AI and Ethics
https://doi.org/10.1007/s43681-025-00812-7

RESEARCH

The 3D model of ethical AI practice

Brian Ball[1] · **Alice C. Helliwell**[1]

Received: 13 January 2025 / Accepted: 29 July 2025

Abstract
In recent years, there have been growing calls to operationalize artificial intelligence (AI) ethics - to move from theory to practice, or (as one group of authors has put it) 'from what to how' (Morley et al. Sci Eng Ethics 26(4):2141–2168, 2020. https://doi.org/10.1007/s11948-019-00165-5). In this paper, we propose a novel account of what ethical AI practice might look like, which we call the 3D model, named for its recognition, within the overall AI design cycle, of the three stages of design, development, and deployment. This model aims to embed ethics throughout this cycle, offering questions that should be addressed at each stage. We articulate the benefits of this approach to ethical AI practice: that it is pro-ethical and value-aware, amenable to implementation, it embeds ethics at every stage of the development process, it embeds a culture and language of ethics in organizations and provides clear decision points. Our model is not a panacea, of course, and we accordingly provide an indication of the context in which the implementation of our model might be most effective in ensuring ethical AI practice.

Keywords AI Ethics · Practice · Operationalisation · AI design · Philosophy

1 The problem

It has been nearly three quarters of a century since Turing published his [46] paper on 'Computing Machinery and Intelligence'; and shortly after that the term 'artificial intelligence' (AI) was coined. During this time, AI has gone through periods of boom and bust, summers and winters [53].[1] The last quarter century in particular, however, has seen significant scientific progress in AI, as well as its widespread commercial adoption,[2] with the current boom being the result of 'deep learning'[3] [25][4]) on 'big data'[5] using new hardware—in particular, graphics processing units (GPUs).[6]

[1] Wooldridge identifies: the first AI boom, 'from about 1956 to 1974' [53] p. 47),then 'the AI winter' (p. 88), which occupied 'the early 1970s to the early 1980s' (p. 88); and 'agent-based' (p. 138) approaches were dominant by the end of the 1990s (p. 162).

[2] According to Wooldridge [53], following Google's purchase of DeepMind in January 2014, '[a]rtificial intelligence was suddenly big news—and big business' (p. 167). This was brought on by 'the twenty-first-century machine learning revolution' (p. 168).

[3] 'The rise of deep learning dates to 2012,' says Pasquinelli, 'when the convolutional neural network AlexNet won the ImageNet computer vision competition.' (2023, p. 14) This is corroborated by [25]: 'the convolutional neural network (ConvNet)… achieved many practical successes during the period when neural networks were out of favour' (p. 439) yet, 'ConvNets were largely forsaken by the mainstream computer-vision and machine-learning communities until the ImageNet competition in 2012' (p. 440).

[4] LeCun et al. [25] found that RNNs were best for sequential (e.g. text) data, and Vaswani et al. [49] introduced the Transformer architecture 'based solely on attention mechanisms' (abstract). It outperformed previous systems on machine translation tasks and paved the way for e.g. OpenAI's 2022 generative AI system ChatGPT.

[5] Wiggins & Jones say that in 2011 'danah boyd and Kate Crawford argued "The era of Big Data has begun."' ([51], p. 5) Google trends suggests an increase in the use of the term 'big data' in the early 2010s. The academic journal *Big Data and Society* published its first issue in 2014.

[6] 'In the late 1990s,' wrote the 2018 Turing Award joint winners 'neural net[work]s… were largely forsaken by the machine-learning community' [25], p. 441). But '[i]nterest in deep feedforward networks was revived around 2006' (p. 439) by researchers who used unsupervised learning for pre-training, with subsequent fine-tuning using backpropagation.

✉ Brian Ball
brian.ball@nulondon.ac.uk

✉ Alice C. Helliwell
alice.helliwell@nulondon.ac.uk

[1] Northeastern University London, London, UK

Published online: 27 August 2025

These developments have led, however, to increasing concern over the ethics of building and using the new data-hungry AI systems. 'In December 2014,' note Wiggins and Jones, Microsoft researcher Hanna Wallach gave a conference talk[7] to computer scientists and machine learning specialists in which 'she proposed that her own field desperately needed to interrogate how the algorithms they were developing, and the technologies the algorithms empowered, challenged our values' [51, p. 3] Others raised related worries[8]; and people took notice. 'Within a few years of the surge of critical concern, the likes of Google, Facebook and IBM all had in-house ethicists.' [51], p. 7). Academia also responded to the growing issue: for example, the journal *AI and Ethics* was founded, and published its first issue in February 2021.

Sadly, the ethical concerns raised by AI development and deployment have not been adequately addressed by these measures. Concretely speaking, AI ethicist Timnit Gebru left Google in December 2020, maintaining she had been fired (Google said she resigned); and Margaret Mitchell was then fired in early 2021 ([51], pp. 233–234). Similar developments occurred in other companies.[9] More abstractly, there has been a growing sense that much of the research done in AI ethics has been too… well, abstract. Theoretical. Impractical. One author has even gone so far as to declare on such grounds '[t]he uselessness of AI ethics' [33].

Perhaps less dramatically, through a study of work undertaken in the previous five years, Jobin and colleagues [22] uncovered 'a global convergence emerging around five ethical principles (transparency, justice and fairness, non-maleficence, responsibility and privacy), with substantive divergence in relation to how these principles are interpreted, why they are deemed important, what issue, domain or actors they pertain to, and how they should be implemented' (abstract). The key point here, as we see it, is that even if there is agreement on principles,[10] there remains unclarity on how those principles can be put into practice. In the words of Morley and colleagues, when it comes to AI ethics, we need to move 'from what to how' ([30], p. 2141). Many others have now made similar diagnoses and calls to action.[11]

We see our contribution here as responding to such calls (we are certainly more sympathetic with the thought that there is a need to supplement ethical principles in the domain of AI than with any demand to reject, repudiate, or replace them). We will not, however, be straightforwardly concerned with interpreting ethical principles in operational terms or translating them into practical guidelines. Rather, we see ourselves as offering a model—the 3D model—of ethical AI practice that we believe will result in the implementation of the relevant principles, if adhered to and applied within an appropriate setting. This is not to say that there is no role for principles in AI ethics. The full operationalization of AI ethics is undertaken by senior figures in corporate governance with a responsibility for ethics—in the manner described by Canca [6]. In this case, fundamental principles will play an important role. But this is about the (governance) context in which developers using the 3D model operate (we discuss this further in §3). However, whilst fully thinking through AI ethics does require working from first principles, we can make headway on implementing ethical design without working from such a high level. In the next section we outline the 3D model and explain its envisioned benefits.[12] Then, in the third and final section, we articulate the way in which we envision the model being embedded in a rich cultural, governance, and regulatory context, thereby responding to objections that might otherwise be levied against it. But we conclude this introductory section with some further motivation for the model through the introduction of a recent case of ethical AI failure, which will guide our discussion throughout.

Let's introduce a case, which will help us throughout this paper to illustrate the potential utility of our model. In recent months, there has been much concern regarding Microsoft's new AI 'Recall' on Windows 11. The system was designed to help to recover or recollect previous

[7] A near transcript of her talk is available at Wallach [50].

[8] 'In 2015 it was discovered that the AI [underpinning AlexNet] mislabels images featuring people with dark skin as "Gorilla" As a solution, Google + removes the Gorilla label entirely. As of 2023, the label is still removed from the database' according to Danielle Williams: https://www.daniellejwilliams.com/_files/ugd/a6ff55_001b0152f3c5448db2d0de3859cad73a.pdf.

[9] For example, see Knight [27].

