An Ecosystem Approach to Ethical AI and Data Use: Experimental reflections

Mark Findlay[1] and Josephine Seah[2]

12 May 2020

Abstract

In response to fears around the risky and irresponsible development of artificial intelligence (AI), the prevailing approach from states, intergovernmental organisations, and technology firms has been to roll out a 'new' vocabulary of ethics. This self-regulatory approach relies on top-down, broadly-stated ethics frameworks intended to moralise market dynamics and elicit socially responsible behaviour among top-end developers and users of AI software. At present, it remains an open question regarding how well these principles are understood and internalised by AI practitioners throughout the AI ecosystem. The promotion of AI ethics has so far proceeded with little input from this group, despite their essential role in choosing and applying this emerging ethical language and associated tools in their project designs and related decision-making. As AI principles shift from normative organisational guides to operational practice, this paper offers a methodology—a 'shared fairness' approach—aimed at addressing this gap. The goal of this method is to identify AI practitioners' needs when it comes to confronting and resolving ethical challenges and to find a 'third space' where their operational language can be married with that of the more abstract principles that presently remain at the periphery of their work life. We offer a *grassroots approach* to operational ethics based on dialog and mutualised responsibility. This methodology is centred around conversations intended to elicit practitioners perceived ethical attribution and distribution over key value-laden operational decisions, to identify when these decisions arise and what ethical challenges they confront, and to engage in a language of ethics and responsibility which enables practitioners to internalise ethical responsibility. The methodology bridges responsibility imbalances that rest in structural decision-making power and elite technical knowledge, by commencing with personal, facilitated conversations, returning the ethical discourse to those meant to give it meaning at the sharp end of the ecosystem. By attending to practitioners, our project aims to better understand ethics as a socio-technical practice, progressing from the appreciation that as a realistic force in regulation, ethics are dynamic and interdependent.

This research/project is supported by the National Research Foundation, Singapore under its Emerging Areas Research Projects (EARP) Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

[1] Professor, Director – Centre for AI and Data Governance, School of Law, Singapore Management University markfindlay@smu.edu.sg

[2] Research Associate – Centre for AI and Data Governance, School of Law, Singapore Management University

Introduction

> *But on a contract and a project for another company, the number one thing that is driving choices is meeting terms of the contract, and there is little room for thinking about ethics, we're not breaking rules, but it boils down to meeting deadlines and getting things ready – rushed – as long as nothing is flagrant you do what needs to get done.*[3]

Artificial Intelligence (AI) and inter-connected big data usage, impacting on all aspects of human life, are under-regulated phenomena. Communities are confused by their complexity and technicality, while perceiving that AI and big data are both increasingly pervasive and represent risks to their social world. Such anxiety feeds off uncertainty as to whom AI will most benefit, what will be lost or displaced or amplified. AI promoters move their justifications for the technology from inevitability to blind faith, gaslit by ethical codes which may have limited deep reach into the AI ecosystem.[4] In response to a significantly negative community consciousness, the prevailing reassurance and legitimisation approach from state administrations, intergovernmental organisations and Big Tech firms – as agents for AI – has been to roll out a 'new' vocabulary of ethics and responsibility. This approach relies on broadly-stated ethics frameworks intended to moralise market dynamics top-down, to elicit socially responsible corporate behaviour among developers and users of AI platforms and tools.[5] The ethical frameworks/codes designs are intended to engender trust across communities, yet all too often are insufficiently engaging with the pressing perceptions and realities of AI anxiety.[6] A cynical reaction to pushing the transfer of human values into machine technology is that it deflects responsibility for risk or data appropriation from creators, commercialisers and regulators, by generating a smokescreen of agreeable but fuzzy principles that travel no further down the value chain than the boardroom or the ministry.[7] In addition, the recent rehabilitation of 'humans in the loop'[8] appears as both belated and bemusing through the inference that human agency was ever expendable, and without it AI would retain user trust. AI is 'artificial' insofar as the human decision-maker presently selects when it is employed and not vice versa.

---

[3] Comment made by a workshop participant – March 2020.

[4] Kris Hammond breaks down intelligence (and its transference to the AI context) to 'sensing, reasoning and communicating'. He quotes John McCarthy who observed that 'as soon as it works, no one calls it AI anymore'. Hammond K. (The AI Ecosystem', *Computerworld* https://www.computerworld.com/article/2918161/the-ai-ecosystem.html Adopting Hammond's three phases this paper understands the AI ecosystem as an endeavour in which AI professionals create technology and use big data to assist human agency in 'sensing, reasoning and communicating'. In this interpretation the ecosystem also incorporates clients who market the application and customers who employ it in their decision-making.

[5] A highly regarded example of this is the Singapore Government's 'Model Artificial Intelligence Governance Framework' https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf

[6] Much of the literature on AI anxiety focuses on labour force and the future of work. https://essentials.news/en/future-of-work/article/ai-anxiety-ethical-challenge-business-7cb52cb39b

[7] Computer Science Department, University of Oxford. 'Problems with Codes of Ethics' https://www.cs.ox.ac.uk/efai/developing-codes-of-ethics-for-ai/downsides-of-codes-of-ethics/

[8] SETI Institute (2019) 'Keeping Humans in the Loop' https://www.seti.org/podcast/keeping-humans-loop

Two fundamental realisations are emerging around the significance of ethics as a regulatory tool for using AI technologies and big data, particularly in this period of global health crisis.[9] Ethics cannot be expected to replace responsible state, agency, industry and community regulation through harder edged, interventionist strategies that demand compulsory compliance. Examples of this regulatory convergence are where immigration, employment and civil order authorities are requiring strict limitations on citizen movement and association.[10] It should never be considered satisfactory (or particularly trust inducing) to leave major decisions impacting on freedom, identity and civil liberties to self-regulatory compliance formed within voluntary and non-accountable codes of conduct.[11]

Despite its recent criticisms, and limitations (highlighted in the empirical section of the paper) ethics, if applied and evaluated contextually, is an important framework against which crucial AI-assisted decisions are made. However, to give ethics sufficient regulatory bite, political, medical, social, operational and sustainability externalities must be recognised as having an equally significant place as determinants of necessary behaviour. As the later summarised research reveals, if young designers are impacted in what they do by organisational power hierarchies and client contract pressures then the operational influence of ethics may be moderated. Starkly, during the COVID-19 crisis hard triage choices regarding the preferential application of limited medical options in emergency treatment settings demonstrate the situational prominence and the relativity of ethics in complicated medical determinations. Such life and death decisions were dependent on access to technology and its application, but not these factors alone.[12]

This paper engages with the ethical regulation of AI at three levels. The first is to generate and share emerging conversations about ethics with AI practitioners and end users[13] lower down the market and production chain so that mutual responsibility in attribution and distribution of ethical considerations in key-decision sites[14] can be evaluated and actuated. As our empirical experience reveals, if ethics is being blind-sided by predictable and recurrent operational pressures and compromises then this information needs feeding into a pragmatic evaluation of the regulatory promises of 'Ethical AI'. That experience also reveals that the general nature, form and

---

[9] For a wider discussion of the issues in this context see Findlay M., (et. Al.) (2020) 'Ethics, AI, Mass Data and Pandemic Challenges; Responsible data use and infrastructure application for surveillance and pre-emptive tracing post crisis'
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3592283

[10] Council of Europe (2020) 'AI and control of covid 19 coronavirus' https://www.coe.int/en/web/artificial-intelligence/ai-and-control-of-covid-19-coronavirus

[11] John Braithwaite expands on this in his enforced self-regulation model. Braithwaite J. (1982) 'Enforced Self-regulation: A new strategy for corporate crime control', *Michigan Law Review* 80/2: 1466 – 1507

[12] Campbell D, Topping A. & Barr C. (2020) 'Virus Patients more likely to Die may have Ventilators taken Away' (1/4/2020), https://www.theguardian.com/society/2020/apr/01/ventilators-may-be-taken-from-stable-coronavirus-patients-for-healthier-ones-bma-says

[13] By 'end user' we here and throughout the paper are referring to those who employ the AI technology or apply associated data in their decision-making environment. These users could include the individual/organisation that commissions a specific AI-tech i.e., the client in the developer-client-user chain while also referring to third parties in that chain such as those who can access the technology or data through open source in any form.

