Agency and Structure:

Conceptualising Applied AI Ethics in Organisations

Zach TAN Zhi Ming and Devesh Narayanan

Key Takeaways

- Organisations that deploy systems or products utilising Artificial Intelligence (AI) are under increasing pressure to address ethical concerns associated with these technologies in a manner that moves beyond mere regulatory compliance.
- Our understanding of these ethical concerns is in constant flux, as we learn more about AI and its impact on society. How might organisations fix in place structures and processes to deploy ethical AI, when ethical AI itself is a moving target?
- We interview individuals within nine organisations in Singapore attempting to deploy ethical AI, focusing on how they refer to existing government and industry frameworks on AI ethics to shape internal processes.
- Organisations find it generally difficult to appraise the downstream effects of AI products and systems in terms of social costs while they are still being developed. Instead, they appear to rely on *prima facie* moral intuitions to evaluate AI systems, and on reframing ethical concerns as business risks.
- Within organisations, three potential job roles are identified which seem best situated to handle the emerging organizational responsibilities related to ethical AI: the AI Ethics Officer, the Product Manager, and the AI Auditor.
- For policymakers, we recommend facilitating a network or roster of experts, who are encouraged to suggest and organise into working groups decided by the members themselves. This would present a dynamic manner of identifying emerging problem areas, combined with a lean process to produce workable outcomes which can then be rapidly iterated on.
- For organisations, we recommend cultivating a flexible, yet robust "ethics infrastructure" to provide context and guidance for employees to make more informed judgments about the AI systems they work with, and to clear the path for ethical action by providing clear structures and processes for concerns to be raised and addressed effectively.

1 Introduction

From the most fleeting online interactions to consequential assessments for loan or credit eligibility, technologies based on Artificial Intelligence (AI) hold great sway over our social and economic interactions. There is a growing recognition of the social costs associated with the typically obscured decisions made by algorithms that are likened to "black-box" systems.[1]

As more attention is paid to the potential and realised harms of AI-based systems, various industry experts, governments, NGOs, and academic experts have been working to create frameworks of AI ethics focused on harm-reduction and optimisation for socially good outcomes. Against the backdrop of growing social and political pressure, companies developing AI-based products are expected to take these frameworks and translate them into action – moving beyond mere regulatory compliance. However, this translation from frameworks to practice, or the "what" to the "how",[2] remains ambiguous for most organisations. Further, the "what" itself is shifting. As we learn more about AI's potential impact on society, our understanding of the relevant concerns for ethical AI is in constant flux. How might companies meaningfully translate these frameworks into practice, when these frameworks themselves are a moving target?

This paper presents a crucial first attempt in outlining how this translation might occur. We do so through a qualitative process of semi-structured interviews with various stakeholders, including data scientists, managers, senior executives, and industry experts. Although our sample size is relatively small, our interviews yielded rich data about how high-level AI ethics principles are interpreted, translated, and implemented into practice in Singapore-based organisations. This data underpins the analysis and discussion presented in subsequent sections.

1 This term is taken to mean a system with known and observable inputs and outputs, but with an obscured internal working. See Pinch, Trevor J. 1992. "Opening Black Boxes: Science, Technology and Society." Social Studies of Science 22, no. 3: 487–510.

2 Morley, Jessica, et al. 2020. "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices." Science and Engineering Ethics 26, no. 4: 2141–68. (https://doi.org/10.1007/s11948-019-00165-5).

2 Background

One central thrust of recent research on AI ethics has been to develop high-level principles that articulate pertinent ethical concerns related to AI systems, with guidance on addressing them. While this paper is focused more on advancing the application of this research, rather than its content, the rest of this section focuses on briefly outlining the shape of the field, some of its key developments, and recent efforts – by way of clarifying what we mean when we discuss AI ethics.

