AI & SOCIETY (2023) 38:411–423
https://doi.org/10.1007/s00146-021-01308-8

OPEN FORUM

Operationalising AI ethics: barriers, enablers and next steps

Jessica Morley[1] · Libby Kinsey[2] · Anat Elhalal[2] · Francesca Garcia[2] · Marta Ziosi[1] · Luciano Floridi[1,3]

Received: 31 May 2021 / Accepted: 13 October 2021 / Published online: 15 November 2021

Abstract
By mid-2019 there were more than 80 AI ethics guides available in the public domain. Despite this, 2020 saw numerous news stories break related to ethically questionable uses of AI. In part, this is because AI ethics theory remains highly abstract, and of limited practical applicability to those actually responsible for designing algorithms and AI systems. Our previous research sought to start closing this gap between the 'what' and the 'how' of AI ethics through the creation of a searchable typology of tools and methods designed to translate between the five most common AI ethics principles and implementable design practices. Whilst a useful starting point, that research rested on the assumption that *all* AI practitioners are aware of the ethical implications of AI, understand their importance, and are actively seeking to respond to them. In reality, it is unclear whether this is the case. It is this limitation that we seek to overcome here by conducting a mixed-methods qualitative analysis to answer the following four questions: what do AI practitioners understand about the need to translate ethical principles into practice? What motivates AI practitioners to embed ethical principles into design practices? What barriers do AI practitioners face when attempting to translate ethical principles into practice? And finally, what assistance do AI practitioners want and need when translating ethical principles into practice?

Keywords AI ethics · Applied ethics · Business ethics · Ethical practices · Ethical principles

1 Introduction

Since the very earliest days of Artificial Intelligence (AI) development researchers have been raising concerns about the ethical implications of its use in society (see e.g., Turing 1950; Wiener 1954). In recent years, as uses for AI solutions have multiplied and have started to impact people's everyday lives in tangible and significant ways, the conversation has moved out of its academic enclave and entered the consciousness of the public and policymakers (Barn 2019). The result has been a rapid proliferation of primarily principle-based ethics statements, frameworks, codes of conduct, and standards (henceforth ethics guides) from industry, academia and both national and supranational Governing bodies. Initially, the production of such documents was viewed as being the key to the creation of the right conditions for the ethical design, development and deployment of AI systems in society. However, over time, it has become increasingly clear that whilst these documents' existence might be necessary for the creation of these much-needed pro-ethical conditions (Floridi 2017), it is far from sufficient (Vidgen et al. 2020).

By mid-2019, there were more than 80 ethics guides available in the public domain (Jobin et al. 2019). Despite this, 2020 saw numerous news stories break related to ethically questionable uses of AI in healthcare (Villarreal 2020); education (Hern 2020); law enforcement (Council 2020); recruitment (Cheong et al. 2020); risk assessment (Guariglia and Tsukayama 2021); and more (Wiggers 2021). In part, the

✉ Jessica Morley
Jessica.morley@oii.ox.ac.uk

Libby Kinsey
libby.kinsey@gmail.com

Anat Elhalal
anat.elhalal@gmail.com

Francesca Garcia
francesca.garcia@digitalcatapult.org

Marta Ziosi
marta.ziosi@oii.ox.ac.uk

Luciano Floridi
luciano.floridi@oii.ox.ac.uk

[1] Oxford Internet Institute, University of Oxford, 1 St Giles', Oxford OX1 3JS, UK

[2] Digital Catapult, 101 Euston Rd, London NW1 2RA, UK

[3] Department of Legal Studies, University of Bologna, Bologna, Italy

limited impact of principle-based ethics guides can be attributed to the fact that there are so many different guides, but very few systems are in place to check for compliance with these guides. This has created a scenario in which those producing, purchasing, or using AI systems can 'ethics shop,' 'ethics wash,' and 'ethics dump' without fear of facing any real consequences (Floridi 2019). However, the impact has arguably been more significantly restricted by the fact that ethics guides remain highly abstract, and of limited practical applicability to those actually responsible for designing algorithms and AI systems. This abstraction tends to encourage a public opinion that there are *good* and *bad* algorithms, rather than *well designed* or *poorly designed* algorithms. The UK Prime Minister, Boris Johnson, for example described an algorithm used to predict A-level grades in August 2020 as a 'mutant algorithm' when it was shown to be highly discriminatory (Coughlan 2020), rather than commenting on the design decisions that led its ethically poor performance.

Portraying algorithms and by extension AI-systems as somehow objectively good or bad, ethical or unethical, incorrectly implies that AI-systems can act as independent moral agents over which human agents have little control (Coughlan 2020). This undermines the agency of AI developers, engineers, and designers (henceforth collectively practitioners) and slows progress from the 'what' of AI ethics (principles) to the 'how' (concrete design decisions) (Morley et al. 2020a, b). Challenges to this perspective have made clear that the impact (negative or positive) of an AI system is defined by the choices made during the design process (Fiore 2020), and practitioners have, therefore, been encouraged to accept their duty of care (van de Poel and Sand 2018) to those on the receiving end of actions made by AI systems (Floridi 2016).

AI practitioners cannot, however, be expected to take on this duty without any support. Individuals making seemingly neutral design decisions need assistance if they are to be expected to understand how these decisions might result in grossly different social or environmental outcomes (Floridi 2016). Such assistance needs to: (i) go beyond general guidelines for professional and ethical practice; (ii) embody a tool-set that does not require a deep background in philosophy; (iii) reflect the normative status of ethical reasoning; (iv) be practically applicable to real-world ethical decisions (Schwarz 2005; Vidgen et al. 2020); (v) enable developers to think through potential future scenarios (Floridi and Strait 2020); and (vi) ensure all stakeholders are engaged and involved in design decisions, rather than simply consulted about them (Durante 2014). The best way to provide assistance of this nature, and therefore implement the practice of AI ethics, remains unknown (Vakkuri et al. 2020).

Our previous research has attempted to start closing this knowledge gap through: the development and use of a question-based ethics framework; the creation of a searchable typology of tools and methods designed to translate between the five most common AI ethics principles (Floridi and Cowls 2019) and implementable design practices (Morley et al. 2020a, b; https://www.digicatapult.org.uk/for-start ups/other-programmes/applied-ai-ethics-typology); and the formalisation of the concept 'Ethics as a Service' (Morley et al. 2021). Whilst the findings from this previous research have proven useful, for the sake of simplicity we have rested on the assumption that *all* AI practitioners are aware of the ethical implications of AI, understand their importance, and are actively seeking to respond to them. In reality, that simplifying hypothesis was useful to develop (and did not undermine) other parts of our research but was always going to be challenged. For we do not know for certain that this is always the case. Nor do we know what the variables are that might affect awareness, understanding and corresponding action. In short, when it comes to the operationalisation of AI ethics, we lack information about the barriers and enablers, therefore we cannot know the best way to encourage widespread adoption of pro-ethical AI practices. It is this limitation that we seek to overcome in the following pages by answering these four questions:

(a) What do AI practitioners understand about the need to translate ethical principles into practice?
(b) What motivates AI practitioners to embed ethical principles into design practices?
(c) What barriers do AI practitioners face when attempting to translate ethical principles into practice?
(d) What assistance do AI practitioners want and need when translating ethical principles into practice?