[14] There is a focus on decision theory in this extension of responsibility for ethical behaviours rather than simply on institutional processes or operational factors. The reason for the importance of decision theory rests in the understanding that the human/machine interface when AI is involved, almost always sees AI and big data enhancing human decision-making capacity.

comprehension of principle-structured ethical discourse is often an impediment to spontaneous, engaged and informed conversations around ethics in use case contexts.[15] Next, and embedded within sustainable conversations, is the pressing requirement to interrogate a meaningful language and understanding of ethics across all stages and decision domains of the AI ecosystem, and thereby create the possibility of evaluating ethics as an inclusive regulatory frame in work life experience. For this purpose, ethics, AI, big data and human agency are seen as a communal enterprise. The responsibility for devising, agreeing and applying a relevant ethical language is mutualised throughout the AI ecosystem and on to its varied applications. Finally, the project aims to prioritise a central element within the ethical panoply, that being fairness, and attempt to model ways in which fairness can not only be shared but effectively and influentially directed to AI and big data applications so that human dignity[16] is maintained and maximised. [17]

The structure of the discussion to follow commences with a brief overview of the vision for ethics as an AI and big data moderator. It works from the assumption that risks are not primarily within the province of the machine or the raw data but at the interface between these, human agency and interpretation. Risks are in the decisions that follow, which of course can be exacerbated by bad tech or faulty data, but not justified by them. The nature of human agency/AI integration will differ in its various institutional, social and market settings,[18] even if the exponential capacity for AI and big data to influence decision-making is common. It is the decisions that are made by humans at this interface which must stand ethical scrutiny.

From here we summarise some thematic controversies in the associated literature, to place the 'ethics push' within a more realist understanding. The empirical foundation of the paper is commenced by a brief comparative examination of the top-down, end-user style advocated globally by the Singapore government; and to explain its aspirations and limitations. In the spirit of expanding the potential regulatory reach of ethics across the AI ecosystem, the paper details a 'shared fairness' approach[19], that looks to the universal (not incremental[20]) attribution and distribution of ethical responsibility within the AI ecosystem.

---

[15] A telling example of this is the principle of 'mal-feasance'. Regularly in the focus group and workshop settings this was immediately reduced to 'harm' which is plain language for the intention contained therein and in that form much more accessible.

[16] The paper recognizes that human dignity as a qualifier of life experience ids often diminished because of its subjectivity, even relativity. In times of global crisis, we take the view that meta measures of life experience particularly those that focus on human's lived experience are valuable analytical aspirations.

[17] At this stage of the project's development 'shared fairness' is being extensively theorized. In the empirical experience several opportunities have arisen to gauge the reaction to fairness as a dominant principle, and the extent to which is attribution and distribution is seen as a shared endeavor.

[18] It is necessary in this vein to identify the need for important work to be done in imagining ethics as facilitating at the human/AI interface, AI assisted decision-making which reflects shared and inclusive governance possibilities through early stage transparency and ethical conversations.

[19] It is recognized that fairness may have different understandings depending on where it is located and what it is meant to influence. Procedural, operational or algorithmic fairness can be viewed differently than fair outcomes or participatory fairness.

[20] Meaning that for the 'shared' component of fairness as a consequence of mutualised responsibility, there cannot be individual or sectoral or hierarchical designations of who within the team, the project, the organisation, is the 'fairness' arbiter or insurer. In addition, shared fairness is not a drip-feed experiment. It depends on responsibility first being mutualised, and this is an important theme discussed in the empirical events.

As mentioned above, *mutualised responsibility for ethical behaviour and principled design* is our overarching methodology for the attribution and distribution of ethical obligations. Appetites for distributing responsibility rather than requiring discrete, hierarchical or operational-based individual compliance grow from the project's wider appreciation of the role of AI and big data in communities of use.[21] That said, discrete individual compliance may still be the outcome of a distributed responsibility, the two are not being mutually exclusive. Mutualised responsibility as a methodology for disseminating fairness does not require some dilution to a point where everyone is responsible for everything. Instead, *shared fairness* is what this mutualising of responsibility intends. Mutualising responsibility remains wedded to specific decisions and decision-making sites wherein participants involved are each and all who achieve and enjoy shared fairness, which then radiates along through other inter-connected decision-sites[22].

Fairness is selected as the pre-eminent ethical value in this holistic consideration because it is contextually and operationally specific and it is one of those essential principles which should bind the human/machine interface, about which definitional singularity can be avoided through more intuitive understanding.[23] These understandings are not definitionally-dependent. Rather, in most decisions and their outcomes fairness is a key determinate of legitimacy. In addition, fairness has specific directions, recipients and benefits depending on what decision site is being considered. There will be further discussion of this understanding of 'shared fairness' in later sections.

It might be considered a tautology to talk of mutualised responsibility for shared fairness, but the emphasis on mutualising responsibility and its shared mission cannot be underemphasised as the glue that binds shared fairness. It is helpful in this regard to think of responsibility not as an ethical principle or a value (which some say it is, we say it is not). Instead responsibility is employed here in the sense of 'responsible to do something - responsibility to be fair'. Responsibility is the attribution for achieving a fair outcome, it is the activation of a duty or responsibility to share processes that achieve outcomes which are fair. Once responsibility is attributed communally and distributed across all players in decision sites in the ecosystem, then 'shared fairness' is a primary objective for those who accept attribution (to be responsible for fairness). Our empirical research suggests that while there is a common view throughout the ecosystem that AI applications and big data use should be fair, there is little mutuality of responsibility, and apparently in the most contentious

[21] More useful than investing human values into machines, the paper argues, is understanding the AI/human interface in terms of *village (kampong) values* that may offer for ethics a collective and communitarian narrative, one in which all participants have an investment. It is necessary in this vein to identify the need for important work to be done in imagining ethics as facilitating at the human/AI interface, AI assisted decision-making which reflects shared and inclusive governance possibilities through early stage transparency and ethical conversations.

[22] The project employs decision theory (Parmigian G. & Lurdes Y. (2009) *Decision Theory: Principles and approaches* Chichester: Wiley) in the interconnected context of AI project management and big data use linkages. An operational argument favouring mutualized responsibility is that it reflects the manner in which the AI ecosystem can be reduced to chains of projects passing decisions on data, one to the other. One reason for an interconnected rather than an independent approach to decision theory here is that AI and big data are developed and used on the edges of uncertainty. These edges can be clarified if the creative process resembles a production line of ideas.

[23] Again, this is put with the qualifier that fairness may have different understandings depending on where it is located and what it is meant to influence. Procedural, operational or algorithmic fairness can be viewed differently than for outcomes or participatory fairness. Even so, fairness in its conceptual ubiquity is not dependent on applications or directions. Fair remains fair.

decisions, less prioritising of fairness as something each player should factor against, or even above commercial and market exigencies.

At this point it is appropriate, having plotted ethical attribution and distribution,[24] to pause and reflect on the theoretical purpose of our endeavour: if ascription to ethical principles realistically prevents social harm from the application of AI and the use of big data, this needs to be a holistic enterprise to maximise its regulatory influence across the AI ecosystem. Ethics guidelines, if they only have contained or exclusive sectoral impact in the ecosystem will be limited in their overall effectiveness. Essential to be engaged in this holistic approach are front-line AI professionals and low-level market users.[25] Externalities working against the shared responsibility model require identification and critical interconnection.[26] In order to ground these aspirations the paper will conclude with the mutual responsibility/shared fairness methodology.

Drawing these thoughts together as the world struggles with an emerging pandemic makes it imperative that the analysis to follow has the capability to contribute to making accountable and transparent surveillance and data sharing externalities that will be responsibly facilitated through AI while not posing unnecessary strains on human dignity that 'shared fairness' considerations can help minimise.

Ethics Vision

Joseph Indaimo in his (2015) *The Self, Ethics and Human Rights,* suggests the utility in broadening the cultural context of ethics and AI. Employing Sun Yat Sen's concepts of 'livelihood' and 'universal brotherhood', Indaimo takes up the latter to argue that a western-focused, individualized human rights paradigm does not completely engage with why we aspire for the protection of human dignity. Where does this link with AI and ethics?

The current, largely corporate narrative of AI and ethics has emerged in part at least from a primarily platform provider/data manager concern to rehabilitate the perception of the mega-information holders/users when it comes to their market applications of personal or secondary data. The Cambridge Analytica[27] scandal has left the billion-dollar business of information management far from any moral frame or even active rights recognition, constrained as these may be in most social media contexts.[28] The trust and confidence once vested by essential public *data-as-product* information providers requires formal reframing. Enter ethics, against a background of commercial confidence building and consumer risk aversion.