2.1 Mapping the AI ethics debate

A recent scoping review analysed a corpus of 84 documents stating principles and guidelines from global efforts related to AI ethics.[3] They report convergence around five principles: transparency, justice and fairness, non-maleficence, responsibility, and privacy – while acknowledging that there remains divergence on how these issues are interpreted, normatively justified, and recommended for implementation. Another recent analysis concurs: describing a similar set of principles as those listed above as the "normative core of a principle-based approach to AI ethics and governance".[4] In this analysis, Fjeld et al. review 36 prominent AI principles documents, and find convergence on similar high-level principles, while once again acknowledging that normative concepts are invoked differently to conceptualise similar principles across various documents.[5]

Singapore's efforts to develop a framework of AI ethics and governance similarly converge towards these high-level principles. The recently released second iteration of the Model AI Governance Framework centres on the principles of explainability, transparency, fairness, and human centricity, and provides broad recommendations for firms to incorporate these principles into their AI systems.[6] Similarly, the Monetary Authority of Singapore issued their own set of principles for AI and data analytics in the financial sector, centred on the principles of fairness, ethics, accountability, and transparency.[7]

One other type of document often overlooked by commentators in the land-

3 Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. "The Global Landscape of AI Ethics Guidelines." Nature Machine Intelligence 1, no. 9: 389–99.

4 Fjeld, Jessica, et al. 2020. "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI." SSRN Scholarly Paper. Rochester, New York: Social Science Research Network.

5 Fjeld et al. find convergence in eight broad themes: privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values. Further elaboration on these eight themes finds that 47 individual principles can be subsumed within these eight – illustrating how these themes are conceptualised differently in different documents.

6 Model Artificial Intelligence Governance Framework-Second Edition (https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf).

7 Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector. (https://www.mas.gov.sg/~/media/MAS/News%20and%20Publications/Monographs%20and%20Information%20Papers/FEAT%20Principles%20Final.pdf).

scape of AI ethics is governance models by standards-setting bodies. Industry standards such as COBIT and ISO/IEC 38500:2015 provide detailed and often technical guidelines on governing IT systems, including advice on how to set up governance bodies to audit and mitigate risks related to the deployment of IT systems.[8] Though these are often skimmed over in favour of flashier national or international principles-based frameworks, standards-setting bodies provide a known and trusted source of guidelines and information to industry practitioners, even if these are not currently aimed specifically at AI systems.

Clearly, there is an oversupply of references for companies to consult when attempting to deploy AI systems ethically. Even as convergence occurs towards a few central themes, organisations seeking to adopt principled approaches to AI ethics must still frame these concerns with respect to their own organisational practices. Furthermore, additional consideration of local social and cultural norms must also be made to understand how they might shape the specific expression of these principles. As most organisations are still in their early stages of attempting this translational exercise, studying these organisations may provide insight into key factors enabling or limiting the wider adoption of these principles in industry.

2.2 How might organisations adapt?

Organizations that adapt well will consult a wide range of sources, referenced to their local context, to then decide on what adaptations might be necessary of their processes to ensure the development of AI-based products and services aligned with ethical principles. Organisations that adapt (deliberately) poorly may use these myriad documents to effectively cherry-pick principles which fit existing practices, contributing to a broader process of performative ethics without accountability, or what is otherwise known as "ethics-washing".[9] Our paper only focuses on the translation problem, although we also acknowledge the risks of such a proliferation of frameworks and guidelines. Regardless, for organisations that wish to make a good-faith attempt at translating these principles into practice, the adaptations they must make depend on what their current processes and decision-making infrastructures look like, as well as the specific AI use-case in question.

One possible adaptation is to strengthen the organisation's ethical infrastructure, comprising of "both formal and informal elements – including communication, surveillance, and sanctioning systems – as well as organisational climates for ethics, respect, and justice".[10] Prior research in business ethics elaborates on various manifestations of this concept, including providing concrete and comprehensible communication about ethical values, and formal rewards for exemplary ethical behaviour.[11] Recent efforts in the philosophy of technology, through coining the term "infraethics", also make a connection between moral behaviour and their surrounding "expectations, attitudes,

8 ISACA. 2019. "COBIT: Control Objectives for Information Technologies." (https://www.isaca.org/resources/cobit). ISO/IEC. 2015. "ISO/IEC 38500 Information Technology – Governance of IT for the Organization." (https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/28/62816.html).

9 Vincent, James. April 2019. "The Problem with AI Ethics." The Verge. (https://www.theverge.com/2019/4/3/18293410/ai-artificial-intelligence-ethics-boards-charters-problem-big-tech).

10 Tenbrunsel, Ann E., Kristin Smith-Crowe, and Elizabeth E. Umphress. 2003. "Building Houses on Rocks: The Role of the Ethical Infrastructure in Organizations." Social Justice Research 16, no. 3: 285–307.