[10] It is not clear that there is. Writing at roughly the same time, Floridi and Cowls also articulated 'five core principles for ethical AI' ([16], p. 2) in the landscape, although their list is slightly different. It includes 'beneficence, non-maleficence, autonomy,… justice… [and] explicability' ([16], p. 2) with the last of these encompassing both intelligibility (and so, one might think, transparency) and responsibility. A recent survey with international reach [9], found convergence around: 1) transparency/explainability/auditability, 2) reliability/safety/security/trustworthiness 3) justice/equity/fairness/non-discrimination, 4) privacy and 5) accountability/liability. For what it is worth, we prefer Canca's list of three core principles: 'respect autonomy, do good, ensure justice' ([6], p. 19). For as Canca notes, these principles of respect for autonomy, beneficence/non-maleficence, and justice concern intrinsic human values, while privacy, transparency—and we would add, responsibility—are about the means to these ends. Focussing first on values and then on the means to achieving these in the context of AI may be helpful to ensure responsiveness of principles to future technological developments.

[11] See, for example, Canca [6], Hagendorff [20], Hickok [21], Kazim and Koshiyama [23], Mittelstadt [29], and Zhou and Chen [54].

[12] Our model has not (yet) been empirically validated; thus, the benefits are currently only anticipated.

Table 1 The 3D model as compared to Morley et al. [31], and Ball and Koliousis [4]

3D model	Morley et al. [31]	Ball and Koliousis [4]
Design	Validate	Objective setting
Develop	Verify	Training
Deploy	Evaluate	Deployment management

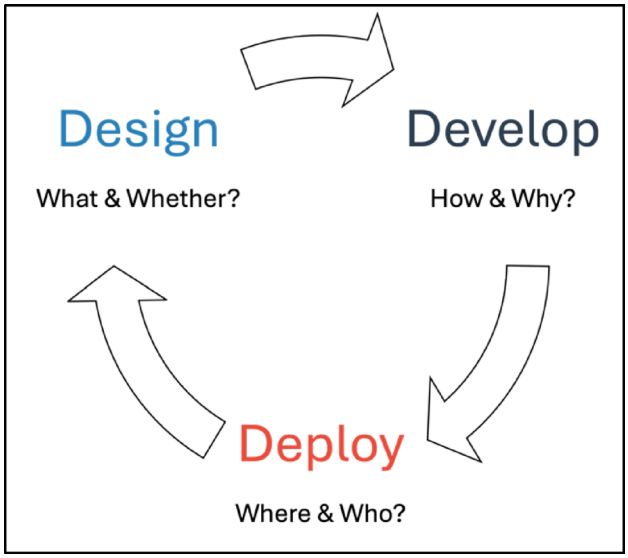

Fig. 1 The 3D model

activity on the computer. The purported utility of this is that Recall can help you find an email from last week, or help you find accidentally deleted work; no need to trawl through files in a panic, simply ask Recall. Unfortunately, Recall has had to be recalled, over security concerns [8]. The problem is that in order to be ready to recollect your actions on the computer, Recall has to monitor and record everything you do. Recall was set up, as Collins writes, to 'literally take screenshots of what you're doing every few seconds, then scrape the information from those screenshots and store it in a database.' [8]. This already might sound alarm bells for those concerned with data privacy, however it does get worse. The data in Recall's database was unencrypted. Malicious agents could potentially access a database of personal data, including financial information and passwords. This possibility was demonstrated by security researchers [8]. As Wes Miller stated when speaking to Forbes: 'as a whole, the approach is wrong… The fact that it created a centralized treasure-trove of data that should have been protected but wasn't, is a fundamentally flawed design.' [8].

The question remains: how did Recall, a system with clear ethical red flags, go all the way through to deployment? We will return to this case as an illustration of how our proposed model can be implemented, and how following our model could have raised the red flag on Recall at each stage of the process.

2 The solution

2.1 Motivation

The AI development process is iterative, and ongoing, and can helpfully be viewed as a cycle. While various accounts might be given of that cycle, in our view it can usefully be thought of as comprising three phases (which give the 3D model its name): design; develop; and deploy (cf. [11]). These correspond to Ball and Koliousis' [4] objective setting (design), training (develop), and deployment management (deploy),[13] and to Morley et al.'s [31, 32] validate (the design), verify (the development), and evaluate (the deployment).[14] Thus, whilst the identification of a cycle of AI development is not new, our approach is corroborated by others, and we propose new, clear labels for each stage. This has the (crucial) advantage of psychological simplicity, which we anticipate will facilitate the adoption of the model amongst practitioners (see below) (Table 1).

At each phase, the 3D model (Fig. 1) presents key questions that those involved in developing AI should ask and answer before continuing. This intervention aims to give structure to ethical decision-making during the design process for AI, facilitating discussions of the ethics of the AI system. This could be described as functioning at 'Level 3' of the *The Digital Catapult AI Ethics Framework* (DCEF). As Morley et al. [31] describe, 'The third level operationalizes 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 [17],' [31, p. 244]. However, unlike the DCEF, the 3D model ties key questions into each of these development stages, and offers indicative follow-up questions to help to articulate the key issues throughout the process.[15] This ensures that

[13] As Ball and Koliousis emphasize, the phases are nested, one within another, making each an essential part of the greater whole. The phases are therefore temporally overlapping and cannot be sequentially disentangled. This accordingly yields an iterative, and ongoing, design cycle [4].

[14] As Morley et al. write: 'The first phase (validation) is concerned with whether the right algorithmic system is being developed; the second phase (verification) is concerned with whether the algorithmic system is being developed in the right way; and the third phase (evaluation) is concerned with whether the algorithmic system is continuing to operate in the right way once deployed, needs to be revised, or can be improved (redacted).' ([31], p. 224).

[15] These follow-up questions have been suggested to draw out the relevant issues in the overarching questions at each stage—however,

ethical discourse occurs every stage of the development of the AI system. Whilst follow-up questions may be altered following validation, we see the overarching questions at each phase as essential to the model.

It is important to note that the envisioned ethical discussion will (ideally) be open-ended and ongoing. Allow us to make three points by way of substantiation of this thought. First, in Ball and Koliousis [4], for example, some stages of the development process are represented as embedded within others, culminating in deployment management. Crucially, AI systems in operation generate new data that interacts with the algorithms it uses in novel ways, with outcomes that are themselves subject to ethical evaluation. Indeed, the very fact that AI systems are sometimes recalled suffices for us to see that the ethical assessment of AI systems during deployment may require a re-appraisal of the answer to the question we present here as occurring at the design stage, namely whether the system should be built (and deployed) at all (see below). Of course, the aim of implementing the 3D model is to avoid the need for recalls and the harms that precipitate them, by anticipating and designing out potential problems in advance. But this should not obscure the present point: ethical decision-making in relation to AI systems is ongoing. Here that fact is represented with the simple cyclical structure of Fig. 1 for ease of uptake.

Second, ethical issues are typically open-ended – including those that concern AI development. As Canca ([6], p.18) notes, there are 'unavoidable value trade-offs' that 'organizations confront' when developing and deploying AI systems (and indeed more generally). Ethical judgments about how to manage those trade-offs in particular contexts are holistic, and must therefore be made in light of all of the evidence available. Accordingly, for this reason also, we cannot specify when ethical deliberation should come to an end, for example by articulating a number of iterations of the cycle that is needed to ensure some notion of ethical compliance.[16] In practice, we anticipate that different organizations will integrate the 3D model into their corporate governance procedures differently (see below).