---

[24] Orr W. & Davis J. (2020) 'Attributions of ethical responsibility by Artificial Intelligence practitioners, Information,' Communication & Society, DOI: 10.1080/1369118X.2020.1713842

[25] Tham I. (2020) 'Singapore's AI Ethics Model needs more Bite' https://www.straitstimes.com/opinion/singapores-ai-ethics-model-needs-more-bite

[26] It became apparent in several of our workshop exchanges that participants, particularly designers, engineers and technicians felt oppressed by profit/contract demands which over-rode sufficient consideration and referencing of ethical obligations.

[27] Wylie C. (2019) Mindf*ck: Inside Cambridge Analytica's plot to break the world London: Profile Books.

[28] An observation such as this cannot be developed fully here beyond perhaps flagging the manner in which privacy within social media realms is often more a negotiable, than a claimed right.

However, the application of ethical codes (and efforts at ethical coding through talk about ethics by design) to real problems of perceived risk, and risky behaviours has progressed in a less than critical crusade, avoiding some essential questions:

1. Is the language of ethics as a medium for AI best practice being dulled by platitudes and vague definitions?[29] How can ethical language be refined and sharpened so it provides a more empirical measure of best practice compliance?
2. Should ethics be culturally relative (or sensitive), particularly when it comes to context specific business practice and institutional behaviours?
3. *Does ethics by design* have it the wrong way around? Should the ethical focus shift from creating tools and technologies to ensure that decision-making outcomes are ethical, to creating applied ethical frameworks that require these tools and technologies operate under ethical agendas in the first place? How can this be achieved outside some abstract push for injecting human values into technology?
4. How can a more definitive approach to the language of ethics operationalise ethical applications so they are seen by innovators as facilitative rather than restrictive?
5. In advancing regulation and governance through ethical codes how do we avoid capture by algorithmic black box elitism?
6. What is the most market/social-effective way of locating applied ethical frames at the decision-making interface between man and machine?

Much of the discussion about ethical AI seems to be constrained by the intention to invest machine intelligence with human values. If you unpack the regularly rehearsed human values in these codes, such as transparency, accountability, explainability etc., they appear to rely on individual rational decision-making and are clearly compatible with issues of individual self-interest. In addition, when ranked by frequency of use, the principles move from the more applicable (transparency, justice and fairness, non-malfeasance) to the communal and more oblique (trust, sustainability, dignity, solidarity).[30] It might even be observed of this progression that it descends progressively to indivisible and more contentious grounds. Some commentators on this 'principles in AI' approach criticise their generality, conflicts in practice, subjectivity of interpretation, and tensions caused through mutual exclusivity.[31]

Moving away from individualised attribution of ethical responsibility, it is unnecessary to geographically or culturally locate *village (kampong) values*[32] that offer for ethics attribution a collective and communitarian narrative, one in which all participants have an investment. It is

---

[29] Mittelstadt, B. (2019). AI Ethics – Too Principled to Fail? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3391293

[30] Jobin A., Ineca M. and Vayena E. (2019) 'The Global Landscape of AI Ethics Guidelines', *Nat Mach Intell* 1: 389-399.

[31] Whittlestone J. (et al) (2019) 'The Role and Limits of Principles in AI Ethics: Towards a focus on tensions', Cambridge: Association for the Advancement of Artificial Intelligence.

[32] For the purposes of this research *kampong* is a cultural shorthand for village spirit, wherein benefits are shared, and business is an etiquette of inclusion. In kampong thinking, the world around us is a living, learning institution and new ideas complement that wider wold. The geopolitical and cultural locations for the term are primarily the Malay peninsula (including Singapore, and Peranakan traditions).

sufficient to shift into a communal appreciation of the AI/human interface that imagines ethics as facilitating AI-assisted decision-making which reflects shared and inclusive governance possibilities through early stage transparency and ethical conversations.[33]

We consider 'ethical commensurability'[34] as the way to resolve the tension between relativist and contextual ethical understandings, and universal values. For the analysis to follow, 'fairness' is the common measure of ethical achievement. Commensurability provides an answer to the question of whether ethics—meant to guide principled AI design—can recognise universal values and respect cultural diversity.[35] This answer is cast in the empirical project below, as reflections on 'kampong' community bonding where AI and human agency must exist in a mutually supportive life-space, relying on fair decisions, protecting the vulnerable and working for fair outcomes.

Unity through a common measure of fairness risks oversimplification and tokenism when arguing the importance of contextual sensitivity in ethical understandings. For instance, there is no infallible coherency about 'Asian' culture or philosophy, as there is no single determinant for whatever is meant by western philosophy despite the cannon being rich and diverse. European thinking around ethics, since the enlightenment, is grounded in individualist/libertarian considerations of rights and duties. Much popular political discourse surrounding the notion of 'Asian family values' and 'Asian business culture' distinguishes itself from individualist rights-based approaches to bonds of cultural obligation.[36] Reduced to a workable duality, global west and east, more relativist aspirations for ethics constructions widens the vision of ethics away from the individualist interpretation of rights and duties to consider:

a) Cultural, contextual locations for AI that emphasise communal place and relationships, and

b) Interpretations of ethical principle which are influenced by philosophies and practices that prioritise communal obligation

---

[33] The conversation methodology is a crucial component of the 'grass-roots' ethics project discussed in the empirical section of the paper.

[34] In its entry 'Comparative Philosophy; Chinese and Western' the Stanford Encyclopedia of Philosophy (https://plato.stanford.edu/entries/comparphil-chiwes/) there is discussion 'ethical commensurability' and how it diverges from virtue ethics. If 'virtues' are seen as discrete and sometimes incompatible (such as transparency and accountability) then a common standard of 'fairness' might be lost.

[35] Alan Chan, founder of oil tanker company Petroships, is a proponent of the Confucian merchant (儒商). He argues that righteousness and benefits are not necessarily opposing elements.
"I gave it some careful thought and I concluded that righteousness is also a benefit, but it is a long-term benefit," he explained at a recent SMU School of Social Sciences (SOSS) seminar, "Confucianism and Business Ethics". Citing a phrase from Confucius' Analects, Virtue is never lonely, it will attract companionship (德不孤, 必有邻), Chan expanded thus: *"You practice righteousness, people will come to you, they will do business with you and support you. If you get support, then you are likely to succeed and you will gain profits. So, righteousness and benefits can be reconciled."* 'The Confucian Merchant', *Perspectives* https://cmp.smu.edu.sg/article/confucian-merchant

[36] 'Comparative Philosophy: Chinese and Western' notes 'Another potential contrast arises from the focus in modern Western moralities on individual rights to liberty and to other goods, where the basis for attributing such rights to persons lies in a moral worth attributed to each individual independently of what conduces to individual's responsibilities to self and others. Confucianism lacks a comparable concept, given its assumption that the ethical life of responsibility to others and individual flourishing are inextricably intertwined (Shun, 2004).'

Accompanying the universalist/relativist debate, another concern with the prevailing iterations of AI and ethics is how they resemble a mimicry of corporate social responsibility (CSR) and claims around corporate self-regulation. As with the *dark side* of CSR, ethics over-reliant on corporate sponsorship,[37] risks becoming an internal management language in danger of normalising otherwise deviant or anti-social corporate cultures and captured within the power frames of corporate networking. Neutralisation through any ethical overlay of the challenges to corporate and market morality could exacerbate and not alleviate these negative cultures which in turn foster AI for purposes and priorities far from social good.

A more inclusive vision for AI ethics does not evolve from positioning ethics in AI governance as a battle between rational individualism and universal brotherhood. Instead, empathetically correlating and co-existing humans and AI in some proactive and productive communal engagement, governed by motivations for behaviour and interaction which are sensitive to, respectful of, and generate social bonding, avoids unrealistically requiring from AI (in its many forms) that it demonstrate or mirror individualist human values.

Literature on 'Ethical AI' development

As already identified on several occasions, to address concerns that the promotion of AI will lead to social harm, many organizations employing these technologies have published high-level principles meant to guide the development of AI tools. Since then, work has emerged tracking and comparing these documents such as Jobin, Ienca and Vayena's study examining 84 of these guidelines and principles. This research team found an emerging convergence around six principles: transparency, justice, fairness, non-maleficence, responsibility and privacy; while also noting substantive differences in their interpretations and methods of implementation.[38] Another such study compared 36 of these documents, recording a similar consensus around eight trends: privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values[39]. There remains room for looking behind these league tables and exploring the reasons for priority and whether these connect with degrees of take-up and operational relevance.[40]

Amid the widespread approval and adoption of these principled approaches, a notable line of critique has been the disproportionate role of industry actors in their crafting and promotion. Private companies like Google, Microsoft, IBM and Tencent have taken the lead in publishing their own ethics documents and principles.[41] Nonetheless, these companies operate in highly competitive markets and, as some have argued, it is ill-conceived to expect that they 'can be trusted to abide by

---

[37] Ochigame, R. (2019, December 20). The Invention of "Ethical AI": How Big Tech Manipulates Academia to Avoid Regulation. *The Intercept*. https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/

[38] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399.