Specifically, Sect. 2 describes our methodology; Sect. 3 outlines the results; Sect. 4 discusses the implications of the findings and makes some recommendations for making AI ethics more practical and implementable; Sect. 5 summarises the limitations of this research; and Sect. 6 concludes the discussion.

2 Methodology

When designing a programme of research, the choice of which method to employ is dependent upon the nature of the research problem (Noor 2008). If our intention was to develop an in-depth understanding of how one method for translating ethics principles into practice might become effective, then a case study approach would be most appropriate, as demonstrated by Vakkuri and Kemell (2019) in their review of the RESOLVEDD strategy for ethical decision-making, or Wong et al. (2018) in their evaluation of mobile Augmented Reality (AR) learning trails—Trails of Integrity and Ethics—in Hong Kong. Alternatively, a

scoping or systematic review approach could be appropriate, as Nicholls et al. (2015) highlighted with their review of empirical research assessing the quality and effectiveness of research ethics reviews. However, in this instance neither the systematic review nor the case study approach would produce results that were generalisable enough. Therefore, we employed a mixed-method qualitative approach designed to produce a high-level general description of the experiences of AI practitioners attempting to implement AI ethics practices by combining a survey with semi-structured interviews.

2.1 Survey

Survey participants were recruited using a snowball method (Babbie 2016) with the link to the survey being distributed across social media and via relevant mailing lists. In total, we collected 54 responses to the survey; 15 respondents were from start-ups, 6 from small-medium enterprises, 9 from large corporations, and 10 from primarily public sector organisations, covering a range of sectors including healthcare, retail, education, media and entertainment, finance, academia, life sciences, and Government. Although this is a relatively small sample size, preventing us from generating statistically significant results, it is a sufficient sample for generating descriptive results—as was our intention. Descriptive statistics were, therefore, used to analyse the results.

2.2 Interviews

Interview participants were recruited using a purposeful sampling method targeting specific individuals from start-ups (via the Digital Catapult's Machine Intelligence Garage), big-tech companies (via AI for People[1]), and the public sector. In total, six semi-structured interviews were conducted online via video conference. Each lasted between 20 and 60 min and covered a range of topics, including: general awareness of AI ethics; interpretation and understanding of AI ethics principles; experience of trying to build AI ethics into AI products; potential uses of translational tools and methods; problems faced when trying to operationalise the principles of AI ethics; perceived benefits of designing AI products pro-ethically; and perceived disadvantages of designing AI products pro-ethically. Thematic analysis was used to analyse the results.

2.3 Ethics

Ethics approval for this research was granted by the Oxford Internet Institute's departmental research ethics committee. Participants were not paid to be part of the study; interviews were not recorded; and notes were blinded. Survey data was collected using the Qualtrics platform which is GDPR-compliant, and no personal details were collected from the respondents.

[1] See https://www.aiforpeople.org/ for more information.

3 Results

Responding to questions in the survey, and in the interviews, about the need for AI ethics, and its operationalisation, participants highlighted a number of common themes. Here we summarise the key findings from this thematic analysis, before discussing the implications of these findings for AI ethics researchers and AI practitioners.

3.1 The theoretical importance of pro-ethical design is well recognised, but its definition remains narrow

When asked directly, 91% of survey respondents believe that designing AI products 'ethically' is very important for a number of reasons, including the positive impact pro-ethical design is perceived to have on consumer trust and satisfaction (43%), but mostly because pro-ethical design is perceived to improve social impact (46%) (Fig. 1). However, it is also clear that for many AI practitioners, this simply means they recognise the importance of being compliant with the tenets of data protection with the most commonly recognised principles being privacy and security (41%), and the least commonly recognised being autonomy and solidarity (7%) (Fig. 2). This is perhaps to be expected. Both interview and survey respondents highlighted difficulties with justifying the additional time and resource costs associated with 'pro-ethical' design, especially when there is no clear return on investment (Fig. 3). With data protection principles, companies have no choice but to accept these costs—and perhaps understand the potential reputation boost that comes from enhanced privacy control—yet this is not the case for other, more abstract, ethical principles which are more focused on 'doing good' rather than 'preventing harm.'

It is possible that this will change over time as consumers become more aware of potential ethical issues and use their purchasing power to persuade AI companies to take a broader range of ethical concerns seriously. Interviewees, for example, repeatedly defined ethical AI as 'AI that isn't biased,' reflecting the impact of media coverage of discriminatory algorithms, and several highlighted the intersection between bias and consumer trust. 82% of those who believe pro-ethical design incurs disadvantageous additional costs also recognise that pro-ethical design improves consumer trust and satisfaction, and 59% think that it improves customer loyalty (Fig. 4). It makes sense that, eventually, increased consumer trust leads to increased profit which