Third, as a conceptualization of an ethical AI development process, we expect that our model – including all of the stages and the associated questions at any given stage—will be available[17] to those implementing it at any given stage. Thus, we expect that: when designing a system, practitioners will anticipate the need to address the questions that arise at development and deployment stages; when developing the system, there will be a recognition of the need to revisit the answers given to questions raised at the design stage, and to anticipate the answers that are relevant to the deployment stage; and that at the deployment stage, development choices will be revisited, possibly in light of new evidence regarding how well the performance of the system matches the initial design (which itself may end up needed to be reassessed). Since each stage of the cycle will in fact occur at least once as a concrete reality, this suffices to ensure that the questions are raised at least twice prior to deployment, and then again at least once following (what might be an initially piloted) deployment.

We would also like to stress that the follow-up questions we have articulated below are each given a clear—even provocative—formulation ('the ethical AI practitioner asks…'), in order to be suitable to generate the kinds of discussion we see as necessary for ethical decision making. In practice, teams will likely need to modify or nuance the questions themselves, and/or formulate others, in order to address their development aims; moreover, organisations may make adjustments for their particular contexts, guiding practitioner teams in the specific manner in which the 3D model of ethical AI practice ends up operationalized within their corporate governance procedures (see below). Once again, though, we present the model in a simple, straightforward manner, with both directness and specificity, in order to facilitate ready adoption. And while the follow-up questions presented may need to be altered following empirical validation of our model, and/or flexed during its implementation, it is important to note that we see the overarching questions at each phase as essential to the model. In particular, we believe the asking of these questions will help to ensure that the key ethical issues arising in the development of AI systems are not overlooked. Accordingly, properly addressing these questions—in part through engagement with appropriate follow-ups—should suffice for ethical AI development; certainly, doing so is necessary.

2.2 The 3D Model

We will now explain the three phases, and the associated questions, in turn. Once again, though, we remind our readers that our model must be presented as having a simple structure so as to (psychologically and culturally) facilitate uptake. The complexities of the issues that will arise in its implementation must not be integrated into the articulation of the model itself—though of course, in the current academic context, we include some discussion of those complexities (e.g. in footnotes), to show that our model is simultaneously both general and flexible enough to

future empirical validation may suggest the need for variant formulations, or even different questions. Furthermore, specific organisations might tailor these follow-ups for the precise contexts in which they operate. See below for further discussion.

[16] See below for the problems associated with tick-box exercises when it comes to AI ethics and the role the 3D model can play in avoiding them.

[17] Ideally through its internalization, though in practice with external e.g. visual aids [supplementary material, excluded for blind review].

accommodate those complexities, and to indicate how we anticipate it may assist in surfacing and resolving them.

i. Design

In the **design** phase we ask:

1. **What** are we building?
2. **Whether** to build it?

In the design phase, we are concerned with what the aims are for the AI system under consideration, and whether a system embodying those aims should be pursued at all. This crucially includes thinking about whether the aims are appropriate. It might be that the aims that characterize the planned system need to be modified, or even that the project should be abandoned. At this stage, practitioners should not be focussed solely on commercial viability, but on broader (e.g. ethical) considerations.[18]
Thus, in a little more detail, when designing an AI system, the ethical AI practitioner asks:
What?

- What is the system that we are proposing aiming to do (i.e. what is its objective)?[19]
- What are the potential benefits of such a system (and to whom do they accrue)?
- What are the associated risks (and who bears them—i.e. who incurs the costs if they materialize)?

Whether?

- Whether such a system would do good? And avoid causing harm?[20]
- Whether the costs or benefits accrue disproportionately to some, at the expense of others?
- Whether, on balance, the benefits outweigh the costs?

ii. Develop

In the **develop** phase we ask:

1. **How** should we go about building the system?
2. **Why** should we build it in that way?

The develop phase is for considering how the system can be operationalized, or implemented—in short, how it can be built—and why it should be done in that way. This includes considering how any such system is trained, and why it should be trained on the selected data.
When developing an AI system, the ethical AI practitioner asks:
How?

- How should we build the system to achieve its objective? (Does that objective need adjusting in light of practicalities about building a system that implements it?)
- How should we collect training data (e.g. to respect autonomy, and avoid bias)?
- How can we implement an appropriate data infrastructure (e.g. to ensure privacy)?

Why?

- Why are we making – and should we make—these engineering choices?
- Do they ensure the benefits to stakeholders that are part of the design brief?
- Do they avoid the potential harms identified in relation to such a system?

iii. Deploy

An absolutely crucial stage in the design cycle is that of deployment: for it is here that the effects of the system will

[18] As Wiggins and Jones note, Pasquale [35] points to 'a "second wave" of algorithmic accountability' ([51], p. 253), in which the question whether certain AI systems should be built at all gets raised. In including the question 'whether' in the 3D model we aim to recognize such second wave concerns.

[19] Objective setting is, of course, a technical notion: it is the task of determining the objective function that the system will approximate or compute. But it is also a non-technical notion: that of determining the purpose of the system (to which that function is pertinent/appropriate). What we are suggesting here is that the ethical AI practitioner engages in objective setting not only in the first of these senses (which, as engineers, they are sure to do), but also in the second (which, since they are ethical, they are obliged to do)—and that these tasks are linked. (This point is related to Marr's [28] computational level of explanation in cognitive science which asks after both what function a (in his case cognitive) system computes and the reason why, i.e. its purpose.) It might be thought that objective setting in the technical sense is a matter of how to meet the antecedently set non-technical objective, and that it accordingly belongs at the second, develop stage. In response we note that objective setting is given as the first stage of the process for Ball and Koliousis [4], and that the boundaries between stages are not rigid and that this is perhaps the point at which the transition between them occurs. In any case, what is key for present purposes is that practitioners should be induced to think of non-technical objectives when engaging with objective setting in the technical sense they are likely to be familiar with. We believe the present formulation of the question achieves this—though, as we have said, we are open to reformulations of the questions following empirical testing.

[20] This question is intended to draw out whether the system is aimed at doing good in the world (e.g. for the climate, for conservation, or tackling poverty. See [42]) or is simply avoiding causing harm (through bias, unfairness, or malevolent uses). We can make a distinction between a system that aims for altruism, or aims to generate profit without negative impact.

be felt by stakeholders; and the use of the system will generate further data that can be used to monitor and evaluate/assess its performance.

In the **deploy** phase we ask:

1. **Where** can (and can't) this system be ethically deployed?
2. **Who** will be affected by its use?

In the deploy phase, we are concerned with where the appropriate (or inappropriate) applications of this system are, and who will be impacted by the system's use. Reflection here should go beyond the intended application, thinking about dual uses. It should also involve thinking about who will be affected beyond the intended user group (including wider society).

When deploying an AI system, the ethical AI practitioner asks:

Where?

Where can/can't the system be ethically deployed?

Where—in what context(s)—were the training data collected? And where—in which context(s)—can we legitimately generalize from the statistical properties of that data to legitimately deploy the model (e.g. with accuracy)?

Where—in what context(s)—might there be dual use potential? Are some alternative (unintended) uses ethically impermissible—and if so, can they be/have they been designed out?

Who?

Who is affected by the deployment of the system in this context? Have they been—or can they be—consulted and engaged in the design process?

What (e.g. legally protected) characteristics do they have?

What are their legitimate interests – i.e., which values are in play? How can those interests and values be navigated?