[39] Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI* (SSRN Scholarly Paper ID 3518482). Social Science Research Network.

[40] In the method of our project we present one of these tables to participants and explore comprehension and relevance, priority and potential operational impact.

[41] Jobin (2019); Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*.

voluntary standards when faced with such powerful commercial imperatives'.[42] These big platform providers, technology giants and mass data users are intent on shaping the debates around highly controversial technologies that they are developing and marketing. Scholars have noted that industry actors are also promoting 'Ethical AI' as a form of principled self-regulation, which then functions as an alternative to legislation and other harder-edged regulatory intervention.[43]

Researchers and commentators have further questioned such voluntary codes as a form of 'ethics washing', which remains a significant challenge to the wider legitimacy of these codes and their principles. Along with this masking function there has been constant reference in critical commentary to conflicts of interest. Many of the promoting companies advocating ethics self-regulation as development and application risk moderator are also at the forefront of developing state-of-the-art AI technologies[44] and incorporating these technologies into both their services and operations.[45] The murky overlap between developer/user and self-regulator demand evaluation of likely contradictions in incentives. To answer such concerns, industry alliances with powerful consolidated messages are asserting a commonality of ethical imperatives to address cut-throat market risk taking. Organisations like the Partnership in AI, bringing together an impressive consortium of companies like Amazon, Apple, Baidu, Facebook, and Google, advance neutral scientific goals of conducting research and sharing insights across market rivalries. Nonetheless they also function as a public validation exercise by suggesting that shared ethical proscriptions will prevail in self-interested, competitive markets. This message, of ethics over profit and collaboration over market advantage, is persuasive in a regulatory climate otherwise not excited by sharp regulatory technologies. It has been argued that such industry giants 'highlight their membership in such associations whenever the notion of serious commitments to legal regulation and business activities need to be stifled.'[46]

More recent regulatory research encourages moving towards making these principles actionable as one might expect duties and obligations in any private law context: shifting, as some have phrased it, from the 'what of AI ethics', to the 'how'.[47] This proposition has resonance where most of the information platforms rely on some form of data-subject consent in the usage agreement.[48]

Despite a broad and high-level consensus around ethical principles, commentators have nonetheless observed that we are yet to witness an similar ethical transition in the design of algorithmic systems

---

[42] Yeung, K., Andrew Howes, & Pogrebna, G. (2019). *AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing* (SSRN Scholarly Paper ID 3435011). Social Science Research Network. This is not to say that in some areas such as the adoption and promotion of facial recognition technologies, some of these companies will take decisions that protect their 'ethical' reputation, in the face of market disadvantage in the short term.

[43] Ochigame (2019); Hagendorff (2020); Yeung et al. (2019)

[44] Rei, M. (2020). ML and NLP Publications in 2019. https://www.marekrei.com/blog/ml-and-nlp-publications-in-2019/

[45] Mirhoseini, A., Pham, H., Le, Q. V., Steiner, B., Larsen, R., Zhou, Y., Kumar, N., Norouzi, M., Bengio, S., & Dean, J. (2017). Device Placement Optimization with Reinforcement Learning. *ArXiv:1706.04972 [Cs]*.

[46] Hagendorff (2020)

[47] Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). 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*.

[48] NB. Facebook's condition on posting photographic material, that the member consent to the platform accessing the phone's media files.

(as evidenced in the algorithmic fairness discussion) despite an emerging literature on technical tools and methods for addressing common ethical challenges.[49] There are two hypotheses for this slippage. The first is that the high-level and abstract nature of AI ethics principles makes it difficult for technicians and designers to use them in their daily activities[50,51] (a matter which is addressed in the paper's empirical considerations), particularly when these activities may themselves be equally oblique but in a different level of technical abstraction. The other suggestion is that there has been insufficient cross-fertilisation between ethical regulatory research in academia on the one hand, and real-life application with developers on the other.[52] For instance, one study found that AI developers, while aware of the ethical challenges in their work, were not organisationally supported with adequate tools or methods for addressing them as they went about their work life in rarefied technical contexts.[53]

This gap between the availability of tools and awareness of them remains one of the key challenges for shifting ethics and principled design into operational concerns. But it is not simply about improving the transition of ethics into product design. More than this is the need for the 'humans in the loop' to agree that ethics has operational advantage and as such it is as important a project requirement as any other. This endeavour is central for the project explained to follow. Work incorporating AI developers own perspectives in the ethics debate remains currently limited and relatively underexplored. That said, a number of studies have emerged in the recent years dedicated to incorporating the voices of AI practitioners. Veale, Van Kleek and Binns interviewed public sector machine learning practitioners working across five countries to understand how they were putting considerations of fairness and accountability into their everyday practices.[54] Holstein and his colleagues similarly sought to reveal the challenges that private sector machine learning practitioners faced when monitoring for bias and fairness, for considering their operational needs in ethical compliance.[55] In another study, Orr and Davis sought to understand how practitioners distributed responsibility across the design of their AI systems, thus focusing in the personal and collective perspectives of practitioners to highlight where they saw themselves (in responsibility terms) regarding other stakeholders in the AI ecosystem[56].

From the research so far mentioned, it is apparent that the development of AI ethics has thus far proceeded with insufficient input from AI practitioners themselves. This group is nonetheless essential to the development of AI products, and, in some instances, the very same group is (or is not) choosing and applying ethical tools in their projects and prioritising the language in their project

---

[49] For an overview of these tools, see Morley et al. (2019)
[50] Mittelstadt (2019)
[51] Holstein, K., Vaughan, J. W., Daumé III, H., Dudík, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 1–16.
[52] Morley et al. (2019)
[53] Vakkuri, V., Kemell, K. K., Kultanen, J., Siponen, M., & Abrahamsson, P. (2019). Ethically aligned design of autonomous systems: Industry viewpoint and an empirical study. *arXiv preprint arXiv:1906.07946*.
[54] Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–14.
[55] Holstein et al. (2019).
[56] Orr, W., & Davis, J. L. (2020). Attributions of ethical responsibility by Artificial Intelligence practitioners. *Information, Communication & Society*, 1–17.

design and its security. One recent line of commentary has argued for the inclusion of AI developers within the strategy-development process for principled design. Madaio and his colleagues, for example, iteratively co-designed a fairness-focused checklist together with AI practitioners – thus enabling them to understand both practitioners *needs* as well as the overall efficacy of such checklists within the wider organisational structures of companies.[57] The approach has two distinct advantages when promoting holistic ethical engagement: first, it addresses the current gap in perspectives from developers on the ground and approaches the discussion from an understanding of what they need; second, like Orr and Davis's work, it helpfully situates the discussion of ethics and responsibility beyond that of a single individual. This line of research findings informs our empirical approach by validating the need for front-line inclusion to achieve a more holistic approach to ethical governance throughout the AI ecosystem, and it assists our argument for the operational importance of mutualised responsibility.

Our approach to ethics as an operational and inclusive language as well as a normative regulatory frame requiring shared responsibility resonates with Habermas's discourse ethics paradigm[58]. Habermas argues that norms emerge from rational-critical deliberation: an inclusive process where opposing views are shared, and parties take part in a reasoned, reflexive, and coercion-free dialogue which ends with an agreeable decision. In any such 'conversation', openness (and as we highlight later, moderating organisational power impediments) is essential if a safe space for mediated decision-making outcomes is to be possible. It would seem from the top-down, set in stone approach to AI ethics broadcasting there has been little internal debate and discourse negotiation. For our purposes the Habermasian model of optimum engagement is the mirror to reveal what does not seem to be happening in the AI ethics transposition, for most companies and state agencies who support principled ethics frames. There are also examples of this discourse of negotiation and meaning sharing in the preparation of best practice guidelines when codes of conduct are struck to encourage internal industry regulatory compliance.[59] There have been assertions from the major advocates of ethics regulation in AI that the development of their principles have been road-tested within the company culture.[60] However, this appears to be more a validation process than any genuine debate about what should or should not stand as an ethical motivator.

At present Top-down ethical principles tend to replace the role of language and communication in operational decision-making. Ethics as a discourse, as a process of communication, and informing decision-making, can act more sustainably and convincingly for identifying and scrutinising embedded assumptions and norms.

---

[57] Madaio, M. A., Stark, L., Vaughan, J. W., & Wallach, H. (2020). *Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI*. 20.

[58] Habermas, J., (1991). *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. MIT press.