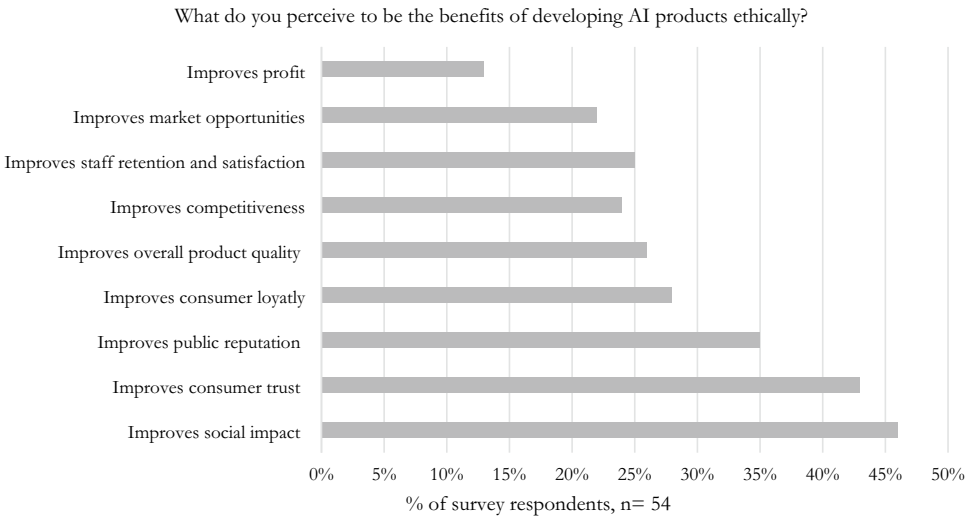

Fig. 1 Perceived benefits of pro-ethical AI design

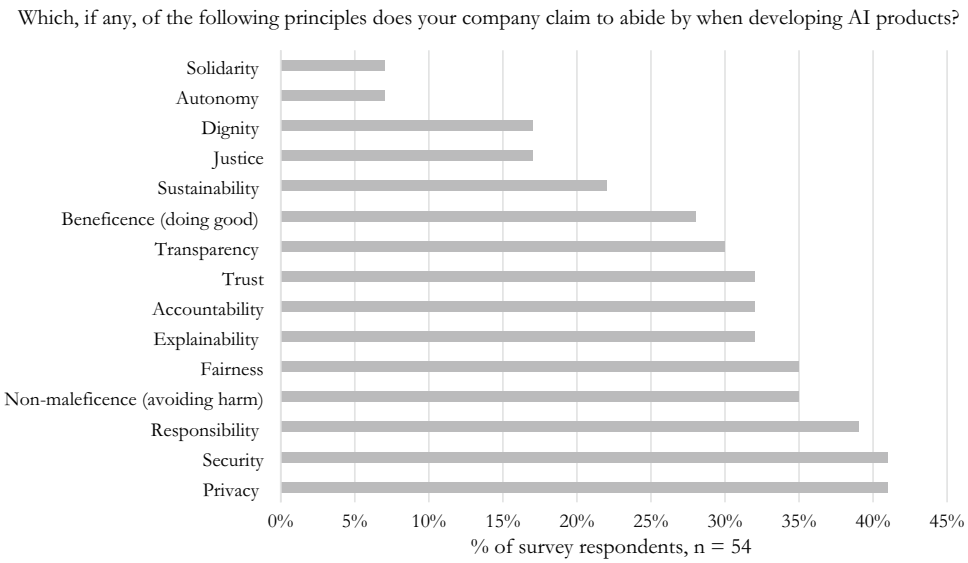

Fig. 2 Ethical principles used by companies to guide the design of AI products

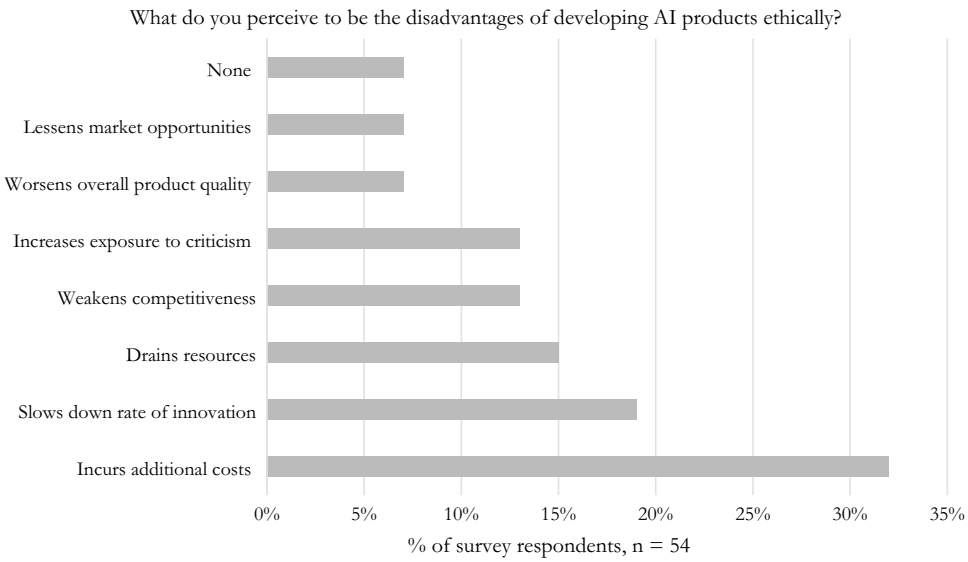

Fig. 3 Perceived disadvantages of pro-ethical design

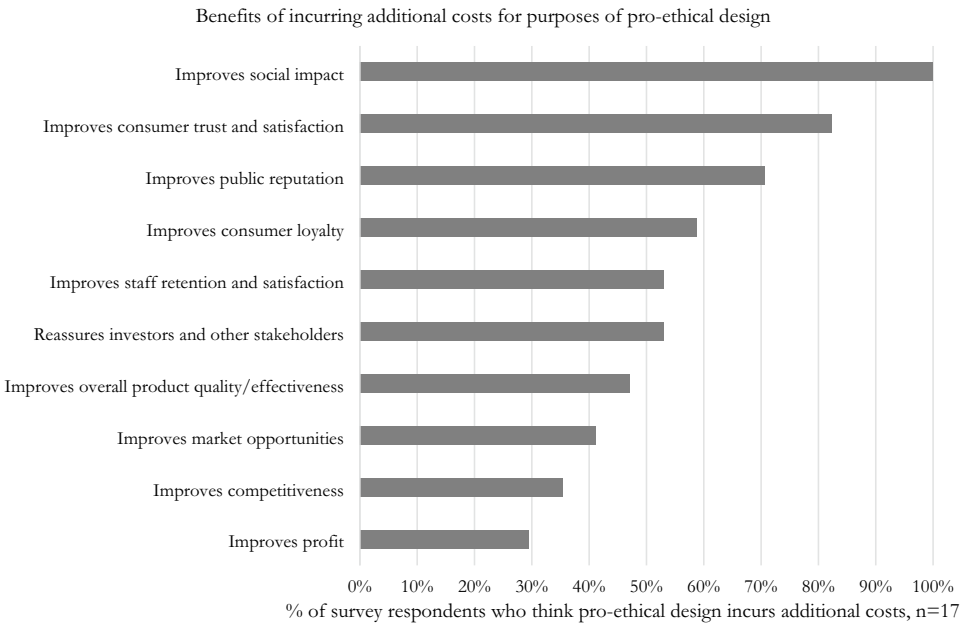

Fig. 4 Benefits of incurring additional cost for the purpose of pro-ethical design

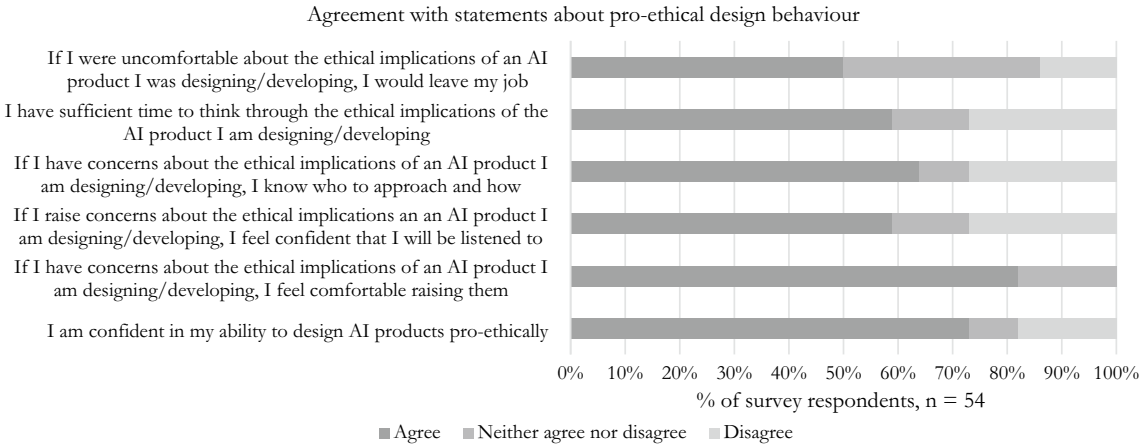

Fig. 5 Pro-ethical design behaviours

would better motivate investment in pro-ethical design. The challenge is in managing the lag between identifying ethical concerns and spreading the necessary level of public awareness needed to enact this cascade effect (Floridi 2018).