At each phase, practitioners should reflect back on the answers to questions in the previous phase. For example, in the 'deploy' phase, practitioners should consider **what** the aims for the system are, **how** they've been implemented, and **why** those choices were made. This will inform the answers to **where** the system can be appropriately deployed. When the deployment phase is reached, monitoring of the system may reveal the need for significant changes to the design, and the cycle begins again. Those involved in the process should not only reflect back on the answers provided at previous stages, but also anticipate forwards to the next one. Whilst the illustration in Fig. 1 does not show additional arrows to capture this (for the sake of representational simplicity), practitioners should not consider each phase in isolation.[21]

Revisiting Recall under the 3D model, we can see that there is failure at multiple points of the design and development of the system, as mapped to the three stages:

1) Design: In the design phase, we are concerned with what the aims are for the system, and whether it should be pursued at all. This includes thinking about whether the aims are appropriate. Whilst one aim is to assist with the recollection of computer activity, to achieve this the system has to have full access to record everything you do (just on a fundamental level, it seems this would be hard to avoid). Whilst having your computer be able to assist you in recovering work or emails is helpful, there are considerable risks associated – it seems that Microsoft did not consider the value of privacy and security, and weigh these up against the potential convenience of Recall in their decision on whether to build the tool.
2) Develop: The develop phase is for considering how the system can be operationalized, and why it should be done in that way. This includes considering how any system is trained, and why it should be trained on the selected data. During this phase, developers might have considered how appropriate data infrastructure could be implemented, particularly considering users of the system and their security (as raised in the design phase). This would likely lead to developers considering encrypting the relevant data (which was not done with Recall), if the company still decided to go ahead.
3) Deploy: In the deploy phase, we are concerned with where the appropriate (or inappropriate) applications of this system are, and who will be impacted by the system's use. This should go beyond the intended application, thinking about dual uses. It should also involve thinking about who will be affected beyond the intended user group (including wider society). There is an obvious concern here, which is that if someone was maliciously inclined, they could access a complete trove of data about a Recall user. The lack of encryption and the sheer volume of personal and financial information that could be accessed through someone's typical computer use is a deadly combination for identity theft, fraud, and even blackmail.

[21] It may also be that a nested structure, as presented by Ball and Koliousis [4], is reflective of the realities here. We are certainly open to this, though for psychological simplicity we are using the cyclical structure in our representation. This is not an unusual move—for instance, in network visualizations, loops that connect a node (corresponding in our case to a stage) to itself are often omitted.

Table 2 Example implementation of the design phase of the 3D model

1 DESIGN: What & Whether?			
What are we building?		**Whether** to build it?	
What is the system we are proposing aiming to do (what is its objective)?	*E.g. We are aiming to build a system that helps people to retrieve past information from their computer*	Whether such a system would do good? And avoid causing harm?	*E.g. The system will be convenient, a more advanced version of "time machine". However, it could ease access to users' private information*
What are the potential benefits of such a system (and to whom do they accrue)?	*E.g. It will benefit our users, to help them find what they need easier. This could be a selling point, benefitting us as a company*	Whether the costs or benefits accrue disproportionately to some, at the expense of others?	*E.g. The benefits will only be to our users, or anyone who misappropriates the technology. The benefit to our users is not at anyone else's expense, but it could be at the expense of our users if privacy and security is not protected*
What are the associated risks (and who bears them—i.e. who incurs the costs if they materialize)?	*E.g. it could help the wrong people retrieve information. This will harm individuals and in turn cause reputational damage to our company*	Whether, on balance, the benefits outweigh the costs?	*E.g. The benefits of the system are at a level of convenience, but the costs are high if done poorly. We should not proceed without robust security measures*
We are building…	*E.g. A system which records what users do, helping them to retrieve information with ease at request*	**We should/ should not…**	*E.g. We should only build this system if we can ensure robust security—if we cannot, we should cease development*

To demonstrate how the structure of the questions can be implemented, see the table below (Table 2). Following the 3D model can be easily recorded in a table or through more detailed records within organizations. The recommendations of each phase (e.g. that recorded at the bottom of the table below) must be passed on to those involved in the next phase of system development, if development continues. It is worth noting that for some questions, practitioners may not be able to answer alone. Accompanying documentation should highlight that further investigation may be necessary in order to answer each question, and that practitioners should not merely give their 'best guess'.[22]

[22] Investigations might follow those recommended in the Value Sensitive Design Framework: conceptual, empirical and technical [18]. Whilst various other stakeholders are not explicitly involved in answering the 3D questions in the first instance, developers might need to consult them in some cases. Note also that the model we present here might be refined through empirical research.

2.3 The benefits of the 3D model

The potential benefits of this model are that it is pro-ethical and value-aware, amenable to implementation, it embeds ethics at every stage of the development process, it embeds a culture and language of ethics in organizations and provides clear decision points. It is worth noting, of course, that we have not (yet) conducted any empirical validation of the model we are proposing, and the benefits we envision are therefore currently only hypothetical. Nevertheless, the approach taken is amenable to further investigation that might test for these benefits.

i. Pro-ethical and value aware

Morley et al. [30] highlight the need for pro-ethical design, that is, design which forces agents to make choices without limitation:

> The difference between ethics by design and pro-ethical design is the following: *ethics by design* can be paternalistic in ways that constrain the choices of agents, because it makes some options less easily available or not at all; instead, *pro-ethical design* still forces agents to make choices, but this time the nudge is less paternalistic because it does not preclude a course of action but requires agents to make up their mind about it. ([30], p2142 fn1)

In our approach then, we suggest a series of questions that must be asked and answered at each stage of AI development. This is a 'pro-ethical approach' as outlined above. In our model, there is no prescription on how the questions ought to be answered. Indeed, in this case we take inspiration from the value sensitive design [18, 47] and value analysis in design [26], which encourages reflection on the values at play from various first and second order stakeholders. We aim for our model to increase the awareness of the values at play in the development and deployment of an AI system.

ii. Amenable to implementation

A key barrier in adopting ethics in AI development has been the gap between high level principles and practice, as stated by Morley et al. [30]: 'closing the gap between principles and practice by constructing a typology that may help practically minded developers apply ethics at each stage of the Machine Learning development pipeline' (p. 2141). Whilst we do not aim to offer a typology, we do aim to provide a tool that can facilitate ethical considerations at every phase of the design and development of AI systems.

Morley et al. [30] found that AI ethics tools were very limited in terms of how easily they could be used in practice. It is clear to us that any ethics tools aimed at the development of AI systems need to have a clear path to implementation and need to be firmly rooted in the development of AI systems. The benefit of this approach over VAD or VSD is that it is simplified in terms of implementation, and fits with the current ML development pipeline. A benefit of this model over principle-focused approaches is that practitioners are not required to have a specialist understanding of ethical theories or concepts.[23] This, we hope, allows for an easier insertion of the approach into the existing practices of AI developers.

iii. Ethics at every stage

As noted by Ball and Kouliosis [4] we need AI 'that engages with human needs, interests, and priorities at every stage of development, shaping the decisions that are taken concerning whether, and if so how, the AI system is built.' (§3.1). Our model ensures that ethics decision-making is required at each stage of the development process. As explained above, our model covers the ML development pipeline as divided into three phases, which we dub 'design', 'develop' and 'deploy', and which align with the stages as outlined in Ball and Koliousis' [4] and to Morley et al. [31]. The benefit of this approach is that questions of ethics can be ensured to be embedded at every stage of the ML pipeline. As Morley et al. [31] highlight, embedding ethics at every stage ensures that this model avoids becoming a 'one-off tick-box exercise' ([31], p. 246), which is completed and then forgotten about. Instead, the design and development of AI systems becomes iterative through an ongoing cycle: reflection on development and deployment should feed back into the design of the system. We are seeing similar approaches to guidance around data protection in AI development, such as the ICO's guidance on 'data protection fairness' throughout the AI lifecycle which divides the lifecycle of data in the AI development process into eight stages.[24]

As noted by Morley et al. ([30], p. 2160) there will be an inherent limitation to the approach of breaking the AI development process into stages, as in practice, technology development cannot be neatly divided into distinct steps. To combat this, our approach focuses on relevant questions at each of three broad stages of development, as opposed to breaking this down into smaller and smaller stages which may become increasingly indistinct in practice.[25] Our approach also tackles the potential weakness highlighted by Morley et al. ([30], p. 2156) who found that insufficient attention is paid to the deployment of ML tools in ethics.

iv. A culture and language of ethics

Drawing from literature on the effectiveness of business ethics more broadly, our model aims to increase a culture of ethics in the development of AI in practice. It is worth noting the relative lack of discussions regarding organisational culture, however, barriers to ethics in business practices will also play out in organisations developing AI tools. As Seger [41] highlights, culture is central to the success of ethical AI practices, even when a principles approach is taken, because these rules do not dictate action, and people are more likely to adhere to policies when they believe in them.