[59] Braithwaite, (1982)

[60] Johnson K. (2020) 'Google Researchers Release Audit Framework to Close AI Accountability Gap', https://venturebeat.com/2020/01/30/google-researchers-release-audit-framework-to-close-ai-accountability-gap/

Another important theme in the literature is the role of 'virtue ethics'[61]. Shannon Vallor in her (2016) *Technology and the Virtues*[62], looking to Aristotelian, Confucian, and Buddhist virtue ethics, notes that all three share conceptions of: 'the highest human good,' 'moral virtues as cultivated states of character, manifested by those exemplary persons,' a 'practical path to moral self-cultivation,' and an conception 'of what human beings are generally like'. Vallor incorporates these cultural comparisons in her argument in favour of 'a global technomoral virtue ethic'. For a culturally sensitive holistic approach to succeed in better embedding ethical regulation it will need to 'resonate broadly enough to motivate significant social cooperation on a global scale'[63]. Tokenistic, or irredentist deference to cultural sensitivity in the distinctly globalised world of AI is unrealistic and may indeed only go to further fragmenting and making more regressive any top-down approaches to ethical injections. The need for this 'technomoral virtue ethic' to function at a global level, Vallor emphasizes, is key as 'no one on the planet today is fully insulated from the failures of human beings to jointly and wisely deliberate about the collective impact of their actions'[64]. The catastrophic emergence of this current global pandemic is tragic evidence not only of disastrously incautious interconnectedness, but conversely the inevitable need for deep empathy and selfless communalism in seeing the virus run its course.

In the next section we describe a top-down, end-user approach to ethics regulation which has received much acclaim and could be said to emerge from the virtue ethics tradition. Equally this approach may emphasize managerial engagement[65] but is yet to permeate down to other levels of validation and application. We take on just this challenge in the empirical section.

Singapore's Approach to Ethical AI: the Model AI Governance Framework

Singapore's state-sponsored approach to 'Ethical AI' has been led in large part by the country's data protection agency, the Personal Data Protection Commission (PDPC). Following the release of a discussion paper by the PDPC in June 2018, the agency promoted Singapore's Model AI Governance Framework (the Framework) in January 2019.

In January 2020, the PDPC updated this with version two of the Framework[66], releasing it along with an Implementation and Self-assessment Guide for Organisations (ISAGO)[67]—which functions as a checklist for organisations to assess their current governance practices—and a compendium of use

---

[61] Singerland E. (2012) 'Virtue Ethics, the Analects and the Problem of Commensurability', https://eslingerland.arts.ubc.ca/files/2013/01/JRE29.1.pdf. The tension between universal moral codes in virtue ethics and commensurability is a constant argument in moral philosophy, as is the struggle between contingent social identity and a characterless moral self. For our purposes, as with the approach taken to decision theory, it is preferred that in the AI ecosystem individuality isn't something that a participant *has* and then choses relationships that suit it. In their working life the AI professional's individuality is determined by dependencies and interdependencies and responses to these. Therefore, with the introduction of universal principles we agree that excellence measured against these is role specific.

[62] Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.

[63] Ibid, p.52.

[64] Ibid, p.53.

[65] It would seem that there are concerns from the framework promoters that even the higher level of engagement and transition into the organisation's governance frame, is not as active as had been anticipated.

[66] PDPC. (2020). Model Artificial Governance Framework (Second Edition). http://go.gov.sg/AI-gov-MF-2

[67] World Economic Forum. (2020). Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organisations. http://go.gov.sg/ISAGO

cases[68] that collects how local and international organisations have aligned their internal AI governance mechanisms with those suggested by the Framework. This revised and expanded version was meant to demonstrate the methodology of industry engagement, where partner organisations and corporations worked with the Framework and detailed how it is incorporated in the life of their individual ethical guidelines.

The Framework is said to be "accountability-based", voluntary in take-up and compliance, and helping to "frame discussions around harnessing AI in a responsible way". In doing so, it provides a structure to help "translate ethical principles into pragmatic measures that businesses can adapt"[69]. According to the Framework, decisions made by systems that use AI should be "explainable, transparent, and fair" and all AI solutions should be "human-centric" (i.e., taking into account the interests, well-being, and safety of human beings)[70]. The document identifies four key points of intervention in an organisation's AI deployment process where specific practices might be made to operationalise these principles: its overall governance structure, level of human oversight in an AI-assisted decision, operations management, and stakeholder communication. There is a notable absence of project-based conversations advocated in the method set out later in this paper, but this will always be a problem with universal ethics agendas that gravitate top-down and are not designed with sectoral flexibility as a central principle.

Similar patterns, discussed in previous sections, can be identified in Singapore's approach to ethical AI. As a largely government-driven framework, it remains an open question as to how well its underlying rationale and guiding principles of fairness, transparency, and explainability have been understood by AI developers. Our conversations suggest reserved uptake, particularly interestingly among government organisations and entities, impedes the voluntary compliance approach. The challenge of operational relevance for project teams is exacerbated by the reality that the document remains directed at organisations as a whole but not at individuals within them who occupy positions carrying the lion's share of decision-making capacities. Higher order managerial positions are intended to institute the framework's governance strategies, adjusting them to "ensure robust oversight over an organisation's use of AI"[71]. While this is an understandable focus in a paternalistic hierarchical implementation mode, one potential consequence, as alluded to above, is that this approach isolates its audience and has limited immediate operational application for other no less significant parties involved in the deployment chain, including engineers and developers building AI-augmented software, as well as end users. The newly introduced stimulus for take up such as a 'trust mark' certification system, and graduated penalties for non-compliance still remain top-down injections

Ethics and Applied Research

Ethics is an interdisciplinary space, welcoming reflections on regulation and governance that are not narrowly legal. While the scope of ethics as a research terrain is encompassing for its many technological applications, the recent 'surge' to intersect AI and big data usage with ethics has gained a somewhat-unruly momentum by not being required to work within operational,

[68] PDPC. (2020). Compendium of use cases: Practical Illustrations of the Model AI Governance Framework. http://go.gov.sg/AI-gov-use-cases
[69] PDPC. (2019) A Proposed Model Artificial Intelligence Governance, p.i.
[70] PDPC. (2020) p.15.
[71] PDPC. (2020) p.21, 3.5.

contractual and market contextual boundaries. In this sense it is not so much that central players are ignoring ethics where AI applications are concerned, but rather that ethics initiatives which fail to recognise project priorities and pressures, wider socio-political contexts, as well as organisational and institutional cultures, cannot be an adequate regulating force. Admittedly, reservations when applying ethics to human and now machine behaviours is not novel, but perhaps something that gets presently passed over in the feverish discourse promoting principled AI is how it has been unable in any applied sense to keep evaluated pace with the scientific explosion of AI technologies. National strategies and corporate commitments largely have tied their regulatory future to one trajectory in the hope that the reservations attendant on AI roll-out can be reined in without slowing the economic imperatives behind the AI revolution. If this singular approach to regulation and governance is to be credible against robust critique, it needs road-testing beyond the incubators of the major information platforms and their corporate language. It is not enough, from an empirical standpoint, to confine the evaluation of ethical governance within the organisational environments so publicly dedicated to its advancement.

If the dominant work model prevailing in AI ethics discourse is a top-down guideline approach, our research identifies the necessity to parallel a *grass-roots* approach to AI ethics offering research subjects and participants the opportunity to tailor-make their aspirations for an ethics conversation, and thereby 'own' its outcomes. Some of the issues that such conversations could confront include:

- Are ethical principles expressed in a way that makes sense?
- If a decision raises an ethical challenge, who's responsibility is it to solve that challenge?
- What worries AI developers about ethical applications?
- What original ideas do developers have about ethical applications?
- When it comes to ethical considerations, how wide is the gap between what consumers/clients expect developers to consider, and what is actually addressed on drawing-boards and innovation incubators?
- Are the explanations consumers expect context specific and if so, how is this so?
- How are ethical considerations complicated through data protection regimes?
- What are the working assumptions about ethical arrangements and obligations on which AI innovators and developers proceed?
- How can these assumptions be better reconciled with the interests of others in their community of users?
- In what ways are ethical responsibility and market profit compatible?
- How do externalities like tight profit margins, contract obligations and client timetables weigh on the time and opportunity for ethical considerations?
- How can ethics be prioritised in project planning?