3.2 Lack of clarity over roles and responsibilities

Overall, AI practitioners are reasonably confident in their own abilities to design AI products pro-ethically when decisions sit with them (Fig. 5). On the surface, this is positive, though it may reflect the narrow understanding of AI ethics as outlined above. However, it is a lot less clear to practitioners *who* are ultimately responsible (and who should be held accountable) for ensuring alignment between product design and ethical principles (Fig. 6), and therefore who sanctions AI practitioners' pro-ethical actions.

Just under a quarter of survey respondents (24%) expressed a view that the responsibility should sit with all those involved in designing, developing or deploying AI systems. This is certainly an ideal view from the perspective of virtue ethics (Kitto and Knight 2019) and the idea that all AI practitioners should be developing responsibility-as-a-virtue (Rochel and Evéquoz 2020) so as to become conscious of the ethical implications of all their decisions. Yet caution is necessary. Interviewees also stressed that without the support of company senior leadership, or appropriate whistleblowing policies (Bolsin et al. 2005), individual practitioners feeling responsible does nothing other than leaving these individuals

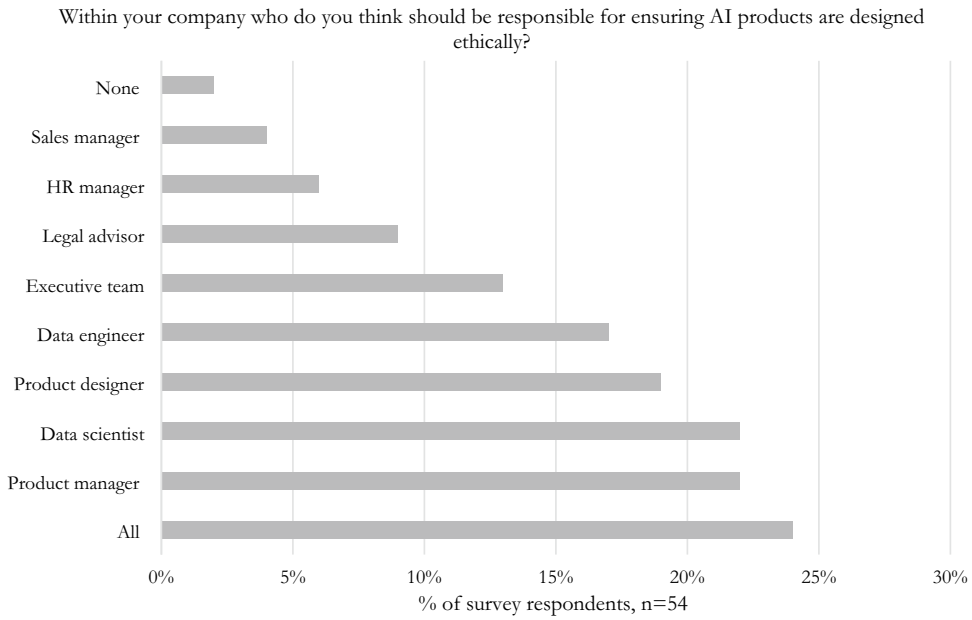

Fig. 6 Individuals responsible for ensuring pro-ethical design of AI products

vulnerable to retaliation (several interviewees cited the recent incident involving Timnit Gebru (Hao 2020)) and feeling burnt out. This is concerning as it is clear that these support mechanisms are frequently not in place. Just 13% of respondents think that the executive team of a company should hold a degree of responsibility for pro-ethical design. This most likely explains why—despite survey respondents saying that they (even though interview respondents said the opposite) feel relatively comfortable raising ethical concerns about products they are involved in designing or developing—less than two-thirds feel confident that they know *who* to approach if they had concerns, feel that they would be listened to, or feel as though they have sufficient time to think through the ethical implications of their decisions (Fig. 5). Unfortunately, cases like that of Timnit Gebru, show that individual workers from ethnic minorities are likely to be at greater risk of the types of retaliation interviewees were describing.

Unionisation could potentially change this. Although unions are not necessarily automatically a 'good' or pro-ethical force, indeed they can behave problematically, by giving workers protection in numbers, unionisation could help those closest to the design and impact-monitoring of AI products, raise concerns publicly and hold companies accountable without fear of ramification. The possibility of this being a lever for change is made especially apparent by the fact that 50% of survey respondents said they would leave their current employment if they had significant ethical concerns (Fig. 5) and 39% believe pro-ethical design practices aid staff satisfaction and retention (Fig. 5). Google's recently created workers' union could represent a watershed moment from this perspective (Koul and Shaw 2021). Alternative, or additional, potential options to unionisation are those falling under the umbrella heading of professionalisation, for example requiring AI practitioners to register with a professional body that holds them accountable for meeting certain ethical standards as is seen in medicine and other safety–critical sectors.

3.3 Disconnect between availability and demand for pro-ethical design resources

When it comes to the availability of pro-ethical design resources, or translational tools and methods, there is a clear disconnect between what is available to AI practitioners, and what they would find useful. In all instances, the proportion of survey respondents who have access to a specific type of resource (e.g., ethical framework) is far smaller than the proportion who would find that particular resource helpful when trying to pro-ethically design AI products (Fig. 7). This gap between 'supply and demand' is most noticeable for the more practical resources (Miller and Coldicott 2019). For instance, 13% of survey respondents indicated that they have used ethical design specifications, but 28% indicated they would find such a resource useful—a gap of fifteen percentage points. In contrast, a fifth of survey respondents indicated that they had access to a Code of Conduct, and 24% indicated they would find such a resource useful. This suggests that the AI ethics community is not yet meeting the needs of AI practitioners, despite the plethora of resources that have been produced (Morley et al. 2020a, b).

Interviewees were particularly keen to stress the limitations of high-level principle-based frameworks, commenting on the confusion caused by the sheer number of frameworks to choose from and the lack of clarity provided by principles. One of the interviews discussed how this issue had left

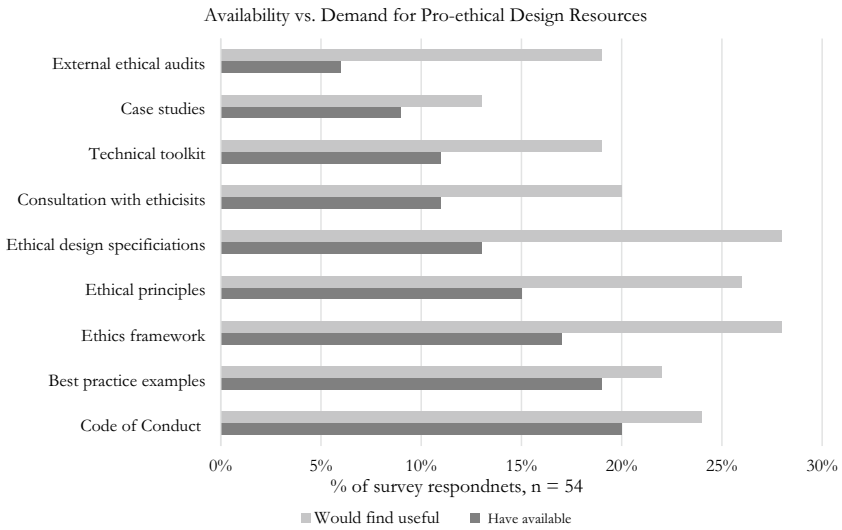

Fig. 7 Which of the following resources do you/have you used to help you develop/design AI products in an ethical manner? And which of the same resources would you find helpful when trying to develop/design an AI product ethically?