As Treviño and Nelson [45] state, "in a strong ethical culture, ethics becomes a natural part of the daily conversation in the organization. Employees feel comfortable talking ethics with each other and with their managers. Organizational values are invoked in decision making. And managers routinely talk ethics with their direct reports." ([45], p 185). They highlight prior research in which 'individuals who discussed their decision-making using ethical language were more likely to have actually made an ethical decision' (p. 186). The culture in tech organizations has also been identified as a key barrier to effective AI ethics [33]. The question-focused model presented here will facilitate the language and culture of ethics in organizations which follow it. Our hope is that this will encourage a shift away from organizations paying lip-service to high level principles to a broader discussion of relevant values and trade-offs in the development of AI. Furthermore, this will help to embed a culture of ethics in organizations, which will allow for a ground-up approach to ethical AI development as opposed to a top-down (principles first) or punitive model of development which tends towards box-ticking or legalising ethical decisions.[26] Whilst the idea of integrating ethical discourse into AI development is not new (see for example [17]), this model however aims to integrate this into the process at every phase of the design process.

[23] This is not to say that ethics should not be discussed at a theoretical level. Indeed, at the level of governance we expect that theoretical discussion will be needed in order to set organisational positions (see [6]). Our aim here is to provide a tool for developers/practitioners that does not require higher level theoretical learning to be effective. For discussion of the different levels at which ethics functions within society and organizations, see Tasioulis [43], Ball [3], and our response to objection four below.

[24] See ICO's website https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/annex-a-fairness-in-the-ai-lifecycle/.

[25] Our stages may well be somewhat blurred in practice, and we imagine both looking forward and reflecting backwards at each stage. We have, however, struck what we hope is an appropriate and useful balance between articulating structure and capturing detailed realities.

[26] This must be accompanied by cultural support from managers to avoid 'moral muteness' [45].

xxii. Clear decision points

The structure of this model also allows for clear decisions to be made and recorded. This combats the issue common to AI ethics cases that there is difficulty in assigning responsibility for harms from AI. Whilst this model cannot account for potential unforeseeable consequences (the typical focus of the responsibility gap, see [24]), a record of responses to the questions in the model, and the addition of final decisions at each stage (e.g. 'Whether we should make this?') allows for corresponding culpability of analysis and decision making. Those involved in the 3D model will not be able to deny having thought about possible consequences or inadvertent uses of their technologies. Note that this cannot protect against unforeseen harms entirely, however this will narrow the culpability gap. As noted by Santoni de Sio and Mecacci [40] 'Culpability gaps are concerning insofar as the more persons designing, regulating, and operating the system can legitimately (and possibly systematically) avoid blame for their wrong behaviour, the less these agents will be incentivised to prevent these wrong behaviours.' ([40], p. 1063). This is similar to the 'problem of many hands' to which computing is vulnerable, as highlighted by Nissenbaum [34].

If the process of the 3D model is properly recorded, it can also demonstrate where in the process of development something has gone wrong. In the table above for example, if the final recommendation of the 'design' phase is that "We should only build this system if we can ensure robust security—if we cannot, we should cease development." then there is a record if the system continues to deployment without robust security and privacy measures in place that this concern was known to the teams involved following this report. The increase in likelihood of blame then should help to incentivise ethical practices in organisations developing AI technologies.[27]

In addition, the recommendations made at each phase of the development process can act as justification for any termination in the development of the system if it cannot be completed in accordance with these recommendations. This can target 'sunk cost' thinking; people are unwilling to abandon projects even when doomed [19]. Suggested strategies to mitigate this include building in frequent decision points, and encouraging reflectivity in practice [37]. The 3D model encourages both ethical reflectivity and has built in decision points.[28]

Finally, it should also be considered that punishment is not always appealing as an approach to governance. For example, punishment which is too severe can encourage actors to attempt to hide or cover up errors (see e.g. [1]). Furthermore, it's possible that misattribution of responsibility can occur to lower-level designers when AI systems go wrong, where these individuals become scapegoats for AI harms ("moral crumple-zones", see [12]). It is possible that the 3D model, besides providing clear decision points so we might appropriately attribute blame, could also provide protection from inappropriate ascriptions of blame. For example, if a team has properly and thoroughly completed the 3D assessment, yet a (truly) unforeseeable harm occurs, the evidence of the 3D process can protect them from organizational consequences such as losing their job. Further to this, if practitioners have properly completed a 3D report they could avoid blame for any issues that arise if recommendations are ignored at later stages of development or deployment. This could be implemented at an organizational or regulatory level. These kinds of policies could help to incentivise use, and particularly serious use, of the 3D model.

3 The context

We are conscious that, if our 3D model of ethical AI practice is to be successful, enabling ethical AI development, it will require embedding in an appropriate (social) context. Indeed, we envision three levels at which norms need to be adhered to, and actions taken to put principles into practice, if ethical AI is to be a reality: appropriate regulation and corporate governance will be needed, alongside a culture of ethical practice on the part of professional AI system developers. In this section, we clarify our view of how the 3D model fits into the ethical/responsible AI ecosystem by responding to objections. As Munn [33] has provided a particularly forceful recent articulation of concerns about the effectiveness of AI ethics initiatives, we will frame our discussion in the terms of his objections. We nevertheless expect that the points addressed will resonate more generally, and that the additional nuance provided will demonstrate how we anticipate the 3D Model of Ethical AI Practice making a positive difference.

[27] This will of course need to be coupled with actual mechanisms for accountability, such as review and punishment. Praise and reward mechanisms should also be considered to encourage ethical practices (rather than simply punishing unethical practices). See below for further discussion.

[28] Better still would be to ensure that decisions are reviewed by a diverse body at a distance to the project itself.

3.1 Objection 1: AI ethics is useless (and so is the 3D model)

As we have already seen, Munn [33] has launched a thoroughgoing assault on the very idea of AI ethics: it has proved to be utterly useless, on his assessment. Looking more closely at Munn's rejection of AI ethics, we see that he understands 'ethics… in the narrow but well-established sense [of] "a set of moral principles"' ([33], p. 869) such as those mentioned above (e.g. wellbeing, justice, and respect for persons), and argues there is 'a gap between principles and practice' ([33], p. 870). He complains in particular of principles that are: meaningless; isolated; and toothless.

The first concern is that, because the ethical principles that get articulated are variously interpretable, and accordingly lack a fixed meaning, those who would be bound by them are able to avoid any substantive demand to improve their AI practices, by engaging instead in what Floridi has called 'ethics shopping'—namely, 'the malpractice of choosing, adapting, or revising… ethical principles… from a variety of available offers, in order to retrofit some pre-existing behaviours' ([15], p. 186). The second worry is that merely articulating principles of AI ethics without embedding them within a culture of respect is guaranteed to be of no avail. 'Unethical AI,' says Munn (with characteristic bluntness), 'is the logical byproduct of an unethical industry.' ([33], p. 871) (The departures of AI ethicists from large tech firms noted above might be thought to provide evidence in support of this claim). And the third point is that 'AI ethical principles have failed due to the lack of consequences' ([33], p. 871) when they are violated. Here the thought appears to be that adherence to the principles of AI ethics requires enforcement,[29] and that ultimately, this needs to come from the state—from outside of organizations, not from within (as AI ethics initiatives are often touted).