Contextualised decision-making and Shared Fairness – Project methodology

The aim for the research initiative explained below is to examine the way practitioners (particularly young millennial designers), team leaders, and project security advisors attribute and distribute (or do not) ethical responsibility for the AI application they help create. Aligned with this is a need to know the way they see responsible ethical subjects and core ethical decision-making in their work. Pragmatically it is also necessary to understand how the webs of deflection are created so that these issues become someone else's responsibility. At these levels and with these insights in view, the project adds to emerging literature that seeks to both understand how practitioners are approaching

the issue of 'ethical AI' and include them into the development of operational ethical practices and principled design[72].

The virtue ethics[73] top-down codification models do not encourage practitioners to think about how they materialize ethics through the artefacts they create, but rather they reflect and amplify prevailing structures of power and inequality in large tech development organisations and platforms, as well as smaller tight-knit consultancy ventures, or government AI development agencies. The Silicon Valley ethics approach critiqued above, does not explore sufficiently how ethical responsibility takes shape and falsely imputes 'blame' for the failure to embody a vague set of accountability norms on an undetermined corporate calculus.

Many of the central and oft-recalled ethical standards (such as accountability, to who/ for what?) lack focus. If AI gives material form to social practices and processes, for instance algorithmic decision-making processes entail a degree of commercial secrecy and mathematical bewilderment, then uncovering algorithmic authorship and ensuring technical transparency for the purposes of accountability is rarely sufficient or feasible to pinpoint attributions of ethical accountability. The 'Shared Fairness' project is interested to assess practitioners perceived ethical responsibilities over key value-laden operational decisions, to identify when these decisions arise and what ethical challenges they confront, and to converse in a language of ethics and responsibility which enables practitioners to internalize ethical responsibility. In directing research attention to practitioners (designers and technicians) it is intended to evaluate and educate ethical potential at an essential operational level, regarding the complex and applied anatomy of AI. Ethics, if it is to be confidently relied on as an active regulating frame, should influence each important decision in the relational construction of AI systems. Along with the practitioner focus, the 'shared fairness' project intends in future to include post-production stakeholders such as users and policy analysts. The methodology bridges responsibility imbalances that rest in decision-making power and technical knowledge, by commencing with personal, facilitated conversations designed to return the ethical discourse to those meant to give it meaning at the sharp end of the ecosystem. By attending to practitioners, the project will better understand ethics as a socio-technical practice, working out from the assumption that as a realistic force in regulation, ethics are dynamic, evolving and interdependent.

It is important for a realist research mission to appreciate *contexts of concern* wherein the folding of algorithms and AI into so many aspects of our lives require understandings as social and market *systems* rather than only talking about responsibilised technologies. In offering research locations which test the regulatory relevance of ethics at many significant stages of the AI/human agency interface, this research is not satisfied with operational outcomes alone. A broader recognition of ethics applications to the AI/human interface across the ecosystem is possible via initiating conversations in many different decision-sites, and thereby revealing whether ethics is or is not a dynamic influence on the social context of AI, its purposes, problems and probabilities.

The research agenda grows from a grass-roots exercise first addressing mundane challenges to responsible machine behaviours managed and manipulated by human agency with identified ethical obligations. Altruism is tempered by market/social needs and ethics is, consequentially, invested with operational clout by better recognising market requirements in settings for the advancement of

[72] Holstein et al. (2019; Orr and Davis (2020); Veale et al. (2018).
[73] Vallor (2016)

social good. Once recognised as counter narratives to the importance of ethics these requirements can be understood and confronted.

Holistic Ethics Project – Research methodology

The initial and foundation method advanced is one *of conversations*: about roles in the creation and use of AI and big data; about whether the AI 'language' makes sense for this operational experience, what is confusing, what seems to be a priority, whether it is just management-speak; about the challenges which arise at particular decision points; and about the creation of a support base with a tailored language for ethics and AI that resonates in project planning, project teams, evaluation exercises, and varied work/life experiences across the whole of the ecosystem.

In this way, the method does not intend to evaluate ethical compliance or to question professional competencies. The project is not an opportunity for organisational management to refine their training agendas or reflect on their ethics compliance expectations, although these outcomes could eventuate once our work is internalised within the participating groups and entities. Our role is to initiate, facilitate and make sustainable conversations in which front-line practitioners, team managers and security/compliance professionals can have their say, express their confusions, work with us to identify challenges, and participate in a process of holistic problem-solving. The project team see themselves as a regulatory resource, offering a safe space for interrogating ethics and principled design primarily within the context of what we express as mutual responsibility for 'shared fairness'.

The initial phase of the project consisted of focus groups and discussion workshops with ecosystem demographics of young designers in a major multi-national technology giant, private consultancy operatives and consultants to industry, and major state-sponsored AI technology and application developers.[74]

The first step in the conversation format is to share among the group each participant's experience in working with AI and big data. This is an exercise in self-reflection and with the facilitation of the moderator, an opportunity to build trust in the personal value of the conversation. It is interesting to see how candour develops as the conversation unfolds.
Following on from the sharing exercise, the conversation moves on to discussing the gaps between how participants conceive of ethics as opposed to how ethics is being presented by the emerging regulatory frame of 'Ethical AI', or more specifically through the training and governance operations within their organisations. At this point, having personalised where AI and big data are important to design work, participants are confronted with some ethics compendia and particular principles are discussed for their meaning and interconnections, and how these should be prioritised in project contexts is debated. In this stage of the conversation it is intended to expose formal virtue ethics to the work experience of the members and their attitudes to applicability, relativity and relevance.

[74] Due to confidentiality undertakings we are not in a position to identify participants, participant organisations, dates of focus group meetings or numbers involved. Suffice to say that consistent with research expectations for focus group coverage we are confident that the scope and professional demographics of participants are adequate for pilot observations. The pilot has extended from November 2019 to April 2020.

The conversation then moves on to explore whether participants feel a sense of responsibility for principled design. Out of this consideration of mutual responsibility on a project basis, the conversation progresses to exploring when ethical challenges may arise in the development of applications and software, or the use of data in progressing their role in a team project, employing some hypotheticals. These exercises act as discussion points about routes for action/reasons for inaction when it comes to ethical challenges and associated problem solving. In terms of ethical principles, they deal with issues such as bias, data integrity, robustness, and accountability/transparency, with a prevailing and unifying focus on fairness. Finally we will talk through what 'language' might make ethical regulation relevant at the front-line of development and use these ideas to offer support in building mutual responsibility for shared fairness.

To date we have road tested this interactive format[75] and it has revealed:

- Confusion about who has responsibility for ethical practice
- Market pressures that are personally felt to reduce the time for thinking about these issues
- Genuine interest in understanding the relationship between ethical standards and principled design
- Uncertainty about whether and to what extent where they sit in the chain of development experiences ethical challenges
- Importance of 'fairness' as a central ethical priority
- Inaccessibility of ethical language and the actioning of principles
- Need for help in identifying ethical challenges and structuring solutions
- Importance of a 'language' that makes operational and social sense

Even though some similarities regarding the pressures surrounding ethical decision-making have recurred in all our conversations to date, there are also clearly different priorities for different organisations (and those working within them) depending on their market positioning, financial security, and institutional complexity. Large multi-nationals can make reputational decisions on an ethical basis which may reduce their market share in the short term, whereas smaller companies, consultancies, and start-ups much more influenced by tight profit margins and tough competition, may not have such flexibility. Bigger organisations may have designated staff and training capacity to advance ethical principles as work practice, but smaller operations will have to engage with ethics in a much more sporadic and crisis-oriented fashion. A more detailed investigation to identify and measure such organisationally relative market pressures will be important if the project is to explain the nature, sources and extent of the differences in ethical engagement between say start-ups, MNCs, and state-sponsored agencies, and assist in countering these impediments in specific development planning.

The project expected and—through its pilot phase—confirmed that these phases offer assistance to AI professionals in applying ethical/principled design to enhance not only regulatory compliance but also operational and market advantage, and as a result may be likely to improve the confidence of customers and clients in future ideas because the developments can be represented as ethically

[75] Due to the intervention of the covid-19 pandemic and attendant restrictions on association and movement, the personalised interaction of the focus group has been modified by slightly more moderated online conversations.

interrogated and imbued. Overall the goal of the method has been to understand AI practitioners' needs and to find a "third space" where their operational language can be married with that of 'AI principles' swirling around in their work life. It would be too grand and perhaps away from our critical agnosticism to intend that such 'ethical thinking' can be developed singularly through conversation/dialogue among otherwise silent ecosystem voices. It is, however, a good place to start, and one so far unfortunately bypassed, if the intention of ethical regulation is to encourage mutualised responsibility for principles such as shared fairness across the ecosystem.