them feeling as though AI ethics is just 'tick-boxy' and 'buzzwordy', whilst several others stressed the fact that ethics, as a whole, as well as the specific principles, mean different things to different people in different contexts and yet it is rare for ethics frameworks to provide definitions or guidance on how to deal with these nuances. For this reason, several interviewees stated that they would prefer enforceable standards to be developed, provided these standards made ethics *relatable* and *actionable.*

3.4 Increasing need to implement external accountability mechanisms

The outlined limitations of current resources available to help AI practitioners translate ethical principles into design practices and the risk of ethics washing introduced by these limitations would, perhaps, be less concerning if there were more external scrutiny of company behaviours. Currently, the use of external accountability mechanisms such as ethics-based auditing (Mökander et al., 2021) is limited. Just 6% of survey respondents indicated that they have used auditing services for this purpose (Fig. 7). This is potentially because the idea is relatively new (Diakopoulos 2015), but uptake is also likely hampered by the fact that, without standardised approaches to pro-ethical design, it is not exactly clear *what* ethical-compliance would look like from an auditability perspective. (The data might be biased as 36% of respondents were from startups, less likely to have the necessary resources to engage in audits).

Insurance is an alternative mechanism for ensuring external scrutiny raised by the interviews. Several interviewees suggested that AI companies could be required to take out insurance against ethical harms which, in turn, may prompt the adoption of impact assessment tools as well as the regular use of ethical verification, validation and evaluation practices (Morley et al. 2020a, b). This has been the case with the introduction of insurance and inspection in other safety–critical areas, such as medical devices. One interviewee even went so far as to suggest a future in which having insurance against ethical harms was a prerequisite for operating as an AI development company. Whether or not this is the best suggestion, or even a feasible one, is beyond the scope of this discussion. What the suggestion does highlight is the growing need for the standardisation of 'pro-ethical design' and the development of readily implementable ethical practices. How the AI ethics community can respond to this call for help is the topic of the next section.

4 Next steps for the operationalisation of AI ethics

The results discussion demonstrates that AI practitioners have an abstract and relatively narrow understanding of ethical principles and how these can be translated into practice. There is recognition that pro-ethical design can improve social impact, particularly from the perspective of avoiding bias. However, for the most part this is interpreted from a risk-based approach as meaning avoiding harm—for example avoiding privacy infringement—rather than actively doing good either for society or for the environment. This suggests that AI practitioners are primarily motivated to translate ethical principles into design practices by the law, which currently seems to provide the only justification for investing additional resources into AI product design. This is a systematic issue since in a competitive environment no

actor can afford to bear such extra costs individually. This is problematic, however, as although legislation might make it easier for AI practitioners to justify additional spend to their shareholders, its existence would not negate the need for ethics. This is because whilst the law provides AI practitioners with answers to 'could' questions, it does not provide them with answers to 'should' questions. Or, to put it another way, the law provides AI practitioners with the rules of the game, it does not provide them with a strategy for 'winning.' In short, AI products that are merely legally compliant will not necessarily be ethically justifiable or socially acceptable. Furthermore, as briefly mentioned previously, new laws have a long lead-time, and they cannot be responsible to changes in social norms or attitudes, which may happen quite rapidly. Ethics will always be needed, therefore, to guide practitioners when they are operating in the 'grey areas.' Only when AI practitioners find themselves facing the same 'grey' issue repeatedly, and society finds itself suffering from the consequences of related poor-decisions, is it likely that legislation will be developed to provide AI practitioners with the reassurance they seek—this is known as the normative cascade (Floridi 2018).

It is, therefore, unrealistic to expect laws to be the 'solution' to all ethical dilemmas. It is also, potentially problematic, to treat pro-ethical design as a 'nice to have' rather than an essential—as is implied by AI practitioners waiting to justify the associated costs on legislation. Yet, other benefits, such as those related to public reputation and consumer loyalty, seem to motivate public declarations of compliance with principles, but do not yet provide sufficient motivation for altering design behaviours or practice in the absence of a clear return on investment or a request for legal compliance. At the same time, the motivation of individual AI practitioners who might be willing to effect change from the bottom-up is undermined by a lack of conceptual clarity about the meaning of ethical principles, and a lack of protection against retaliation from unsupportive senior stakeholders. As one interviewee said: 'we want to do the right thing, we just don't know what that is or how to do it.' This particular barrier could, perhaps, be lowered by greater use of translational tools and/or methods. However, there appears to still be limited uptake of those available in the public domain—or at least limited availability—and, therefore, uptake of the translational tools that practitioners would find most helpful. It seems that, as convincingly argued by Rességuier and Rodrigues (2020) there is a need to provide a mechanism other than the law that gives AI ethics 'teeth' (or makes pro-ethical design enforceable) to ensure its adoption.

If read from the perspective of 'what has not yet been achieved' these findings could be perceived as being demoralising. However, the findings are a reflection of the state of the industry/research field and only represent a generalisation from a few stances. They should not be seen as suggesting that no progress has been made in the field of pro-ethical AI, nor as a denial of the fact that there are examples of excellent pro-ethical practice. Instead, the findings should be seen as identifying the foundation from which we, as the pro-ethical AI community, can now build by taking a series of both macro actions focused on cultural change and micro actions focused on further developing existing translational tools—particularly ethics frameworks.

4.1 Encouraging a cultural shift

At a macro level, there is a need for actions that support a broader cultural shift in the realm of data-driven innovation. First, all AI practitioners (and indeed technologists in general) need to be encouraged to develop an understanding of the ethical implications of the products that they design by combining ethics theories (Kitto and Knight 2019) in mandatory courses provided to all data science, computer science, engineering, etc., trainees. Much can be learned here from the way that medical ethics is taught to all those in medical school (Concannon et al. 2019). The focus should be on practical ethics, such as the use of empathy exercises (Montonen et al. 2014), and the development of critical thinking skills that will help future AI practitioners develop the ability to fight against Hume's guillotine where 'is' is confused with 'ought' (Roff 2019).

Second, AI ethics researchers, in collaboration with journalists and public engagement specialists, should focus on making AI ethics *relatable*—both to AI practitioners and to the public. As we have discussed, highly abstract principles are potentially hindering rather than helping attempts to ensure AI products are developed pro-ethically. Bringing the principles down to a lower level of abstraction and focusing on more readily understandable questions such as 'is this the right solution for the problem?' 'is the solution working in the right way?' 'is the product having the right kind of impact?' can better support discussions about potential ethical implications (Hoffmann 1993).

