Routledge
Taylor & Francis Group

Information, Communication & Society

ISSN: 1369-118X (Print) 1468-4462 (Online) Journal homepage: https://www.tandfonline.com/loi/rics20

Attributions of ethical responsibility by Artificial Intelligence practitioners

Will Orr & Jenny L. Davis

To cite this article: Will Orr & Jenny L. Davis (2020): Attributions of ethical responsibility by Artificial Intelligence practitioners, Information, Communication & Society, DOI: 10.1080/1369118X.2020.1713842

To link to this article: https://doi.org/10.1080/1369118X.2020.1713842

Published online: 26 Jan 2020.

Submit your article to this journal

Article views: 134

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=rics20

INFORMATION, COMMUNICATION & SOCIETY
https://doi.org/10.1080/1369118X.2020.1713842

Attributions of ethical responsibility by Artificial Intelligence practitioners

Will Orr and Jenny L. Davis

School of Sociology, The Australian National University, Canberra, Australia

ABSTRACT

Systems based on Artificial Intelligence (AI) are increasingly normalized as part of work, leisure, and governance in contemporary societies. Although ethics in AI has received significant attention, it remains unclear where the burden of responsibility lies. Through twenty-one interviews with AI practitioners in Australia, this research seeks to understand how ethical attributions figure into the professional imagination. As institutionally embedded technical experts, AI practitioners act as a connective tissue linking the range of actors that come in contact with, and have effects upon, AI products and services. Findings highlight that practitioners distribute ethical responsibility across a range of actors and factors, reserving a portion of responsibility for themselves, albeit constrained. Characterized by imbalances of decision-making power and technical expertise, practitioners position themselves as mediators between powerful bodies that set parameters for production; users who engage with products once they leave the proverbial workbench; and AI systems that evolve and develop beyond practitioner control. Distributing responsibility throughout complex sociotechnical networks, practitioners preclude simple attributions of accountability for the social effects of AI. This indicates that AI ethics are not the purview of any singular player but instead, derive from collectivities that require critical guidance and oversight at all stages of conception, production, distribution, and use.

ARTICLE HISTORY
Received 2 September 2019
Accepted 21 December 2019

KEYWORDS
Artificial intelligence (AI); accountability; AI ethics; inequality; professions; organizations

In April 2018, 3,000 Google employees issued an open letter to CEO Sundar Pichai. The letter condemned Google’s involvement with the Pentagon in developing Artificial Intelligence (AI) for targeted predator drone technology. Declaring that Google ‘should not be in the business of war,’ the letter became a successful yet isolated campaign obstructing the development of autonomous technologies on ethical grounds (“Letter to Google C.E.O.,” 2018). The case represents a neat and tidy anomaly in the world of AI ethics. A multinational company sought to produce an ethically questionable product, the employees evaluated that decision as unethical, engaged in collective action, and the project halted. Rarely are ethical decisions so straightforward, nor corporations so readily responsive.

CONTACT Jenny L. Davis ✉ Jennifer.Davis@anu.edu.au School of Sociology, The Australian National University, Canberra ACT 0200, Australia

This article has been republished with minor changes. These changes do not impact the academic content of the article.

© 2020 Informa UK Limited, trading as Taylor & Francis Group

Beyond its swift resolution, what stands out about the Google case is its focus on responsibility and accountability. This letter was a call for companies to recognize their own ethical responsibilities for the technologies they create, rather than outsourcing ethical concerns to those involved in products' implementation or use. That is, accountability was firmly and clearly placed, and it rested with Google Inc.

Issues of responsibility and accountability are crucial for the future of AI. Artificial Intelligence, in particular data-driven Machine Learning (ML) systems, have become embedded in most major institutions and are increasingly present as part of everyday life, with profound social effects in the near term and further down the line. What those effects are, however, remains uncertain. The apparent inevitability of AI's continued growth, paired with broad public anxiety about its social implications, have prompted a flurry of attention and channelling of resources towards understanding AI systems, predicting their effects, and devising social and technical interventions to minimize harm and optimize social good. Yet, questions persist about who will be held to account.