In summary the project frame involves:

1. Thesis: When ethics are applied to an AI ecosystem this is usually a top-down exercise. For ethical considerations to provide a facilitating and encompassing regulatory impact they should be introduced at a foundation/operational level, so that the relevance of responsible AI applications is viewed like any other operational component in the early life AI development or big data usage.
2. Method: The project revolves around several conversation formats where strategically selected participants are engaged in a phased conversation, at the conclusion of which applied ethical facilitation is offered to address specific ethical challenges and their market/consumer ramifications.
3. Output: At the end of the conversation exercise the project intends to develop a holistic analysis of problems, possibilities and potentials for applied ethics throughout the AI ecosystem.

In focusing on the attribution and distribution of responsibility as a mutualised and communal (Kampong) exercise the project reflects its theoretical predisposition that participants in the AI ecosystem do not essentially approach responsibility purely as an individual obligation but in a more operationally reflective sense, as a mutualised endeavour to achieve 'shared fairness'. If this can be accomplished it will bring with it a better understanding about the way operational ethical applications will render ethics as much more than a regressive compliance frame.

Tangentially, the project challenges an 'ethics by design' approach in treating applied ethical application not as a toolkit for ethical decision-making, but rather as any other operational measure which judges, through the responsible use of AI, how potential market/consumer preferencing can be enhanced (among other operational concerns). In this way ethics moves from normative considerations to operational components across key decision-sites in the ecosystem.

Reflections on Method Building

The entire conversation methodology is dependent on the importance of a bottom up ethics initiative in a context where professionals—particularly 'rank and file' workers—have limited control over what they are developing in terms of the extent to which they adopt responsibility for ethical operationality.[76] The power differentials which characterise most commercial organisations, such as those in which these professionals work, militates against empowering all ecosystem 'citizens' to attribute responsibility when it is distributed to them (voluntarily or through some compliance

[76] This sense of hierarchical dependency has been confirmed in the empirical experience to date.

frame). There are two practical conditions which must be achieved for the methodology to produce a mutualised responsibility dynamic with 'shared fairness' outcomes:

1. *buy-in from top and middle managers*. A potential problem in gaining and maintaining managerial buy in is that managers and leaders are reluctant to give up power, or to mutualise their claimed power even if it results in a more effective and pragmatic distribution of ethical responsibility; thus, it is
2. *important to develop and offer an outcome for management to be interested in a freer flow of ethics and ethical reputation such as an attraction for the best and most committed young minds*. One outcome which has become apparent in our work with managers is the realisation that the more representative and inclusive is the ethical framework of a company or organisation, the easier it will be to recruit the best AI professionals, who themselves value working in an organisation with a high and tested ethical profile in the market

In order to insure against the insidious disruption of power differentials, it is necessary in parallel with the project method to ask what middle managers want from their enabling of the conversations to take place. This forensic of where organisations' power structures currently stand and how, within their protected hierarchies, the organisational policy speaks to address and value ethics and governance, needs to be understood in terms of the life experience of the less powerful. Perhaps a diagnostic from this ancillary investigation of power impediments to ethics free flow will be the experience of the conversations feeding into training protocols for how to better the activation and sustainability of ethics through the organisation.[77]

Another approach to the challenge of organisational acquiescence is to allow limited management participation in the conversation format, so that they can appreciate first-hand the experiences of their junior staff. It is clear, however, this comes at a cost to less inhibited conversation up and down power hierarchies. Controlling who is in the room remains a challenge even without management participation, even by focusing on teams and projects. These sub-hierarchies all have leaders, and followers, and power dynamic externalities. This is where, through generating trust, a conversational ethic will emerge that has indicated on many occasions the surfacing of honest reflection over towing a corporate line.

Finally, the down-side of a methodology working out of institutional case-studies, is that it limits empirical standardisation. In determining the demographics of our focus group populations, the project had no concern for ecosystem representativeness, and as such any potential to speak about broad application across ecosystem decision-sites is speculative, at this stage.

[77] The process of encouraging interest, generating support and gaining approval for the workshop method has been one of the most intuitive and time-consuming components in project's ground-laying. In one instance it was the possible impact of the conversation experience on internal training protocols which won the day.

Appendix 1 – Random Reflections from the Focus Groups/Workshops[78]

- …there is too much talk about ethics, all organisations have an ethics code the same way that they have a motto — but when it comes to ethics and AI, the starting point should be about data — should we be masking certain attributes when profiling to avoid unfairness and harm?
- it seems that people are mostly unconcerned about ethical breaches. But when it comes to combining big data, for everyone there is a line where they ask, 'Why are we doing this?' — but that line is quite far off and sometimes it is very grey
- ethics is mostly reactionary, an after-thought, everything that Google has done to react to ethical challenges has been after influence of the GDPR — it really is about defining ethical data use, e.g., is it okay to consider gender when looking at data for credit evaluations? How can we set boundaries for ethical data use?
- ethics is very much principle based, and differs geographically but also shifts due to profit considerations or competition in particular market settings — but also, computer scientists are not sufficiently learning ethics or secure coding in their curriculum, and most of the technical training is focused on efficiency — recently companies are expected to have ESG reports, perhaps something similar should happen for ethical AI?
- people at the work face want to be ethical, and companies know that an ethical reputation will attract the best workers.
- data is about awareness, but the definition of harm is so difficult — defining boundaries of use is so important — too caught up in harm to humans, what if we changed this to acknowledging harm to humanity?
- regarding the details of these principles and how they are understood: 'if you were talking to R&D developers then they would know, because their peers might strike them down.'
- on explainability and transparency: experience working on explanations for automated decisions — ultimately explained through examples: a compromise between what is inside the model — not really maths, but how the decisions is determined by its inputs.
- the principles are dependent on market contexts - removing competition will companies be ethical? Most likely not, competition prompts ethics.
- on bias: look at sustainability, corporations must have a sustainability report, can there be something like that for AI Ethics? Something that is trackable. Look at the recent case of Apple's Card and its credit scoring algorithm, data issues here were due to biased data.
- data privacy issues are the challenge here, we currently have too much of an opt out system right now, we should have an opt in system instead. We are told that our data is being shared with third parties, but you need to know what kind of third parties and for what purposes.'

[78] Due to confidentiality undertakings, and a commitment that the output from the conversation format should be 'owned' as much as possible by participants, the following restructured views and observations are offered as an insight into information accumulated to date. The material was selected (with no thematic agenda in mind) from workshops/focus groups conducted in early 2020. The participants ranged from young designers, technicians, project administrators, team leaders, client services personnel, trainers, and management representatives.

- on shared responsibility: Post-GDPR, when getting client data, a data governance council was involved so that the data team were liable — we need platforms/forums for discussion and raising these issues. Another issue is: what happens when developers are external to your organisation?'
- Responding to the notion that data cannot be audited because it can't be explained: 'this is unprofessional.'
- On shifts in company culture 'Yes, start-ups in particular usually operate in an extreme market environment — idea is to solve everything now and quickly — only once they've grown to a size that they cannot side-step the question of ethics (reputation) do they start to address this — previously the idea of 'we are just a tool' cannot stand up to scrutiny and reputation damage from a loss of trust — had to intervene in communications between the platform and its users'
- Have thought about ethical issues a lot and agree that ethics is important – hardest thing is that there are so many angles to tackle when considering the impact and application of ethics, and it is difficult to prioritise one when others are also important e.g., privacy issues with data use, issues with bias – all of which have been outlined by the slides shared.
- Also, issues around what problems you choose to tackle in the first place e.g., is it ethical to work on stock portfolio optimization when there are medical problems that need to be addressed, a lot of compute and money going into the former, so I think that one of the steps that's necessary in driving change in terms of how the world interacts with AI research is pinpointing critical ethical concerns that affect real people in terms of outcomes of the AI projects – and then focusing specifically on those because it's easy to get bogged down by the different aspects of the problem
- In terms of day to day work, the work is very removed from outside effects. E.g., just building a model might not involve thinking about ethical concerns because it isn't something that has ethical effects – so it isn't pressing to think about ethics at that stage.
- But on a contract and a project for another company, the number one thing that is driving choices is meeting terms of the contract, and there is little room for thinking about ethics, we're not breaking rules, but it boils down to meeting deadlines and getting things ready – rushed – as long as nothing flagrant you do what needs to get done.
- So, there is a little bit of a disconnect there. Obviously if I saw something that had ethical challenges – it varies from person to person – there was one issue where one of the protocols for how to transfer data was being violated and I did raise concerns, but it ended up that everyone including those that drafted those rules still did carried on anyway, because there was no feasible way of doing the data transfer that otherwise matched the contract. People on both sides of the contract were when the protocol are not being upheld. It was not about leaking personal data. But these things happen, and even people drafting documents and who set the rules do things beyond the rules, if push comes to shove, often they look past it.
- in terms of discussion about the relationship between humans and machines, one of the reasons why AI is a big ethical issue is the idea of community, risk and fear. One of the problems in terms of selling AI and big data for the wider community, is that people who are not informed about AI are apprehensive about decision choices being taken away from them. That introduces a clear issue for the discussion of ethics. How to introduce the client/citizen into the use process as early as possible so they have a say – that is a challenge

particularly when technicians have unique training and don't have time to spend with outsiders. There would be far less public resistance to visual recognition if the common person was involved earlier in the process, so aware of benefits and parties of the ways in which we can minimize risk. So, inclusion is important.