Third, policymakers and legislators need to push against the false logic of the Collingridge dilemma (Genus and Stirling 2018). This is the idea that, when trying to govern emerging technologies, they face a double-bind problem of information and power whereby impacts cannot be easily predicted until a specific technology is extensively developed and widely used, but technology cannot be controlled or changed once it has become entrenched. Believing that their actions are restricted by this double bind allows policymakers and legislators to take an unjustified laissez-faire attitude to technological development and enables technology

companies to exert undue influence over both society and government. It is true that the impacts of technology—including AI technology—are hard to predict from the outset and that, therefore, inflexible legislation would be inappropriate and potentially even harmful. Nevertheless, some technological impacts remain foreseeable and when they are not, they hardly occur suddenly, they become increasingly visible gradually, and hence can be addressed at an earlier stage when correcting measures are more easily implementable. The European Commission has already recognised this, and recently proposed the first-ever legal framework for AI (European Commission 2021). Other nation states or supranational organisations may follow suit. Alternatively, other governing bodies may make the perfectly reasonable decision to take interim steps that encourage greater reflexivity throughout the whole accountability chain (Floridi 2016; Genus and Stirling 2018) and thus mitigate risk. Examples of such interim measures include: requiring AI companies to have an auditable whistleblowing policy and procedure in place; developing insurance policies against algorithmic harms; fostering the development of a more diverse tech workforce; and adapting copyright legislation appropriately so that more code can be shared openly for the purposes of error-checking, reuse and validation, without companies having to be concerned about the loss of intellectual property rights. Such interim actions can help foster a culture built on what Benrimoh et al. (2018) term Meticulous Transparency where everything about an AI product from the motivations for its development, its implementation, and its interaction with people and systems is recorded and made available for scrutiny.

4.2 Supporting the implementation and use of ethics frameworks

Macro-level changes will not happen overnight. For now, we must accept that the majority of AI practitioners do not have access to ethics frameworks. We cannot let this continue to be a reason for not being more proactive when it comes to ethical harms associated with AI. Instead, practical steps should be taken to enhance the utility of these frameworks, including our own (see Box 1).

Box 1: The digital catapult ethics framework (DCEF)

The DCEF was developed by the Digital Catapult's independent Ethics Committee following consultation with a number of Digital Ethicists and other experts. The framework consists of fourth levels and is intended to help AI start-ups working with the Digital Catapult to define and translate, transparently and contextually, high-level ethical principles into practice. The first level, therefore, consists of the five unifying high-level principles identified by Floridi et al. (2018): beneficence, non-maleficence, autonomy, justice, explicability. The second level consists of seven interpretations (or contextual definitions) of these principles identified through documentary analysis consultation with AI practitioners and those affected by AI systems. The third level operationalises Habermas's concept of discourse ethics (Buhmann et al. 2019), i.e. an approach that seeks to establish normative values and ethical truths through open discourse, and consists of a series of questions that are designed to encourage AI practitioners to conduct ethical foresight analysis (Floridi and Strait 2020). The fourth level provides access to more practical, and less discursive tools e.g. python libraries designed to identify bias in data. The connections between the levels are shown below. Companies using the DCEF to translate high-level ethical principles into practice are encouraged to consult it at validation, verification and evaluation stages of their product development pipeline, to ensure that at each stage time is dedicated to thinking through the ethical implications of all decisions made. This discussion is supported by members of the independent AI ethics committee through consultations which also provide a vehicle for reviewing the efficacy of the Framework itself.

L1	Beneficence: promoting well-being, preserving dignity, and sustaining the planet	Non-maleficence: privacy, security and 'capability caution.'	Autonomy: the power to decide (whether to decide)	Justice: promoting prosperity and preserving solidarity	Explicability: enabling the other principles through intelligibility and accountability
L2	Be clear about the benefits of the product or service Consider the business model	Know and manage the risks Use data responsibly	Be open and understandable in communications	Promote diversity, equality and inclusion	Be worthy of trust
L3	For example: What are the goals, purposes and intended applications of the product or service? Who or what might benefit from the product/service? Consider all potential groups of beneficiaries, whether individual users, groups or society and environment as a whole	For example Is the training data appropriate for the intended use? Have potential biases in the data been examined, well-understood and documented and is there a plan to mitigate against them?	For example: Does the company communicate clearly, honestly and directly about any potential risks of the product or service being provided? Are the company's policies relating to ethical principles available publicly and to employees? Are the processes to implement and update the policies open and transparent?	For example: Are there processes in place to establish whether the product or service might have a negative impact on the rights and liberties of individuals or groups? Does the company have a diversity and inclusiveness policy in relation to recruitment and retention of staff?	For example: Is there a process to review and assure the integrity of the AI system over time and take remedial action if it is not operating as intended? Does the company have a clear and easy to use system for third party/user or stakeholder concerns to be raised and handled?
L4	See: https://www.digicatapult.org.uk/for-startups/other-programmes/applied-ai-ethics-typology				

From the perspective of implementation, more should be done to match the 'tasks' associated with ethics frameworks (including discussing answers to open questions posed by frameworks) to normal stages in the workflow of software development. As one interviewee described it, thinking of ethics as 'something else to do' increases the cognitive load and resource burden too much, but if ethical considerations become standard aspects of software verification, validation of evaluation processes and prompts to ask '*x*' question at '*y*' stage are embedded in software project management tools such as Jira,[2] this will be less of a barrier. In short, we should move to a development pipeline where 'ethics review' becomes as much a standard part of the workflow as 'code review' and the use of ethics frameworks is perceived to be as essential as code review checklists. One way to do this is to tie more closely ethical consideration with user research and expand the concept of the latter so that the focus of UX and UI is not simply on user need but user impact. Additionally, tools like the consequence scanner developed by DotEveryone can help AI practitioners think through the impacts of the systems they design (DotEveryone).

From the perspective of utility, until AI practitioners have gained more experience considering ethical implications and translating these considerations into pro-ethical design decisions, they will require more specific guidance on 'what good looks like.' Eventually, this may come in the form of standards—something which some AI practitioners are waiting for—but in the meantime, this guidance can be provided by detailed case studies of where excellent pro-ethical design has been achieved (Kitto and Knight 2019) and the impact that this had on product success. A request for the latter was also made by one of the interviewees, who voiced the need to be confronted with such excellent cases as a yardstick to judge one's own ethical performance. These will be especially impactful if developed in collaboration with the responsible AI practitioners themselves, and those affected by the relevant AI system. These may seem like small suggestions, but if acted upon they have the potential to make a relatively big impact in the near future.

This does not, however, mean to imply that once AI Ethics Frameworks—and methods for translating the principles into practice—are unilaterally available, then the 'problem' of AI Ethics is 'solved.' To imply this would be to subsequently imply that it is possible to come to an unambitious technical solution to any potential area of ethical controversy. This is simply not possible, ethics, and by extension ethics frameworks, are not hard rules that, if followed, will always result in the 'right' outcome. There is no 'one' way to be ethical—even if

[2] Jira is a software development tool developed by Atlassian designed to help software developers plan, track and manage agile software development.