11 Fernández, José Luis, and Javier Camacho. 2016. "Effective Elements to Establish an Ethical Infrastructure: An Exploratory Study of SMEs in the Madrid Region." Journal of Business Ethics 138, no. 1: 113–31.

rules, norms and practices", in relation to ethical decision-making about information and communication technologies.[12] By strengthening their ethical infrastructures, organisations may allow employees working on AI systems to raise concerns without hesitation and ensure that these concerns are channelled to and effectively addressed by the relevant parties.

12 Floridi, Luciano. 2017. "Infraethics – on the Conditions of Possibility of Morality." Philosophy & Technology 30, no. 4: 391–94.

3 Methodology

Considering the challenges previously outlined, organisations seeking to practise AI ethics must simultaneously look outwards to understand and interpret high-level principles, while also building up internal capabilities to address the translation of these principles into practice for their own specific use-cases. Structured linearly, this presents three potential problem areas. First, the translation of high-level principles must result in actionable tasks to be carried out by people. Second, the roles which are best positioned to take up these tasks must be appropriately identified or created. And finally, to ensure that these roles meaningfully contribute towards the ethical deployment of AI products and systems, organisations need to clearly articulate the influence of ethical considerations within their decision-making processes.

However, given the relatively new emergence of formalised structures and job roles in AI ethics within organisations globally (Microsoft, for example, only formed their first full-time position in AI policy and ethics in 2018),[13] such neat linearity is unlikely to manifest across various organisations in Singapore. Therefore, instead of comparing static indicators like organisational charts or job descriptions, our methodological approach aimed to produce rich qualitative data that could speak to the complex experiences of the people within organisations attempting to translate principles into action, so that we might extrapolate the root causes enabling or limiting the implementation of AI ethics.

We utilise semi-structured interviews with our participants to gather data on four central themes. These are: what organisational factors might empower employees to raise concerns about AI ethics; how do employees raise and act on these concerns; what kinds of processes (if any) surround the appraisal of these ethical concerns; and how are these processes maintained for long-term sustainability? In addition, we also leverage prior work done at the Lee Kuan Yew Centre for Innovative Cities to reconstruct key job roles as identified by our participants in a co-production exercise for the purposes of better integrating AI ethics concerns into existing organisational structures.[14] Finally, we selected participants with a bias for those close to or directly involved in current ethical decision-making processes. Over the course of this study, we interviewed ten individuals across nine organisations. These organisations are varied in terms of both size and reach. Two organisations are leading

13 Davenport, Thomas H. 2020. „What Does an AI Ethicist Do?“ MIT Sloan Management Review. (https://sloanreview.mit.edu/article/what-does-an-ai-ethicist-do/).

14 Ong Teng Cheong Labour Leadership Institute, and Lee Kuan Yew Centre for Innovative Cities. 2018. “Polarising of Job Opportunities: Charting New Pathways and Adopting New Technologies.” Labour Research Conference 2018 Proceedings. Singapore: Labour Research Conference 2018.

Fortune 500 companies with a large presence in Singapore; four are mid-sized or large companies with operations across multiple Asian countries, and the remaining three are smaller organisations operating primarily within Singapore. Our participants also represent a diverse range of experiences. While all self-identify as being active participants in the implementation of AI ethics, their backgrounds and job domains include data science, engineering, law, corporate investing, product management, and technology marketing and management. In the following section, we present our findings from these interviews, grouped into three common themes.

4 Findings and Implications

4.1 The importance of ethical infrastructure

One clear finding from our interviews is that in order for AI ethics to gain purchase within an organisation (in terms of being factored into decision-making), it needs to be encompassed within a larger ethical framework alongside a well-developed organisational infrastructure that supports and encourages the voicing of ethical concerns more generally. This infrastructure further needs to be supported and endorsed by senior members in the organisation and must contain multiple channels for raising concerns and acting upon them. Finally, such an infrastructure also needs to unequivocally support and protect whistle-blowers.

Public support from senior leadership of the organisation is essential as a signalling mechanism. If senior leadership does not explicitly support the prioritisation of ethical concerns over commercial outcomes, junior members of the organisation will not feel comfortable bringing up ethical concerns, lest they conflict with commercial imperatives. Instead, these members feel the need to either repress these concerns or couch them in terms of "business risk", as one of our participants reported. Moreover, another participant, a supervisor in a data science team, notes that "having a meetup of 300 people could be less impactful than two extremely senior vice-presidents having a chat over coffee" – further illustrating the importance of getting buy-in from senior leadership who can set organisational priorities (and thus, culture) around ethics.