- Trust and transparency – important but difficult to determine.
- Another thing that contributes to bias – the limited base data that clients have. Often, they want solutions like acoustic or visual recording, but they have not kept data up until now, so they have only just started to collect data. And this is another source that creates bias in the data in certain cases.
- That's why it's important to make clients understand machine learning is not magical, it is a process done on both sides: the solution provider, and the clients as well. So that is a process and a journey rather than having an accurate result from the first point of implementation.
- Tough to treat these ethical principles all at once - the shakiest ones in are dignity and solidarity. Very tricky/unsure about what they mean and who they relate to. Dignity = not exposing personal information? More for the sake of it being embarrassing for them, distinct from privacy? But similar with privacy with many more social connotations or something like that. So perhaps subsumed into privacy. If you have privacy do you have dignity as well?
- Also, a strong relationship between justice and fairness and beneficence. See a distinction, but definite overlap. Main difference in that two parties should be treated identically from objective POV; but how do you define social good, which is much more a society related concept rather than the concept of consistency. [79]
- Tough call, but not having consistency is slightly worse. At least knowing what to expect out of the model and application of the AI is good because you are an outside observer; you can at least predict and understand what the process is doing and address it. Whereas if the models are making decisions that have a huge impact on everyone – then it's hard.
- Also, there could be a distinction between developers and the AI solutions expected by the clients. Between clients and users.
    - E.g., transparency: transparency is between the solution providers and the client. But justice, privacy, and trust, might be between client and the users of the client – e.g., smart cities hosted by a company, but we are the solution provider, so transparency is between us and the client. But justice and fairness relate to the impact to people in society. So, it might be made clearer. [80]
- If the researchers were the ones who decided what they want to work on, they would not be working on the things they are working on. Not to say researchers have no control; researchers are spoiled compared to other divisions in development environments. But

[79] Observer comment; Here if there is understandable confusion surrounding 'social good' and a distinction is made between human benefit and operational consistency (which is perfectly arguable), we need to have a clearer understanding of the connection between good operational practice and human benefit.

[80] Observer comment; Another important comment on the way in which relationships and expectations can give different attributions to different ethical principles depending on perspective.

there is a huge balance between money making and actual topics of interest to the betterment of society.[81]

- Most of the time ethics and profit don't align with each other and would take a pretty big change – medicine and environment and energy optimization would be the biggest ones for AI to tackle, most are not incentivized heavily. Energy usage is a little bit e.g., for data centers, but most of the time those tasks are not the ones AI professionals end up working on. How do you make companies prioritise social good in a real operational sense?
- It's going to take some crazy political revolution before those things get aligned.
- But in terms of usage, by design it's difficult to look from a perspective of when AI professionals create the solution in the first place. It is easier/more comprehensive if coming from the use-case perspective e.g., use-case is on automotive companies and purpose is detecting defects, and monitoring defects. Professionals know the purpose of the solution. But if the use-case is a country that wants to spy, we know it isn't right, so that is more straightforward to look at use-cases rather than from the technology side.[82]
- For governance to be put in place, so requiring "responsibility" from the platform operator, there is potential liability. Each party needs to be protected such that data is not being used, beyond what users know, so the breach is on the side that is committing the wrong action.
- Openness with the client is a challenge
- AI professionals always need to come back to the client's side of how they currently conduct their processes, otherwise it becomes garbage in and out, with is dangerous with biased data. Having clients that are open to the professional is key [83]
- Clients understand that whatever data provides teaches the model; open communication with client on their provided data is important
- An individual data privacy issue here too – utopian scenario where in order for your data to be used, the individual about who the data is concerned would have to opt in otherwise the data could not have been transferred in the first place. But that doesn't solve the problem because even if personal data is not being used, the individual can still be profiled by the model constructed on other people's data.
- Ownership and privacy aren't the fundamental issues here[84]
- Is this more of a contract policy? There are strict guidelines for drafting contracts – prohibitions on passing that data to third party data brokers. The default is that the professional explicitly agrees not to be able to use any insights for purposes other than what was built for the client.[85]

---

[81] Observer comment; This is a challenge for management to ensure that ethics and social good are not sectioned out of areas where profit and client satisfaction dominate.

[82] Observer comment; the use-case applications approach is indeed more targeted and critical. If we take this approach does it mean the application of universal principles to all situations is unrealistic?

[83] Observer comment; perhaps openness is best ensured where the client and the designer are required to abide by the same ethical standards, and not have different standards subservient to the conditions of a contract. Ethics embedded in AI contracting is a key issue

[84] Observer comment; this is no doubt correct in a total sense. If it is not about privacy and data ownership, then how do we direct ethical involvement in data use where the data object has no say?

[85] Observer comment; if we are understanding this correctly some companies have specific guidelines for the way in which they contract BUT can these have influence on the way the client contracts beyond an agreement of terms?

- Trust – from the designer and to the corporation as a whole. Designer might not see this. Tech is dual-use, and the moment that people use tech not in a good way, their corporation can step in and decide whether or not to use something.

Preliminary Observations

Moving past the identification of ethical challenges
How might one intervene once they see something that might be problematic? It's clear that the challenge is not simply identifying problematic uses of data or the ethical quandary, but rather what do we (as individuals and teams) do to help people understand both why that challenge occurs (in the context of production, in the context of their own organisation) and what they might do once they see something that should be questioned.

When it came to suggesting solutions, participants mentioned not using data presented, or correcting for bias within the dataset, or removing key attributes like gender/ethnicity. No one brought up the use of existing toolkits such as IBM's AI Fairness tools or Google's What-If Tool, both of which were launched with relative fanfare. If one way of addressing ethical challenges is to look at possible points of intervention — during data collection, during model development, during testing — it would be useful to know whether these toolkits are on practitioners' radars, and whether they're seen as a viable option for companies to audit their models.

Expanding the notion of 'mutualised responsibility'
If, again, it is less a question of identifying ethical challenges than having safe space to engage with them, then this goes back to considering whether the issue is more about developing mutualised responsibility within teams building an AI-enabled product/service.

What we're seeing is that int is not so important to focus on what ethical principles a company might have in their ethics documents, but what routes exist within the organisation to enable those certain principles to be realised — and conversely what gaps or blockages (internal and external to the service provision) prevent this. For developers, their position/seniority/status/longevity/value within the hierarchy of the organisation might be made explicit as a problem for ethical distribution, attribution and assertion. Is there a risk for their job security by wanting to bring ethics into the conversation? Is that possible to generalise when some organisations value ethical reputation more than others? No doubt there is a need to have platforms and forums discussing and identifying challenges for ethics so that concerns might be raised safely, openly and without repercussions such as perceived disloyalty to the team. How do we address developers' individual responsibility and their specific company cultures?

All of this speaks to transparency within the organisation itself — perhaps lateral and vertical processes of accountability and pressure-valves are required for ethical identification preceding the more dramatic whistleblowing policies? How does one decide who should sound the alarm up to hierarchy of the organisation? Mutualised responsibility would collectivise that obligation and decision.

Aspirations for transparency can break down between organisations when they share design activities, also connecting the end user —being less the case where a single organisation is involved in data collection, cleaning, labelling, augmentation, classification, and profiling. If so then what are the lines of communication between organisations when a data set gets handled off, or when a

model is built and sold to a user? What would be the reach of mutualised responsibility be in that scenario and how could it be policed? If responsibility stretches to the client and beyond what binds the party in a shared fairness ethic after the product/data has been produced and monetised in the market? This is an important consideration for the influence of ethics in determining social good.

Tension between production pressures and reputation-preserving tactics

This was a recurrent implication in the discussions without often being directly addressed other than, "If so, I'm not sure what to do". For example, the profit motive/competition motive/contract commitment leads to companies taking dubious decisions, or the start-up environment that tends to be extreme in a 'move fast and break things' way. And what about the commercial/operational value of the connection between ethical practice and reputation as an individual and organisational benefit?