Technical systems pass through multiple hands over the trajectory of conception, design, implementation, and use. Myriad actors and organizations come in contact with a given AI product, and each has formative effects upon it. It remains unclear who the stewards of these technologies are, and where the burden of social responsibility lies.

Through interviews with AI practitioners ($N = 21$), we take up the problem of ethical responsibility. Reflecting the heterogeneity of the AI field, we define 'practitioner' broadly, referring to the professionals who develop AI products and services (Holstein, Vaughan, Daume, Dudik, & Wallach, 2019). Despite a quickly growing literature on AI (and AI ethics in particular), practitioners have received surprisingly scant attention. Our research is thus motivated by the following question: *how do AI practitioners attribute and distribute ethical responsibility for AI systems?*

Ethics are the value-based principles of right and wrong that guide behavioral decisions.[1] The meaning and standard of ethics in AI remains contested, polyvalent, and in flux (Ananny, 2016). For example, deontological orientations adhere to rule-based ethical frameworks (Murphy & Woods, 2009), consequentialist positions assess ethical systems by their outcomes (Awad et al., 2018) and sociologists argue for an AI ethics that can overcome structural inequalities (Greene, Hoffmann, & Stark, 2019). Rather than impose these (or other) *a priori* definitions, we are instead interested in ethics as part of practitioners' in-situ processes. Practitioner's tacit definitions speak not to any objective ethical ideal, but reveal how ethics manifest into procedures, decisions, and material goods, and the related responsibilities these manifestations entail.

The purpose of this project is to elucidate norms of ethical accountability within the AI sector. Those whom we interviewed are all, in some capacity, building 'smart' machines. These technologies – like many emergent technologies – will have profound and unknowable effects. Should the effects be detrimental, adjustments will be required to rectify and mitigate harm. Norms of accountability will affect adjustment capacities by either providing a clear map of responsible subjects and core decisions, or alternatively, creating webs of deflection in which both problems and solutions remain someone else's burden.

The fraught and ambiguous nature of ethics-in-practice came through across interview subjects, as did complex strategies of attributing and distributing responsibility for design decisions. The practitioners with whom we spoke reached for external conventions, codifications, and figures of authority as guides for ethical decision-making. At the same time, practitioners expressed personal moral prerogatives and recognized the autonomy bestowed upon them through their technical expertise. What emerged was an entangled understanding of ethical accountability, entrenched within conflicting perceptions of stakeholder interests and responsibilities. In what follows, we trace the ways that AI practitioners act as constrained autonomous agents, deferring judgment to extrinsic bodies while actively translating broad mandates into tangible products that will go out into the world.

AI as materialized action

Rooted in the critical tradition of science and technology studies (STS), we begin with the assumption that technological artifacts, among them AI systems, give material form to social processes and practices. This recognizes the complex interrelation of technological artifacts and the social relations that they emerge from and are embedded within. In particular, we follow Schraube's (2009) materialized action approach. Schraube (2009) poses that technologies are imbued with human subjectivity. The values of designers, engineers, and corporate bodies materialize through the artifacts they create. Yet, the direct and flow-on effects of an artifact remain unknown and unknowable.[2] Artifacts embody human subjectivity, and then take shape in myriad ways through encounters with diverse user-publics. This perspective grounds artifacts in social subjectivities and at the same time, infuses artifacts with a liveliness all their own. In this way, people and technologies – or subjects and objects – are mutually constitutive.

The notion of human-technology co-constitution is integral to mainstream perspectives in STS. Most eminent among these perspectives is actor network theory (ANT), which situates humans and technologies in actor-network assemblages (Callon & Law, 1997; Latour, 1988; Latour, 2005; Latour & Porter, 1993; Sayes, 2014). ANT and related theories posit that people and technologies collaborate to mutually shape and affect each other. A materialized action approach adheres to this configuration. What distinguishes a materialized action approach, however, is its assumption of subject-object asymmetry. Schraube (2009) positions humans (subjects) as responsible parties, even in the face of considerable technological (objects) forces.