We will address these concerns separately in what follows, explaining how they might be thought to apply to our 3D model in particular, before responding. But first we would like to respond to the overarching concern that AI ethics is useless. To this end, it is important to note that Munn is not some sort of value relativist, or nihilist: his complaint is not that there is no such thing as doing (absolutely) better in value-relevant dimensions when it comes to the development and deployment of AI systems. Rather, Munn thinks that improvements will not come from (the pretense of) grappling with ethical principles; and he suggests two specific alternatives which he thinks will work better. One is more technical: what is needed, he thinks, is (something like) implementations of key value decisions in code, so that developers can make use of the solutions they provide to avoid ethical harms and produce ethical benefits. The other is more social—and external to the companies that employ AI developers: we need regulation, and we need it enforced.

The first of these ideas, in a way, answers the objection that ethical principles are variously interpretable, and therefore meaningless. For technical packages such as code libraries are not so variously interpretable: they implement value decisions in specific ways; or so one might think. Munn's second positive proposal, equally, answers to the concern that the principles of AI ethics are toothless: for what better than external regulation—and e.g. state bodies with powers of enforcement—to give teeth to any ethical principles that might be held to be societally important? Thus, in a way, Munn's rejection of AI ethics ultimately embodies a kind of scepticism about the possibility of addressing his remaining concern, the worry that the principles of AI ethics are isolated. Munn appears to think that the culture of the AI industry cannot be markedly improved, in such a way as to render AI ethics useful. It is here that we are more optimistic, as we will explain below. But the key point for present purposes is that the first, overarching objection that we would like to address—namely, that AI ethics is useless—simply rests on a *non sequitur*. Munn's argument has, as its structure, the premise that (the articulation of principles of) AI ethics on its own is *insufficient* for the ethical development and deployment of AI systems, and the conclusion that it is therefore *unnecessary*, or useless. Similarly, it might be thought that merely having our 3D model available will not, on its own, adequately address the various ethical crises recently witnessed in the industry. But it cannot be concluded that the model is therefore of no help. This conclusion, like Munn's, simply does not follow.

Ultimately, context matters when it comes to implementing AI ethics in practice: this includes aspects of corporate governance (structures, policies, and procedures), and legal setting (including regulation and enforcement), as well as the culture amongst practitioners in the sector. In what follows we aim to show how the 3D model can be integrated with appropriate corporate governance, developer education, and state regulation to facilitate ethical AI practice.

3.2 Objection 2: meaningless principles (no operationalization)

The 3D model of ethical AI practice is just that: a model. As such, it can fail to be implemented, and in this way fail to yield any real benefits. But some might harbour a more specific worry about it: it does not adequately *operationalize* the principles of AI ethics. That is, the 3D model does not serve to *translate* the abstract concepts employed in ethical

[29] It is worth mentioning that incentives are also needed. Munn does not address this possibility, but this is an area where organisations are well-placed to contribute. However, we may instead need a new business model, which expressly follows regulation.

principles into more meaningful operational terms—either those of computer code that developers might use, or those of corporate procedures to be followed. Accordingly, in this somewhat more specific way, it might be thought that the 3D model is useless in operationalizing AI ethics.

Indeed, we opened our paper by citing Morley et al. [30] with approbation, endorsing their call to move AI ethics from the 'what' of principles to the 'how' of practices. But a closer look at that paper reveals that, as Morley and colleagues put it a few years later, it '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.' ([32], p. 411) The idea appears to have been that, if an AI developer is asked, for example, to make a 'fair' (i.e. unbiased) classification tool (that respects the principle of justice), they will need some way of understanding concretely what they are being asked to do, so that they can comply with the request: and they can use the taxonomy to search for tools and methods—whether code packages, or algorithmic design strategies—that will enable them to do so.[30]

We concede that the 3D model does not provide translations, in this sense, from the principles of AI ethics to specific coding practices.[31] Of course, AI development is not just a technical process. But neither, it might be objected, does our model indicate how to translate from the principles of AI ethics to appropriate practices at e.g. the corporate level to ensure ethical AI development either. That is, there are good questions that arise at the level of corporate governance about how to ensure that AI development results in systems that adhere to principles of AI ethics—for instance: what procedures should be followed; what policies should be implemented; and what structures should be in place. Should an ethicist be embedded in every R&D team? Or should the technical development team be required to check in with an independent ethics committee periodically, or at certain development steps or stages? These are not questions that our 3D model aims to answer either.

In short, our model is not aimed at the operationalization of AI ethics in the strict sense of providing tools[32] for the translation of abstract concepts into more concrete terms.[33] We are more concerned with *implementing* AI ethics: ensuring that there is *conformity* with ethical principles, rather than guidance by them.[34] In fact, it may be an advantage of our approach that AI developers will not need extensive training in theoretical ethics—they need not understand the differences between deontological and consequentialist approaches, for instance—in order to ask the questions posed by the 3D model at the appropriate stages in the development cycle. Nevertheless, in earnestly attempting to answer these questions, they are likely to uncover, and begin to address, the key ethical issues the development of the envisioned AI system raises.

Thus, we acknowledge that, without appropriate support in the form of structures, policies, and procedures that are ultimately a matter of corporate governance, developers are likely to struggle to implement ethical AI solutions. And we also concede that technical tools will often be needed, once key ethical decisions are made. In short, the 3D model of ethical AI practice can only succeed when embedded in an appropriate context. Nevertheless, fostering a culture amongst developers of seeking to implement the 3D model, by asking the questions it raises at the appropriate development stages, will help to ensure that the principles of AI ethics are adhered to. We elaborate on this in what follows next.

3.3 Objection 3: isolated principles (no implementation)

Another obstacle to the implementation of the 3D model of ethical AI practice that might be suggested by Munn's discussion is an unethical culture in the industry. And this worry is perhaps particularly pressing given our response above to the objection that the 3D model does not operationalize AI ethics: for we have conceded this, while emphasizing that success will come from implementing ethical AI practice; accordingly, if the 3D model does not help with this, it might fairly be said to be of no use at all.

Above we suggested that measures need to be taken at the corporate governance level to facilitate ethical AI practice: and that point is pertinent here too. But in the present

[30] The typology is available in Morley et al. [30].

[31] As we have seen, Munn [33] appears to want something like the technical solutions provided in the taxonomy presented by Morley et al. [30] in order to make real progress in addressing the harms caused by AI systems. It is unclear to us how, though, in the absence of the principles of AI ethics, such tools could be appropriately used: for the value-decisions themselves that are ultimately to be encoded through technical means need to be made at the non-technical level,and so it seems the principles are needed to guide the process. (How else could developers find the appropriate tools, if not by looking for some code that implements some particular interpretation of a principle?).

[32] Many such tools and toolkits exist. In addition to those mentioned in Morley et al.'s [31] taxonomy, see also Canca's [6] 'box' and Vallor's [48] toolkit, available at the following link: https://www.scu.edu/ethics-in-technology-practice/ethical-toolkit/.

[33] See e.g. Chang [7].

[34] This distinction between conforming to and following a rule stems from the scholarly discussion of Wittgenstein [52]. See e.g. Reiland [38, 39], Ball [2], Ball, Helliwell and Rossi [5], and Pelland, Träctler and Lowe [36].

context we would like to emphasize a different response: while appropriate corporate governance can synchronically facilitate ethical AI practice, appropriate education and training for AI developers can also help; and it is perhaps here above all that we envision the 3D model as being useful.