Having multiple channels for reporting ethical concerns is similarly crucial as it presents people with a way to circumvent managerial elements who may not share the same concerns. These channels collect feedback both internal and external to the organisation and can manifest in both informal and formal manners. Informal channels regularly consist of cross-team mailing lists and interest groups, enabling information sharing for ethical concerns across functions like engineering, product, and opera-

tions. Some formal channels may be set up for specific projects in the organisation, while others collect more general feedback and concerns from employees. Typically, this manifested in a split between project managers and people managers, so that an employee could raise a concern to one without having to go through the other.

It further helps when feedback is sought from the end-users of the product. As mentioned by one of our participants who is an executive at a popular ridesharing company, "a lot of our drivers in the early stages personally knew or were connected via WhatsApp to the founders [and] senior executives, and ... a lot of feedback came directly from them." Having a close connection to the end-users further helps to factor in the social outcomes of an AI application throughout multiple layers of the development process, which also enables a greater sense of ownership among employees.

4.2 Ad-hoc ethics

Almost all our participants highlighted that AI ethics concerns are presently being treated by people in an informal and ad-hoc manner within their organisations. This is partly due to a general lack of education around AI ethics, but also partly because it remains difficult to appraise the downstream effects of AI products and systems in terms of social costs while they are still being developed. The appraisal process, therefore, resembles more instinctual rather than procedural recognition. Additionally, high-level principles and frameworks do little to alleviate this issue, as they are generally viewed as being too generic to be of practical use within these organisations.

When queried about why AI ethics remains instinctual rather than procedural, our participants described how there were little resources or time allocated to seriously treat these concerns in a structured manner, in large part due to the "start-up" nature of their organisations. As a result, and in the absence of a system to standardise the appraisal of potentially thorny ethical issues, there is a reliance on individual appraisals, which are often highly ambiguous and ad-hoc in nature, and necessarily subjective. Our participants mentioned that this process can often be simply condensed into asking if something is being done "weird" or by asking the question: "would you be comfortable telling your mother what you have done"?

As to why high-level frameworks were not useful in this case, our participants pointed to the large differences between the applications of AI within and across companies, often commenting that the frameworks are "too generic ... [and] too watered down". Furthermore, these frameworks often assume that AI is presented and deployed as a final product, but the reality is that something as simple as credit scoring requires a tremendous amount of iteration to get [to] a model where [the] training data gets ... an outcome which can be verified and [is] accurate".

In these cases, it is generally unclear where one would find the time or resources (amidst constant iteration) to ensure alignment with these high-level principles in an efficient manner.

4.3 Efficiency, Competitiveness, and Ethics

Our conversations with interview participants often turned to the competitive logic underlying the overall business environment, and how it frames the ways in which employees think about ethical concerns.

One participant analogised that the various processes in a company resemble the flow of a river, moving according to the logics of efficiency and competitiveness. Processes related to ethics, here, resemble a net – trying to catch problems as they emerge throughout the company, but in turn, slowing down the flow of the river. The larger and more fine-toothed the net, the more likely it is to catch all the problems, but also the greater the impediment presented to the river's flow.

In almost all the interviews, some version of this antagonistic relation between ethics and business imperatives presented itself. A common pragmatic view is that ethical processes may be easier to establish if they do not "clog up" other business functions. The resulting operationalisation of ethics was then a series of interconnected responsibilities diffused across different roles in the product life cycle. To return to the analogy: this would resemble various small nets, strategically placed at various points along the river, thereby leaving the rest of the river to flow largely unimpeded.

Further, our participants also suggested that ethical processes should be built to fit existing structures, rather than create new ones. One of our participants suggested that ethical considerations be folded into and normalised as part of software development or managerial processes as far as possible, to minimise the role of ethics as a perceived "external" force on decision-making. For software engineering, this would involve including checks for bias or fairness (if these could be operationalised in a technical manner), alongside the usual testing for stability, uptime, usability, and other such parameters.