This assumption of asymmetry provides a critical analytic lens, centralizing politics and power in sociotechnical systems. While technologies can have immense shaping effects, their outcomes ultimately trace back to human subjects. Such a lens is well suited for the driving question of the present project, which addresses how AI practitioners distribute and attribute responsibility. The materialized action approach foregrounds specific and difficult questions about *which* humans are responsible, *how,* and *under what circumstances?* These questions of responsibility zoom out from moralizing investigations of what is (or is not) ethical in the construction of intelligent systems and instead, explore how ethical accountability takes shape.

Social dynamics of AI systems

Social accounts of AI have sought to identify the values embodied within these systems. Repeatedly, studies show that in-built values most often reflect and amplify prevailing structures of power and inequality (Angwin, Larson, Mattu, & Kirchner, 2016; Buolamwini, 2017; Eubanks, 2017; Lambrecht & Tucker, 2019; West, Whittaker, & Crawford, 2019). Trained on existing human data, AI confound and enact biases that pervade the societies from which they derive. Through an inequality of access and outcome, intelligent systems become stealthy 'weapons of math destruction' (O'Neil, 2016) – opaque autonomous algorithms that encode historical power relations. Understanding the culture and practices through which value laden decisions materialize in the design of intelligent systems, and perceptions of ethical responsibility therein, thus remains a critical project.

Developing AI systems is at once a product of developer choices and at the same time, cloaked in layers of imperviousness. Whether they are subject to intellectual property restrictions, require a high degree of computational literacy, or are merely unknowable beyond their inputs and outputs, algorithmic decision-making processes entail a degree of secrecy and bewilderment (Burrell, 2016; Mackenzie, 2018; Pasquale, 2015). Technology is, therefore, developed through a '*predicament of mutual opacity*' (Bruun Jensen, 2010, p. 72); no single practitioner may understand every technical detail or be present at every conversation during the creation of a system. This is exacerbated by the complex interrelation of actors that constitute a team of designers – with processes, practices, and perspectives that culminate in a 'loosely coordinated confusion' (Seaver, 2017, pp. 3–4). As such, distilling agency from this obscurity, and attributing accountability to any one author, actor, or stakeholder, are prohibitively intricate tasks (Reddy, Cakici, & Ballestero, 2019).

Indeed, uncovering algorithmic authorship and ensuring technical transparency is rarely sufficient nor feasible to elucidate attributions of ethical accountability. Blame does not rest easily with designers, users, hardware, or code, but rather, somewhere in the spaces between (Shank, DeSanti, & Maninger, 2019). These complexities have prompted a philosophical reformulation of agency and the disentanglement of morality from responsibility, recognizing that autonomous technological systems are neither value-neutral nor blameless (Floridi & Sanders, 2004). Many STS formulations of accountability have thus moved beyond attributions of blame and authorship towards an interrelation of social, legal, ethical, and disciplinary norms (Ananny & Crawford, 2018; Neyland, 2016; Reddy et al., 2019). By assessing practitioners' perceived ethical responsibilities over key value-laden decisions, our research interrogates the professional culture of those who build intelligent machines.

Although many studies work to identify and mitigate biases and unfairness in AI systems (Chouldechova, 2017; Dwork, Hardt, Pitassi, Reingold, & Zemel, 2012; Feldman, Friedler, Moeller, Scheidegger, & Venkatasubramanian, 2015), the practitioners behind these systems' creation and the processes through which these systems materialize, remains under explored (Holstein et al., 2019). The perspectives of those on the ground, making intelligent machines, is a crucial piece of the puzzle within a broader effort to understand and implement ethical AI. Two recent works, each addressing AI practitioners, provide key insight into this important dimension. Holstein et al. (2019) assess the practical needs of private sector AI practitioners, while Veale, Van Kleek, and Binns (2018) ask public sector AI practitioners what is fair, and what challenges they face achieving fairness standards. Together, these works illuminate the experiences of AI professionals and the meanings of fairness that inform production processes.