Indeed, in articulating his objection that the principles of AI ethics are isolated, Munn writes: 'If the tech industry lacks ethics, so does the education of the software engineers and technologists who will soon join it. Undergraduate data science degrees emphasize computer science and statistics but fall short in ethics training.' ([33], p. 871)

We agree with Munn's (implicit) assessment that the education of AI developers must incorporate ethics and/or value-sensitivity. Indeed, we welcome both Ball and Koliousis's [4] call to train 'philosopher engineers' (at post-secondary level) who are versed in ethical AI design, Dabbagh et al.'s [10] demand for AI ethics education in (primary and secondary) schools, and TechUK's [44] call for investment in educational pathways that include both technical and ethical components. This will go some way towards addressing the concern that the culture in the industry is problematic.[35] Indeed, we regard the 3D model as itself a contribution to this goal: for in our view, it may constitute a mnemonically simple, and therefore psychologically effective, part of an ethics training for AI developers. If students can be encouraged to remember to embed ethics into the AI design cycle in the simple manner proposed in the 3D model, this will provide an important impetus to culture change within the industry.

3.4 Objection 4: toothless principles (no consequences)

A final concern that might be raised about the 3D model again parallel's Munn's objection to AI ethics in general, namely that without any consequences for failure to conform (to the principles, by implementing the model effectively), tech companies and their AI development teams will not be incentivized to improve their practices in ethical respects. As we have seen, Munn suggests that what is ultimately needed is regulation, with legal penalties for ethical failures.

It is worth saying that, all things considered, we agree. But the 3D model can readily be integrated with legal initiatives. For example, in describing the 3D model, we have said that AI development teams should pause to ask key questions at each step in the iterative design cycle; and we have also indicated that the answers to some of these questions should typically reference the answers to others—so that, for example, when explaining why certain technical solutions were implemented at the development stage, developers should point to the answers to the question of what the objective of the AI system in question is. Part of the aim of this is to ensure that ethics is not regarded as a tick-box exercise, considered once and then set aside. We did not say in the model itself that development teams must write down their answers to the questions raised at each stage, though obviously this would help them in recalling their previous decisions, and in making such cross-references. We do indicate however that recording answers can be helpful, particularly to ensure adequate oversight (and corresponding reward, exculpation, or punishment).

What we would like to stress in the present context is that it is consistent with what we have said that governments should require companies in the tech sector to be prepared to undergo a form of ethical audit for their product development. There are efforts to implement effective AI regulation in progress, such as the EU's risk-based approach in the AI Act, which requires organisations to undergo a conformity assessment with high-risk technologies (European Commission [13]). However, this is markedly different from the requirement to undergo an ethical audit of all AI development, particularly as low risk AI systems are not subject to the same requirements in the act (European Commission [14]).[36] If there were such a legal requirement, this might well incentivize the adoption of appropriate corporate governance policies—for example, policies requiring notetaking when answering the very questions proposed in the 3D model. Thus, while the 3D model is not a panacea, it can be embedded within a broader social context, in such a way as to implement ethical AI practice.

Thus, Munn's three specific objections have helped us to identify three things that need to change: practitioner culture; corporate governance; and legal regulation. Our 3D model provides a key piece that can help to address the culture problem in particular: for it can be incorporated into the education of AI developers, during their university careers, or in corporate training sessions for career and professional development. It can also be integrated with appropriate corporate governance and regulatory initiatives, be they negative (in terms of punishment), or positive (in terms of reward).

Accordingly, our response to the broad objection that AI ethics (in general) is useless—and that our 3D model in particular is as well—is that it needs to be embedded in an appropriate context. That means using it as a tool for education and training to address worries about corporate culture,

[35] Some post-secondary institutions are making some headway on this: for instance, Northeastern University's Khoury College of Computer Science expects its graduates to take an oath: see https://www.khoury.northeastern.edu/about/mission-and-vision/.

[36] Furthermore, we are yet to see this, or other approaches tested, so it is unclear whether they will indeed have 'teeth'.

integrating it within appropriate systems of corporate governance to facilitate its implementation, and supplementing it with legal regulation and enforcement. If such measures are undertaken, we see no impediment to the implementation of the 3D model, and thereby of ethical AI practice.

4 Conclusion

We have articulated a simple model of ethical AI practice that ensures that ethics is embedded and implemented iteratively at every stage of the design cycle. We first framed the contemporary problem of AI ethics, namely that there are ethical violations in AI practice, which a focus on high-level ethical principles have failed to tackle. In this climate, there has been a growing call for the operationalization and/or effective implementation of AI ethics at the practical level. We have responded to this call for an articulation of the 'how' of AI ethics with our 3D model—named for the three stages of the AI cycle: design, development, and deployment—comprising critical, but readily understandable questions to be answered iteratively at each phase. And we have articulated a number of potential benefits of this model: that it is pro-ethical, implementable, embeds ethics at every stage, enables a culture and language of ethics, and provides clear decision points. Future work should validate this approach, and the particulars (e.g. of follow-up questions) should be refined in light of such empirical work. Finally, we have suggested that our model, if embedded within appropriate supporting structures, can help address the four objections to AI ethics as put forward by Munn [33], that AI ethics is in general useless, that it has no operationalization, and that its principles are both meaningless and toothless. In this way, we envision addressing the need for implementable models of ethical AI practice.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s43681-025-00812-7.

Acknowledgements We are grateful to an anonymous reveiwer who took great care in giving lots of helpful feedback on this manuscript. We also want to thank students and colleagues in Philosophy and Computer Science at Northeastern University London who gave helpful feedback on this project, as well as our colleagues in Boston, John Basl, Meica Magnani, Vance Ricks and Riccardo Baeza-Yates for their continued support. Finally we would like to acknowledge the support from Northeastern University's Tier 1 grant programme.

Author contributions Both authors contributed equally to the development of the model presented here, and to the writing and reviewing of the paper.

Funding Open access funding provided by Northeastern University Library. This research was developed from the Building AI and Data Ethics Infrastructure project funded by a Northeastern University Tier 1 grant.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

Ethical approval Not applicable. There was no ethical approval needed for this research.

References

1. Anderson, R.: Security Engineering: A Guide to Building Dependable Distributed System, 3rd edn. Wiley (2021)
2. Ball, B.: Playing games, following rules, and linguistic activity. In: Philosophical Insights into Pragmatics 79 (2019)
3. Ball, B.: Putting ethical AI into practice. [Blog]. Z/Yen (2024). Available at: https://www.zyen.com/community/the-ethical-ai-thought-space/putting-ethical-ai-into-practice/
4. Ball, B., Koliousis, A.: Training philosopher engineers for better AI. AI Soc. **38**(2), 861–868 (2022). https://doi.org/10.1007/s00146-022-01535-7
5. Ball, B., Helliwell, A.C., Rossi, A.: Wittgenstein and artificial intelligence, volume II: values and governance. Anthem Press, London (2024)
6. Canca, C.: Operationalizing AI ethics principles. Commun. ACM **63**(12), 18–21 (2020)
7. Chang, H.: Operationalism. In: Zalta, E. N. (ed.) The Stanford encyclopedia of philosophy, Fall 2021 edn. Available at: https://plato.stanford.edu/archives/fall2021/entries/operationalism/
8. Collins, B.: Recall recalled: is AI on windows 11 already doomed? Forbes (2024). Available at: https://web.archive.org/web/20240724152643/https://www.forbes.com/sites/barrycollins/2024/06/14/recall-recalled-is-ai-on-windows-11-already-doomed/. Accessed 30 Nov 2024
9. Corrêa, N.K., Galvão, C., Santos, J.W., Del Pino, C., Pinto, E.P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., Terem, E. de Oliveira, N.: Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. Patterns **4**(10) (2023). https://www.cell.com/patterns/fulltext/S2666-3899(23)00241-6
10. Dabbagh, H., Earp, B.D., Mann, S.P., Plozza, M., Salloch, S., Savulescu, J.: AI ethics should be mandatory for school children. AI Ethics (2024). https://doi.org/10.1007/s43681-024-00462-1
11. De Silva, D., Alahakoon, D.: An artificial intelligence life cycle: from conception to production. Patterns (2022). https://doi.org/10.1016/j.patter.2022.100489

