ethics. https://doi.org/10.1007/s43681-021-00084-x

Baxter P, Jack S (2008) Qualitative case study methodology: study design and implementation for novice researchers. Qual Rep 13:544–559. https://doi.org/10.46743/2160-3715/2008.1573

Bellamy RKE, Dey K, Hind M et al (2018) AI fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. ArXiv181001943 Cs

Benjamens S, Dhunnoo P, Meskó B (2020) The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. Npj Digit Med 3:1–8. https://doi.org/10.1038/s41746-020-00324-0

Blackman R (2020) A practical guide to building ethical AI. Harvard Business Review. https://hbr.org/2020/10/a-practical-guide-to-building-ethical-ai. Accessed 15 Oct 2021

Bogina V, Hartman A, Kuflik T, Shulner-Tal A (2021) Educating software and AI stakeholders about algorithmic fairness,

accountability, transparency and ethics. Int J Artif Intell Educ. https://doi.org/10.1007/s40593-021-00248-0
BSA (2021) Confronting bias: BSA's framework to build trust in AI. In: BSA Artif. Intell. BSA Artif. Intell. Policy Overv. https://ai.bsa.org/confronting-bias-bsas-framework-to-build-trust-in-ai/. Accessed 29 Aug 2022
CDC (2019) Health Insurance Portability and Accountability Act of 1996 (HIPAA) | CDC. https://www.cdc.gov/phlp/publications/topic/hipaa.html. Accessed 17 Mar 2022
Cohen IG, Mello MM (2018) HIPAA and protecting health information in the 21st century. JAMA 320:231–232. https://doi.org/10.1001/jama.2018.5630
de Swarte T, Boufous O, Escalle P (2019) Artificial intelligence, ethics and human values: the cases of military drones and companion robots. Artif Life Robot 24:291–296. https://doi.org/10.1007/s10015-019-00525-1
Etikan I (2016) Comparison of convenience sampling and purposive sampling. Am J Theor Appl Stat 5:1. https://doi.org/10.11648/j.ajtas.20160501.11
FDA (2021) Artificial intelligence and machine learning in software as a medical device. FDA
Floridi L, Cowls J (2022) A Unified Framework of Five Principles for AI in Society. In: Carta S (ed) Machine Learning and the City, Wiley & Sons, p 535–545. https://doi.org/10.1002/9781119815075.ch45
Floridi L, Cowls J, Beltrametti M et al (2018) AI4People—an ethical framework for a good AI Society: opportunities, risks, principles, and recommendations. Minds Mach 28:689–707. https://doi.org/10.1007/s11023-018-9482-5
Generation R Consulting (2018) Ethics analysis of predictive algorithms: an assessment report for technical safety BC. In: Tech. Saf. BC. https://www.technicalsafetybc.ca/ethics-analysis-predictive-algorithms-assessment-report-technical-safety-bc. Accessed 5 June 2022
Gerke S, Minssen T, Cohen G (2020) Ethical and legal challenges of artificial intelligence-driven healthcare. Artif Intell Healthc. https://doi.org/10.1016/B978-0-12-818438-7.00012-5
Goodwin GP, Darley JM (2010) The perceived objectivity of ethical beliefs: psychological findings and implications for public policy. Rev Philos Psychol 1:161–188. https://doi.org/10.1007/s13164-009-0013-4
Hagendorff T (2020) The ethics of AI ethics: an evaluation of guidelines. Minds Mach 30:99–120. https://doi.org/10.1007/s11023-020-09517-8
Harrison G, Hanson J, Jacinto C, et al (2020) An empirical study on the perceived fairness of realistic, imperfect machine learning models. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. ACM, Barcelona, Spain, pp 392–402
Health Canada (2004) Medical devices. https://www.canada.ca/en/health-canada/services/drugs-health-products/medical-devices.html. Accessed 30 Aug 2022
Health Canada (2021) Health Canada's action plan on medical devices: continuously improving safety, effectiveness and quality. https://www.canada.ca/en/health-canada/services/publications/drugs-health-products/medical-devices-action-plan.html. Accessed 5 June 2022
High-Level Expert Group on Artificial Intelligence (2019) Ethics guidelines for trustworthy AI
IBM (2020) Precision regulation for artificial intelligence. In: THINK-Policy Blog. https://www.ibm.com/blogs/policy/ai-precision-regulation/. Accessed 29 Aug 2022
IBM (2022) AI ethics. https://www.ibm.com/artificial-intelligence/ethics. Accessed 29 Aug 2022
Institute for the Future Ethical OS. https://ethicalos.org/. Accessed 5 June 2022
Jobin A, Ienca M, Vayena E (2019) Artificial Intelligence: the global landscape of ethics guidelines. Nat Mach Intell 1:389–399. https://doi.org/10.1038/s42256-019-0088-2
Katell M, Young M, Dailey D, et al (2020) Toward situated interventions for algorithmic equity: lessons from the field. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, NY, USA, pp 45–55
Knight CG (2001) Human–environment relationship: comparative case studies. In: Smelser NJ, Baltes PB (eds) International encyclopedia of the social and behavioral sciences. Pergamon, Oxford, pp 7039–7045
Madaio MA, Stark L, Wortman Vaughan J, Wallach H (2020) Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In: Proceedings of the 2020 CHI conference on human factors in computing systems. Association for Computing Machinery, New York, NY, USA, pp 1–14
Microsoft (2022) Artificial intelligence solutions and services. In: Microsoft AI. https://www.microsoft.com/en-us/ai. Accessed 17 Mar 2022
Montréal Declaration Responsible AI (2017) Montréal declaration for a responsible development of artificial intelligence. https://www.montrealdeclaration-responsibleai.com/. Accessed 31 Oct 2022
Morley J, Floridi L, Kinsey L, Elhalal A (2020) From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci Eng Ethics 26:2141–2168. https://doi.org/10.1007/s11948-019-00165-5
Musschenga AW (2005) Empirical ethics, context-sensitivity, and contextualism. J Med Philos Forum Bioeth Philos Med 30:467–490. https://doi.org/10.1080/03605310500253030
Noble S (2018) Algorithms of oppression: how search engines reinforce racism. New York University Press, New York
Office of the Privacy Commissioner of Canada (2018) PIPEDA in brief. https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/pipeda_brief/. Accessed 17 Mar 2022
Open Roboethics Institute (2019) Foresight into AI Ethics Toolkit
Peters D, Vold K, Robinson D, Calvo RA (2020) Responsible AI—two frameworks for ethical design practice. IEEE Trans Technol Soc 1:34–47. https://doi.org/10.1109/TTS.2020.2974991
Ribeiro MTC (2022) lime
Ribeiro MT, Singh S, Guestrin C (2016) "Why Should I Trust You?": Explaining the predictions of any classifier. arXiv
Santa Clara University An Ethical Toolkit for Engineering/Design Practice. https://www.scu.edu/ethics-in-technology-practice/ethical-toolkit/. Accessed 17 Mar 2022
Schiff D, Rakova B, Ayesh A et al (2021) Explaining the principles to practices gap in AI. IEEE Technol Soc Mag 40:81–94. https://doi.org/10.1109/MTS.2021.3056286
Theodorou A, Dignum V (2020) Towards ethical and socio-legal governance in AI. Nat Mach Intell 2:10–12. https://doi.org/10.1038/s42256-019-0136-y
Treasury Board of Canada (2021) Algorithmic Impact Assessment Tool. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html. Accessed 16 Mar 2022
Tricot R (2021) Venture capital investments in artificial intelligence: analysing trends in VC in AI companies from 2012 through 2020. OECD, Paris
Vakkuri V, Kemell K-K (2019) Implementing AI ethics in practice: an empirical evaluation of the RESOLVEDD strategy. In: Hyrynsalmi S, Suoranta M, Nguyen-Duc A (eds) Software business. Springer International Publishing, Cham, pp 260–275
Vakkuri V, Kemell K-K, Jantunen M, Abrahamsson P (2020a) "This is Just a Prototype": how ethics are ignored in software startup-like environments. In: Stray V, Hoda R, Paasivaara M, Kruchten P