AI practitioners might like this to be the case. Instead, applying the concepts of pro-ethical design requires recognition of the fact that the 'most ethical' solution depends entirely on the wider socio-cultural context. To take the risks associated with discriminatory or 'biased' algorithms, demonstrated by the highly popular and oft-quoted ProPublica investigation into algorithms used to calculate recidivism risk (Angwin et al. 2016). One way for minimising the potential for algorithms to be discriminatory is to ensure no protected characteristics, for example, race, are included in the datasets used to train the algorithm. This might work, for example, when designing algorithms that analyse a person's creditworthiness but, in other contexts—notably medical contexts—it might be essential that such characteristics *are* included to ensure the accuracy of the algorithm and so avoid harm. Similarly, there may be instances in which it is necessary to prioritise the protection of 'one' right or ethical principle above the others. For instance, when protecting public health, it might be necessary to prioritise justice at a population level over individual-level autonomy, but in a different context doing this might be highly unethical. Finally, in a democratic society different people might reasonably disagree over the different values that should be embedded in different algorithms in different contexts—we must allow for value pluralism (Binns, 2018). Thus, AI ethics frameworks—no matter how much they come to be relied upon—must always be seen as guardrails, designed to stop AI practitioners from crossing social red-lines but not specifying exactly what to do to do this in each individual instance. Instead, AI practitioners should take an approach, inspired by Habermas's discourse ethics where the aim of AI ethics frameworks is to guide open discussions in which all sides of an argument are listened to and considered until a decision that is acceptable to all can be reached (Morley et al. 2020a, b). It is the discussion, and the process followed to ensure this discussion is held in an open, transparent, and 'fair' way, that is important. Both Whittlestone et al. (2019) and Terzis (2020) discuss the complexities of encountering tensions in AI ethics, in more detail.

5 Limitations

All research has limitations, and this is no exception. The relatively small number of research participants means that the results cannot be taken to be statistically significant and must be assumed to represent only a part of the overall AI ethics landscape. This is especially true as the interview participants were not representative of all industries in which AI is being developed, nor of all roles that an AI practitioner might take on when an AI product is being developed, and were all UK based so may not represent the experience of AI practitioners working elsewhere. In addition, as the interviews revealed a much greater lack of clarity regarding the meaning of 'AI ethics'—especially ethical principles—than expected, it is possible that both interviewees and survey respondents did not fully understand the questions being asked and therefore provided answers they thought that we would want. This might undermine the internal validity of some of the questions. Finally, we did not explicitly recruit interview or survey participants to ensure they were representative of the diversity of the AI practitioner workforce. This means that, although we do mention potential issues surrounding vulnerable minorities, we were unable to delve into the complexities associated with discrimination in the workplace and the expectation often placed on individuals from ethnic minorities to act as the 'moral conscience' of the company that they are working for. This should be a topic of further research. We have taken steps to minimise the impact of these limitations on the discussion, augmenting our findings with existing literature and our own experiences as AI practitioners, policymakers and ethicists and we believe the results are still useful for prompting further discussion. However, these limitations should not be seen to undermine the value of this research. Rather, they mean that the results should be seen as those from a pilot study, providing a starting point for an important conversation, and a point from which future research can build by asking similar questions in different settings with a more diverse range of participants.

6 Conclusion

Identifying the optimum mechanisms for implementing the tenets of AI ethics will take time and is likely to be an ongoing task that will require regular reflection as both the field of AI ethics and AI technology itself develop. In other words, the implementation of AI ethics should be underpinned by a learning governance model where regular reflection on impact is embedded in the research and decision-making cycle and overseen by those most affected by AI products and yet excluded from the development pipeline—lay members of the public, especially those from traditionally marginalised groups (Banner 2020). If those involved in aiding the development of pro-ethical AI design know that 'solutions' put in place are only for now, they are more likely to act sooner and more boldly which, ultimately, will ensure that when it comes to practical AI ethics, 'perfect' does not become the enemy of the good. Embedding this reflexivity into the AI ecosystem will require the AI ethics community, policymakers, and AI practitioners to collaboratively consider how best to monitor the progress of pro-ethical design,

evaluate processes put in place and methods used to enable pro-ethical design practices, and iterate based on feedback. We hope that our research can play a small, but significant, role in this process.

Acknowledgements We would like to thank the anonymous reviewer for their incredibly valuable comments, they added significant depth to the discussion and the paper is much better as a result. LK, AE, and JM conceived the paper. JM wrote the survey, designed the interviews and drafted the paper. MZ helped analyse the interview results, and helped recruit participants for both the survey and the interviews. FG helped draft the survey questions and helped recruit participants. All authors edited and reviewed the paper. LF oversaw the research and contributed to the paper.

Funding JMs work for this paper was funded by the Digital Catapult. JM has also received research funding, unrelated to this project from the Wellcome Trust, Google and Vodafone.

Availability of data and material Results from the survey may be available upon request from the corresponding author.

Code availability Not applicable.

Declarations

Conflict of interest JMs work for this paper was funded by the Digital Catapult. LK, FG and EA were all employed by the Digital Catapult at the time of writing. LF was the chair of the Digital Catapult Ethics Advisory Board.

Ethical approval Ethics approval was provided by the Departmental Research Ethics Committee of the Oxford Internet Institute.

References

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine Bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Babbie ER (2016) *The practice of social research* (Fourteenth). Cengage Learning

Banner N (2020) A new approach to decisions about data. Understanding Patient Data. https://understandingpatientdata.org.uk/news/new-approach-decisions-about-data

Barn BS (2019) Mapping the public debate on ethical concerns: Algorithms in mainstream media. J Inf Commun Ethics Soc 18(1):38–53. https://doi.org/10.1108/JICES-04-2019-0039

Benrimoh D, Israel S, Perlman K, Fratila R, Krause M (2018) Meticulous transparency—An evaluation process for an agile AI regulatory scheme: vol 10868 LNAI. Scopus, p 880. https://doi.org/10.1007/978-3-319-92058-0_83

Binns R (2018) Algorithmic accountability and public reason. Philos Technol 31(4):543–556. https://doi.org/10.1007/s13347-017-0263-5

Bolsin SN, Faunce T, Oakley J (2005) Practical virtue ethics: healthcare whistleblowing and portable digital technology. J Med Ethics 31(10):612–618. https://doi.org/10.1136/jme.2004.010603

Buhmann A, Paßmann J, Fieseler C (2019) Managing Algorithmic Accountability: Balancing Reputational Concerns, Engagement Strategies, and the Potential of Rational Discourse. J Bus Ethics. https://doi.org/10.1007/s10551-019-04226-4

Cheong M, Lederman R, McLoughney A, Njoto S, Wirth A (2020) Ethical implications of AI bias as a result of workforce gender imbalance. University of Melbourne. https://about.unimelb.edu.au/__data/assets/pdf_file/0024/186252/NEW-RESEARCH-REPORT-Ethical-Implications-of-AI-Bias-as-a-Result-of-Workforce-Gender-Imbalance-UniMelb,-UniBank.pdf

Concannon M, Gillibrand W, Jones P (2019) An exploration of how ethics informs health care practice. Ethics Med 35(1):27–42

Coughlan S (2020) A-levels and GCSEs: Boris Johnson blames 'mutant algorithm' for exam fiasco. BBC News. https://www.bbc.co.uk/news/education-53923279

Council J (2020) Facial recognition companies commit to police market after Amazon, Microsoft Exit. Wall Street J. https://www.wsj.com/articles/facial-recognition-companies-commit-to-police-market-after-amazon-microsoft-exit-11591997320

Diakopoulos N (2015) Algorithmic accountability: journalistic investigation of computational power structures. Digit J 3(3):398–415. https://doi.org/10.1080/21670811.2014.976411

DotEveryone (n.d.) The DotEveryone consequence scanning agile event. https://doteveryone.org.uk/project/consequence-scanning/