This notion of reframing ethical norms as business norms appeared to be commonplace. One participant spoke about how they would reframe privacy concerns in terms of "business risks" to get their point across more effectively – citing how it was important to "use commercial reasoning and thought processes" in commercial spaces. This business reframing of ethics appeared to be a sticking point in the current discourse on AI ethics, with another participant lamenting how ethics is used as a profitable proposition … [and] not as an altruistic good to serve society".

Finally, our participants also told us that, on the individual level, incentive structures need to change to accommodate ethical decision-making. This relies on commercial reasoning: if performance incentives or measurements are defined by conventional business outcomes, employees might not be interested in raising ethical concerns that are unrelated to, or even counterproductive to, these outcomes. Participants called for business ethics to be somehow quantified and measured as part of an employee's internal performance metrics wherever possible. Where this is not possible, organisations would need to set up alternate structures to ensure that these concerns are appropriately addressed.

5 Job Roles

As mentioned in Section 3, one important theme in our interviews centred on new responsibilities and tasks related to translating AI ethics principles into action, and which job roles might be best suited to take these on.

Considering that many of these roles and responsibilities are not yet formalised in most Singapore-based organisations, we approached this theme in a hypothetical and co-produced manner, working together with our participants to imagine (or reimagine) these job roles. As to why these roles were not yet present in our participants' organisations: many felt that their companies were not yet at the stage where such roles were either necessary or viable. For example, an organisation with a flexible, rules-averse culture would find it preferable to let ethics permeate throughout, rather than try to formalise it in one place (many small nets). Smaller companies would find it difficult to justify employing a dedicated full-time employee solely to perform an ethics function. And organisations deploying third-party AI solutions feel that the responsibility for ethical AI falls on the provider of these solutions, rather than themselves.

These barriers notwithstanding, our interviews surfaced three job roles that seemed to be particularly important to ensure the development and deployment of ethical AI: the AI Ethics Officer, the Product Manager, and the AI Auditor. We summarise these roles and their importance below.

5.1 The AI Ethics Officer

In most conversations, this role was conceptualised as a mid- to senior-managerial position, overseeing the deployment of AI systems throughout the organisation. One participant recommended that these officers report directly to the C-suite and the Board, to ensure that they can meaningfully provide oversight without being bogged down by bureaucratic processes.

The AI Ethics Officer would take on a few key responsibilities. First, they would survey the burgeoning developments in AI ethics principles, regulations and research, to pick out those that are relevant to the company, and through conversations with various stakeholders in the company, carve out a space for ethical considerations within existing business and technical decision-making processes. They would also serve to champion greater AI ethics education within the organisation. Further, they would act as a centralised point of contact for the issues raised by employees working on AI systems to stream up to. They would then consolidate these concerns, develop strategies to address them, and report all this to senior management. Finally, they

would participate in national and regional conversations about AI ethics – advocating for their company's position on these matters and helping shape policy outcomes related to AI ethics.

5.2 The Product Manager

Product Managers (PMs) already play an important role in many organisations today and appear to be especially well suited to take on responsibilities related to AI ethics. They work closely with development teams and would have intimate knowledge about AI systems being deployed, their technical specifications, and potential vulnerabilities and weaknesses. They coordinate constantly with various other teams and would, therefore, have sight of the objectives and key results that others are striving towards. One participant raised the example of how, for instance, considerations about bias are already important in Human Resources departments, and a PM could help translate these considerations into technical requirements for another team working on developing hiring algorithms. Third, PMs are also well situated to understand customers and end-users of the AI systems being deployed, to assess how their needs and requirements might bear on the development processes of these systems.

Traditionally, the PM's role centres on assessing how the concerns of other stakeholders – both internal and external to the company – bear on technical processes, and then translating these into requirements for developers to execute. Concerns related to AI ethics, then, seem to be a natural fit, inasmuch as these are raised to the attention of PMs by the stakeholders they interact with. AI Ethics Officers, or other senior managers involved in ethical AI, could also work closely with PMs to translate principles and policies into technical requirements for development teams.

5.3 The AI Auditor

In our conversations, the role of the AI Auditor was conceived as an extension to both compliance functions – related to, for instance, complying with data protection or privacy regulations – as well as Quality Assurance functions – related to testing algorithms and their use-cases.