12. Elish, M.C.: Moral crumple zones: cautionary tales in human-robot interaction. In: Engaging Science, Technology, and Society (2019)
13. European Commission: Artificial Intelligence – Questions and Answers. European Commission (2024). Available at: https://ec.europa.eu/commission/presscorner/api/files/document/print/en/qanda_21_1683/QANDA_21_1683_EN.pdf
14. European Parliament: EU AI Act: first regulation on artificial intelligence. European Parliament Topics (2025). Available at: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence
15. Floridi, L.: Translating principles into practices of digital ethics: five risks of being unethical. Philos. Technol. **32**, 185–193 (2019). https://doi.org/10.1007/s13347-019-00354-x
16. Floridi, L., Cowls, J.: A unified framework of five principles for AI in society. Harvard Data Sci. Rev. **1**(1) (2019). https://doi.org/10.1162/99608f92.8cd550d1
17. Floridi, L., Strait, A.: Ethical foresight analysis: what it is and why it is needed? In: The 2020 Yearbook of the Digital Ethics Lab, pp. 173–194. https://doi.org/10.2139/ssrn.383048
18. Friedman, B., Kahn, P., Borning, A.: Value sensitive design and information systems. In: Zhang, P., Galletta, D. (eds.) Human-Computer Interaction in Management Information Systems: Foundations, pp. 348–372. M.E. Sharpe Inc, New York (2006)
19. Garland, H.: Throwing good money after bad: the effect of sunk costs on the decision to escalate commitment to an ongoing project. J. Appl. Psychol. **75**(6), 728–731 (1990). https://doi.org/10.1037/0021-9010.75.6.728
20. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. Minds Mach. **30**(1), 99–120 (2020). https://doi.org/10.1007/s11023-020-09517-8
21. Hickok, M.: Lessons learned from AI ethics principles for future actions. AI Ethics **1**(1), 41–47 (2021). https://doi.org/10.1007/s43681-020-00008-1
22. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**(9), 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2
23. Kazim, E., Koshiyama, A.S.: A high-level overview of AI ethics. Patterns (2021). https://doi.org/10.1016/j.patter.2021.100314
24. Königs, P.: Artificial intelligence and responsibility gaps: what is the problem? Ethics Inf. Technol. (2022). https://doi.org/10.1007/s10676-022-09643-0
25. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015). https://doi.org/10.1016/j.patter.2021.100314
26. Kopec, M., Magnani, M., Ricks, V., Torosyan, R., Basl, J., Miklaucic, N., Muzny, F., Sandler, R., Wilson, C., Wisniewski-Jensen, A., Lundgren, C., Baylon, R., Mills, K., Wells, M.: The effectiveness of embedded values analysis modules in computer science education: an empirical study. Big Data Soc. (2023). https://doi.org/10.1177/20539517231176230
27. Knight, W.: Elon Musk has fired twitter's 'ethical AI' team. Wired (2022). Available at: https://www.wired.com/story/twitter-ethical-ai-team/
28. Marr, D.: Vision: a computational investigation into the human representation and processing of visual information. W. H. Freeman and Company, San Francisco (1982)
29. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. Nat. Mach. Intell. **1**(11), 501–507 (2019). https://doi.org/10.1038/s42256-019-0114-4
30. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci. Eng. Ethics **26**(4), 2141–2168 (2020). https://doi.org/10.1007/s11948-019-00165-5
31. Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., Floridi, L.: Ethics as a service: a pragmatic operationalisation of AI ethics. Minds Mach. **31**, 239–256 (2021). https://doi.org/10.1007/s11023-021-09563-w
32. Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., Floridi, L.: Operationalising AI ethics: barriers, enablers and next steps. AI Soc. **200**, 1–13 (2023). https://doi.org/10.1007/s00146-021-01308-8
33. Munn, L.: The uselessness of AI ethics. AI Ethics **3**(3), 869–877 (2023). https://doi.org/10.1007/s43681-022-00209-w
34. Nissenbaum, H.: Accountability in a computerized society. Sci. Eng. Ethics **2**, 25–42 (1996). https://doi.org/10.1007/BF02639315
35. Pasquale, F.: The Second Wave of Algorithmic Accountability. LPE Project 25 November (2019). Available at: https://lpeproject.org/blog/the-second-wave-of-algorithmic-accountability/
36. Pelland, J.C., Trächtler, J., Love, H.: Practice makes human: why we can't understand black-box artificial intelligence. In: Ball, B., Helliwell, A.C., Rossi, A. (eds.) Wittgenstein and Artificial Intelligence, Volume II: Values and Governance (2024)
37. Perignat, E., Fleming, F.F.: Sunk-cost bias and knowing when to terminate a research project. Angew. Chem. (2022). https://doi.org/10.1002/anie.202208429
38. Reiland, I.: Rule-following I: the basic issues. Philos Compass **19**, e12900 (2024)
39. Reiland, I.: Rule-following II: recent work and new puzzles. Philos Compass **19**, e12976 (2024)
40. Santoni de Sio, F., Mecacci, G.: Four responsibility gaps with artificial intelligence: why they matter and how to address them. Philos. Technol. **34**, 1057–1084 (2021). https://doi.org/10.1007/s13347-021-00450-x
41. Seger, E.: In defence of principlism in AI ethics and governance. Philos. Technol. **35**, 45 (2022). https://doi.org/10.1007/s13347-022-00538-y
42. Taddeo, M., Floridi, L.: How AI can be a force for good. Science **361**(6404), 751–752 (2018). https://doi.org/10.1126/science.aat5991
43. Tasioulis, J.: First steps towards an ethics of robots and artificial intelligence. J. Pract. Ethics **7**(1), 49–83 (2019)
44. TechUK: Mapping the responsible AI profession, a field in formation. [Report] (2025). Available at: https://www.techuk.org/resource/techuk-paper-mapping-the-responsible-ai-profession-a-field-in-formation.html
45. Treviño L.K., Nelson, K.A.: Managing Business Ethics: Straight Talk About How To Do It Right, 5th edn. Wiley (2011)
46. Turing, A.M.: Computing machinery and intelligence. Mind **59**, 433–460 (1950)
47. Umbrello, S., De Bellis, A.F.: A value-sensitive design approach to intelligent agents. In: Yampolskiy, R. (ed.) Artificial Intelligence Safety and Security, pp. 395–409. Chapman and Hall/CRC (2018)
48. Vallor, S.: An ethical toolkit for engineering/design practice. Markkula Center for Applied Ethics (2018). Available at: https://www.scu.edu/ethics-in-technology-practice/ethical-toolkit/
49. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
50. Wallach, H.: Navigating the Broader Impacts of Machine Learning Research. Medium (2021). Available at: https://hannawallach.medium.com/navigating-the-broader-impacts-of-machine-learning-research-f2d72a37a5b
51. Wiggins, C., Jones, M.L.: How data happened: a history from the age of reason to the age of algorithms. WW Norton & Company, New York (2023)
52. Wittgenstein, L.: Philosophical Investigations. Wiley-Blackwell, New York (1953)

53. Wooldridge, M.: The road to conscious machines: the story of AI. Penguin, London (2020)
54. Zhou, J., Chen, F.: AI ethics: from principles to practice. AI Soc. **38**(6), 2693–2703 (2023). https://doi.org/10.1007/s00146-022-01602-z