(eds) agile processes in software engineering and extreme programming. Springer International Publishing, Cham, pp 195–210
Vakkuri V, Kemell K-K, Kultanen J, Abrahamsson P (2020b) The current state of industrial practice in artificial intelligence ethics. IEEE Softw 37:50–57. https://doi.org/10.1109/MS.2020.2985621
Wachter S, Mittelstadt B (2019) A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. Columbia Bus Law Rev 2019:494–620. https://doi.org/10.7916/cblr.v2019i2.3424
Whittlestone J, Nyrup R, Alexandrova A, Cave S (2019) The role and limits of principles in AI ethics: towards a focus on tensions. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and Society. Association for Computing Machinery, New York, NY, USA, pp 195–200
WHO (2021) Ethics and governance of artificial intelligence for health. https://www.who.int/publications-detail-redirect/9789240029200. Accessed 18 Mar 2022
Yin RK (2003) Case study research: design and methods, 3rd edn. Sage Publications, Thousand Oaks
Yin RK (2009) Case study research: design and methods. SAGE, Thousand Oaks
Zuloaga L (2021) Industry leadership: new audit results and decision on visual analysis. In: hirevue.com. https://www.hirevue.com/blog/hiring/industry-leadership-new-audit-results-and-decision-on-visual-analysis. Accessed 5 June 2022