AI Auditors would be brought in at several key junctures during the development and deployment of AI systems: when a system is deemed complete and ready for deployment, when significant changes have been made to existing systems, or when systems start to behave errantly. Auditors would then run a series of checks on these systems. Depending on the type of AI system in question, these checks could include data lineage and bias checks, system access and authorisation checks, or checks for the coherence of the model's logic, to name a few.

One participant also pointed to two quite different types of audits. One, where the auditors run through a list of existing checks and procedures related to the AI system in question, and another, where a "red hat" auditor takes an adversarial approach to the system, attacking it in various ways to find its faults. Participants also recommended that auditors should be independent and disinterested, such that the company's interests and priorities do not get in the way of a thorough audit.

6 Discussion

Earlier in the paper, we discussed a crucial tension between the call for AI ethics to move from abstract principles to routinised, fixed practices, and the recognition that "ethical AI" is a continually moving target as we learn more about AI and its impacts on society. Analysing a moving target, using ethical concepts that are themselves shifting, requires flexibility to be baked into approaches to ethical AI in practice. In this section, we discuss what this flexibility might look like and how it can help policymakers to advance the conversation on AI ethics, and organisations to better implement ethical AI.

6.1 Flexibility in the policy-making process

Though this paper focuses on the problems that organisations face when attempting to translate ethical principles into practice, we must also acknowledge that AI ethics – both in terms of how it is conceptualised and applied – is fundamentally an open-ended, multi-stakeholder problem. Policymakers play a significant role in driving forward the conversation about AI ethics, often acting as the nexus between academia, industry, and civil society, themselves translating various stakeholder inputs into various priority areas and workable policy solutions. Additionally, many of the most influential committees and councils that have produced high-level AI ethics frameworks have been formed through the actions of policymakers seeking to gather expert opinions and feedback on important topic areas.

However, the success of such efforts turns on the ability of policymakers to identify and invite the most relevant stakeholders, matched to a problem that is of the appropriate scope and workability. Identifying the right experts for the right problems is a difficult endeavour – given the constantly moving target of ethical AI *vis-à-vis* changing social considerations, alongside the continual emergence of new applications of AI. How might policymakers ensure that the most recent technical and social developments on AI are captured in their policies – even as the field is expanding all around them?

One possible solution is for policymakers to facilitate the formation of a *roster of experts* representing diverse stakeholders, who are encouraged to organise into working groups decided by the members themselves. Since much of the fluidity surrounding AI ethics is based on the continual evaluation and re-evaluation of issues by practitioners in the industry, academics, and civil society groups, facilitating a process in which these individuals can put forward working group suggestions without restrictions, which are then evaluated and agreed upon collectively or discarded, presents a dynamic manner of identifying emerging issues and prioritising policy development.

Such a process is not without precedent. In fact, we base this idea on existing frameworks found in internet governance, drawing lessons specifically from the internet standards-setting body of the Internet Engineering Task Force (IETF). Consisting of volunteers, the IETF works on technical issues pertaining to the underlying protocols of the internet. They are organised into working groups, suggested either by area leads or individual volunteers, and are designed to be short-lived in nature: typically expiring after achieving a specific goal or deliverable. We believe that this example points towards one way in which to provide an overarching structure to facilitate the organic identification and prioritisation of emerging problem areas, combined with a lean process to produce workable outcomes that can be rapidly iterated on.

6.2 Flexible ethics-infrastructures for organisations

Even as principles and policies move towards more complete representations of the concerns relevant to ethical AI, organisations deploying AI systems must already implement processes to catch and act on as many of these concerns as possible. To this end, diffusing the responsibilities related to AI ethics seems like the ideal option. There are both practical and commercial reasons for this preference. Practically, employees working on AI systems are best placed within the organisation to identify their benefits and harms. Further, it would be competitively advantageous to have strong ethical checks and balances, as being able to anticipate social harms would very likely result in a reduction of the costs associated with social harms in the first place (re-development, audits, enforcement and checking costs).[15] Finally, the diffusion of responsibilities also presents a potential solution to the problem of tracking developments in an ever-shifting field: avoiding the rigidity of formal approaches based on job roles. If more people are involved in conceptualising and acting on AI ethics, there is a greater chance that more diverse and important concerns are raised and addressed.