Durante M (2014) The democratic governance of information societies. A critique to the theory of stakeholders, vol 28

European Commission (2021) Proposal for a Regulation laying down harmonised rules on artificial intelligence. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

Fiore E (2020) Ethics of technology and design ethics in socio-technical systems investigating the role of the designer. FormAkademisk. https://doi.org/10.7577/formakademisk.2201

Floridi L (2016) Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions. Philos Trans R Soc a: Math Phys Eng Sci 374(2083):20160112. https://doi.org/10.1098/rsta.2016.0112

Floridi L (2017) The logic of design as a conceptual logic of information. Mind Mach 27(3):495–519. https://doi.org/10.1007/s11023-017-9438-1

Floridi L (2018) Soft ethics, the governance of the digital and the General Data Protection Regulation. Philos Trans Ser A Math Phys Eng Sci. https://doi.org/10.1098/rsta.2018.0081

Floridi L (2019) Translating principles into practices of digital ethics: five risks of being unethical. Philos Technol. https://doi.org/10.1007/s13347-019-00354-x

Floridi L, Cowls J (2019) A unified framework of five principles for AI in society. Harvard Data Sci Rev. https://doi.org/10.1162/99608f92.8cd550d1

Floridi L, Strait A (2020) Ethical foresight analysis: what it is and why it is needed? Mind Mach 30(1):77–97. https://doi.org/10.1007/s11023-020-09521-y

Genus A, Stirling A (2018) Collingridge and the dilemma of control: towards responsible and accountable innovation. Res Policy 47(1):61–69. https://doi.org/10.1016/j.respol.2017.09.012

Guariglia M, Tsukayama H (2021) Questions remain about pretrial risk-assessment algorithms: year in review 2020. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2020/12/questions-remain-about-pretrial-risk-assessment-algorithms-year-review-2020

Hao K (2020) We read the paper that forced Timnit Gebru out of Google. Here's what it says. MIT Technology Review. https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

Hern A (2020) Ofqual's A-level algorithm: why did it fail to make the grade? https://www.theguardian.com/education/2020/aug/21/ofqual-exams-algorithm-why-did-it-fail-make-grade-a-levels

Hoffmann DE (1993) Evaluating ethics committees: a view from the outside. Milbank Quart 71(4):677–701. https://doi.org/10.2307/3350425

Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1(9):389–399. https://doi.org/10.1038/s42256-019-0088-2

Kitto K, Knight S (2019) Practical ethics for building learning analytics. Br J Educ Technol 50(6):2855–2870. https://doi.org/10.1111/bjet.12868

Koul P, Shaw C (2021) We built Google. This is not the company we want to work for. The New York Times. https://www.nytimes.com/2021/01/04/opinion/google-union.html

Miller C, Coldicott R (2019) People, power and technology: the tech workers' view. Doteveryone. https://doteveryone.org.uk/report/workersview/

Mökander J, Floridi L (2021) Ethics-Based Auditing to Develop Trustworthy AI. Minds Mach 31(2):323–327. https://doi.org/10.1007/s11023-021-09557-8

Montonen T, Eriksson P, Asikainen I, Lehtimäki H (2014) Innovation empathy: a framework for customer-oriented lean innovation. Int J Entrep Innov Manag 18(5/6):368. https://doi.org/10.1504/IJEIM.2014.064719

Morley J, Cowls J, Taddeo M, Floridi L (2020a) Ethical guidelines for COVID-19 tracing apps. Nature 582(7810):29–31. https://doi.org/10.1038/d41586-020-01578-0

Morley J, Floridi L, Kinsey L, Elhalal A (2020b) 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(4):2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Morley J, Elhalal A, Garcia F, Kinsey L, Mökander J, Floridi L (2021) Ethics as a service: a pragmatic operationalisation of AI ethics. Mind Mach 31(2):239–256. https://doi.org/10.1007/s11023-021-09563-w

Nicholls SG, Hayes TP, Brehaut JC, McDonald M, Weijer C, Saginur R, Fergusson D (2015) A scoping review of empirical research relating to quality and effectiveness of research ethics review. PLoS ONE. https://doi.org/10.1371/journal.pone.0133639

Noor KBM (2008) Case study: a strategic research methodology. Am J Appl Sci 5(11):1602–1604. https://doi.org/10.3844/ajassp.2008.1602.1604

Rességuier A, Rodrigues R (2020) AI ethics should not remain toothless! A call to bring back the teeth of ethics. Big Data Soc 7(2):205395172094254. https://doi.org/10.1177/2053951720942541

Rochel J, Evéquoz F (2020) Getting into the engine room: A blueprint to investigate the shadowy steps of AI ethics. AI Soc. https://doi.org/10.1007/s00146-020-01069-w

Roff HM (2019) Artificial intelligence: power to the people. Ethics Int Aff 33(2):127–140. https://doi.org/10.1017/S0892679419000121

Schwarz TSJ (2005) Teaching ethics and computer forensics: the Markkula center for applied ethics approach. In: Proceedings of the 2nd annual conference on information security curriculum development—InfoSecCD '05, 66. https://doi.org/10.1145/1107622.1107637

Terzis P (2020) Onward for the freedom of others: marching beyond the AI ethics. Scopus, pp 220–229. https://doi.org/10.1145/3351095.3373152

Turing AM (1950) Computing machinery and intelligence. Mind 59(236):433–460

Vakkuri V, Kemell K-K (2019) Implementing AI ethics in practice: an empirical evaluation of the RESOLVEDD strategy: Vol. 370 LNBIP. Scopus, p 275. https://doi.org/10.1007/978-3-030-33742-1_21

Vakkuri V, Kemell K-K, Jantunen M, Abrahamsson P (2020) "This is Just a Prototype": how ethics are ignored in Software Startup-like environments: vol 383 LNBIP. Scopus, p 210. https://doi.org/10.1007/978-3-030-49392-9_13

van de Poel I, Sand M (2018) Varieties of responsibility: two problems of responsible innovation. Synthese. https://doi.org/10.1007/s11229-018-01951-7

Vidgen R, Hindle G, Randolph I (2020) Exploring the ethical implications of business analytics with a business ethics canvas. Eur J Oper Res 281(3):491–501. https://doi.org/10.1016/j.ejor.2019.04.036

Villarreal A (2020) US healthcare workers protest chaos in hospitals' vaccine rollout. The Guardian. https://www.theguardian.com/world/2020/dec/21/us-healthcare-workers-protest-chaos-hospitals-vaccines-vaccinations

Whittlestone J, Nyrup R, Alexandrova A, Cave S (2019) The role and limits of principles in AI ethics: towards a focus on tensions. https://doi.org/10.17863/cam.37097

Wiener N (1954) The human use of human beings: cybernetics and society (Revised). London

Wiggers K (2021) Outlandish Stanford facial recognition study claims there are links between facial features and political orientation. Venture Beat. https://venturebeat.com/2021/01/11/outlandish-stanford-facial-recognition-study-claims-there-are-links-between-facial-features-and-political-orientation/

Wong EYW, Kwong T, Pegrum M (2018) Learning on mobile augmented reality trails of integrity and ethics. Res Pract Technol Enhanc Learn 13(1):22. https://doi.org/10.1186/s41039-018-0088-6