However, considering what our participants reported to us, we suggest that the diffusion of responsibilities is incompatible with the ad-hoc fashion in which most conversations about AI ethics are presently held. The reliance on *prima facie* moral intuitions and rules like "*do what feels right*" or "*do what you might be comfortable telling your mother about*" is unlikely to yield generally credible appraisals,[16] making meaningful disagreement when intuitions differ difficult.[17] Further, moral judgement itself is not sufficient for moral action. Several obstacles could stand in the way of an employee who identifies a concern with an AI system, and wants to bring this concern up to the relevant authorities for resolution.

15 Thomsen, Steen. 2001. "Business Ethics as Corporate Governance." European Journal of Law and Economics 11, no. 2: 153–64; Agafonow, Alejandro. 2017. "Transaction Costs and Business Ethics." In *Encyclopedia of Business and Professional Ethics,* 1–4. Cham: Springer International Publishing; King, Andrew. 2007. "Cooperation between Corporations and Environmental Groups: A Transaction Cost Perspective." Academy of Management Review 32, no. 3: 889–900.

16 A full discussion on moral intuitionism is outside the scope of this paper. However, usually, moral intuitions rely on the object being appraised – AI systems in this case – impressing upon the appraiser an unambiguous, self-evident moral judgement. This is usually not the case with AI. It is generally difficult, for instance, for a developer working on some feature of an AI system to think of the lines of code on their screen in terms of moral impact, much less intuit with certainty whether this feature leads to good or bad outcomes. See Stratton-Lake, Philip. 2020. "Intuitionism in Ethics." In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

17 Specifically, competing subjective intuitions cannot be meaningfully weighed against each other – since it is difficult for one person to share evidence about how they arrived at their intuitions with another. See Frances, Bryan, and Jonathan Matheson. 2019. "Disagreement." In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

For these reasons, we recommend focusing on cultivating a robust "ethics infrastructure" to ensure that the diffusion of responsibilities meaningfully contributes to the development and deployment of ethical AI. This ethics infrastructure would serve two purposes:

1. Provide context and guidance for employees to make more informed judgements about the AI systems they work with, and
2. Clear the path for ethical action – providing clear structures and processes for concerns to be raised and addressed effectively.

Our interviews surfaced some ideas about what such ethics infrastructures would look like. Since ethical norms are often compatible with and related to technical or business norms, organisations can fold ethics into existing technology development and management processes. In doing so, however, organisations must be careful to avoid prematurely reifying and holding stable abstract ethical ideas like fairness and privacy, despite there being considerable ambiguity about what these terms mean. Other ideas – such as the separation of managerial responsibilities, or the creation of multiple parallel channels for reporting concerns – complement this routinisation of ethics, to ensure that those concerns that cannot be folded into extant business and technical processes also have a place to be raised.

Foregrounding the agency of workers by diffusing responsibilities while simultaneously establishing robust ethics infrastructures requires a delicate balancing act. Here, the job roles identified through our interviews may be of help. AI Ethics Officers can work to actively align their organisation to the shifting landscape of AI ethics – both by tracking developments in the field and by actively shaping policies through collaborations with policymakers and other stakeholders. AI Auditors can update their checks to include new ethical concerns as they are unearthed. Product Managers can ensure that the most up-to-date desiderata for ethical AI are meaningfully translated into requirements for product development teams. In this way, we call for two simultaneous movements – the formalisation and standardisation of ethical principles into formal structures wherever possible, as well as continued efforts to keep these formal structures mobile, as we learn more about AI and its impact on society.

7 Conclusion

Dynamism and agility must underpin how organisations and policymakers approach ethical AI. For organisations, we recommend the implementation of ethics infrastructures that foreground and enable the agency of employees to take meaningful action on ethical concerns. A similar balancing of structure and agency may be appropriate for policymakers, by, for instance, setting up a roster of experts to (re)evaluate policymaking priorities and setting up working groups for incremental, tangible policy advancements. While further research will be needed to enable AI ethics principles to better reflect prevailing social concerns, and organisations to better implement these principles, we hope to have provided some seminal strategies and conceptual clarity that may guide these future efforts.

Authors

*Zach TAN Zhi Ming**

Zach is a Senior Research Assistant at the Lee Kuan Yew Centre for Innovative Cities (LKYCIC) at Singapore University of Technology and Design. He has a background in the social sciences and critical internet studies. At the LKYCIC, his research focuses on the gig economy; the future of work and emerging tasks; and how social stratification might affect Singapore's future digital society. Prior to joining the centre, Zach was a master's student at the Oxford Internet Institute, where he wrote his dissertation on reputation and rating systems in online labour markets. He has also worked at the Digital Ethics Lab at the University of Oxford, and as an intern at the Berkman Klein Center for Internet and Society at Harvard University. He holds an MSc from the University of Oxford, and a BA from Wesleyan University.

*Devesh Narayanan**

Devesh is a Research Assistant at the Lee Kuan Yew Centre for Innovative Cities (LKYCIC) at Singapore University of Technology and Design. His background is in engineering and moral philosophy. His research interests are primarily in science and technology studies, philosophy of technology, and applied ethics. At the LKYCIC, he works on a few projects: forecasting trends in the future of work with an emphasis on identifying new tasks that might be created by emerging technologies; understanding how social stratification bears on how Singaporeans conceive of and form aspirations about the future digital society; and understanding how smart planning can be deployed to achieve sustainable development in future cities. He is also currently pursuing an MA in philosophy at the National University of Singapore, where he writes about the normative underpinnings of Explainable AI.

* Both authors have made an equal contribution to this article.

References

A Agafonow, Alejandro. 2017. “Transaction Costs and Business Ethics.” In *Encyclopedia of Business and Professional Ethics*, 1–4. Cham: Springer International Publishing.

D Davenport, Thomas H. 2020. “What Does an AI Ethicist Do?” MIT Sloan Management Review. (https://sloanreview.mit.edu/article/what-does-an-ai-ethicist-do/).

F Fernández, José Luis, and Javier Camacho. 2016. “Effective Elements to Establish an Ethical Infrastructure: An Exploratory Study of SMEs in the Madrid Region.” Journal of Business Ethics 138, no. 1: 113–31.

Fjeld, Jessica, et al. 2020. “Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI.” SSRN Scholarly Paper. Rochester, NY: Social Science Research Network.

Floridi, Luciano. 2017. “Infraethics – on the Conditions of Possibility of Morality.” Philosophy & Technology 30, no. 4: 391–94.

Frances, Bryan, and Jonathan Matheson. 2019. “Disagreement.” In *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University.

I ISACA. 2019. “COBIT: Control Objectives for Information Technologies.” (https://www.isaca.org/resources/cobit).

ISO/IEC. 2015. “ISO/IEC 38500 Information Technology – Governance of IT for the Organization.” (https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/28/62816.html).

J Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. “The Global Landscape of AI Ethics Guidelines.” Nature Machine Intelligence 1, no. 9: 389–99.

K King, Andrew. 2007. “Cooperation between Corporations and Environmental Groups: A Transaction Cost Perspective.” Academy of Management Review 32, no. 3: 889–900.

M Model Artificial Intelligence Governance Framework-Second Edition (https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf).

Morley, Jessica, Luciano Floridi, Libby Kinsey, and Anat Elhalal. 2020. “From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices.” Science and Engineering Ethics 26, no. 4: 2141–68. (https://doi.org/10.1007/s11948-019-00165-5).

O Ong Teng Cheong Labour Leadership Institute, and Lee Kuan Yew Centre for Innovative Cities. 2018. “Polarising of Job Opportunities: Charting New Pathways and Adopting New Technologies.” *Labour Research Conference 2018 Proceedings*. Singapore: Labour Research Conference 2018

P Pinch, Trevor J. 1992. “Opening Black Boxes: Science, Technology and Society.” Social Studies of Science 22, no. 3: 487–510.

Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore’s Financial Sector. (https://www.mas.gov.sg/~/media/MAS/News%20and%20Publications/Monographs%20and%20Information%20Papers/FEAT%20Principles%20Final.pdf).

S Stratton-Lake, Philip. 2020. “Intuitionism in Ethics.” In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

T Tenbrunsel, Ann E., Kristin Smith-Crowe, and Elizabeth E. Umphress. 2003. “Building Houses on Rocks: The Role of the Ethical Infrastructure in Organizations.” Social Justice Research 16, no. 3: 285–307.

Thomsen, Steen. 2001. “Business Ethics as Corporate Governance.” European Journal of Law and Economics 11, no. 2: 153–64.

V Vincent, James. April 2019. “The Problem with AI Ethics.” The Verge. (https://www.theverge.com/2019/4/3/18293410/ai-artificial-intelligence-ethics-boards-charters-problem-big-tech).