Following Holstein et al. (2019) and Veale et al. (2018), the present work focuses on AI practitioners. While Holstein and colleagues assess practitioner needs and Veale and colleagues address practitioner values, we work to understand practitioners' perspectives on ethical accountability. We seek to identify how practitioners conceptualize norms of accountability in the design of AI systems, and specifically, who is deemed responsible for defining and ensuring a system's ethical character or deciding which formulations of fairness will be embodied within it. Consequently, this research looks beyond the practitioners themselves to the perceived roles and responsibilities of stakeholders involved in the full 'anatomy' of AI (Crawford & Joler, 2018). The path towards ethical AI is one of immense complexity; in accordance with Holstein et al. (2019) and Veale et al. (2018), it is crucial that practitioners' practical needs, realities, and perceptions remain central.

Methods

Twenty-one practitioners operating within Australia agreed to be interviewed for the purpose of this research. In line with existing studies of the AI profession, we define practitioner broadly as those whose professional practice includes the development of AI products and services (Holstein et al., 2019). Given this broad definition, our sampling frame includes participants from the private (seven participants) and public sectors (seven participants), as well as academics (seven participants) working in AI research. This cross-sector inclusion criteria distinguishes our work from similar studies which, despite broad definitions of 'practitioner,' focus exclusively on private (Holstein et al., 2019) or public sector (Veale et al., 2018) professionals.

Bolstering the value of cross-sector sampling, the practitioners we interviewed rarely fit cleanly within any one category (academia, private sector, public sector). Instead, participants moved between sectors, with concurrent or past positions in more than one. Further, their work often included cross-sector collaborations such that industry, government, and academic practitioners maintained regular contact with each other, and sometimes occupied positions on the same project teams. Practitioners across sectors are thus aware of, and subject to, a shared set of understandings and normative conventions, including complex trade-offs between various stakeholder interests. With that said, each participant in our sample did identify primarily with one of the three sectors, and each sector maintains distinct norms, practices, and demands. In short, professional sectors operate not as independent silos nor as a homogenous block, but as interrelated constituent parts of a differentiated whole. Our sampling frame reflects this.

The field of AI encompasses a varied range of technologies, processes, and applications (Russell & Norvig, 2003). Participants exemplified this heterogeneity, working on systems for financial broker monitoring, governmental chatbots, self-driving cars, autonomous weapons, and bespoke systems that serve specific client needs as they arise. Despite this heterogeneity in professional focus, the sample remains predominately male (~75%) and white (~85%), reflecting a demographic homogeneity that plagues the field of AI, and STEM fields more broadly (i.e., the 'white guy problem') (Crawford, 2016).[3] To address this, we grant particular attention to the voices of underrepresented participants, as they offer a standpoint unavailable to those in the majority (Harding, 2004; Harding, Grewal, Kaplan, & Wiegman, 2008).

We conducted a total of twenty-one interviews. Of those, sixteen took place face-to-face in a location selected by each participant and five took place via video or phone call to accommodate geographic distance. Interviews lasted between thirty-one and eighty-three minutes with an average length of forty-eight minutes. Given the sensitive nature of the subject matter, coupled with legal and professional proprietary restrictions, participants were often reluctant to share specific information about their systems, practices, and organizations. Reflecting similar concerns and challenges faced by Veale et al. (2018) and Holstein et al. (2019), participants often sought anonymity to avoid professional repercussions for their organizations and/or themselves. For this reason, each respondent is anonymized in our presentation of the findings (R1 – R21) and professional and personal details remain intentionally vague. Due to the small numbers of racial minorities and women in the sample, gender and racial markers are removed to avoid identification. To provide context within these constraints, the sectors to which participants belong are retained (R1 – R7 from academia, R8 – R14 from the public sector, R15 – R21 from the private sector).

Interviews examined the specificities of each practitioner's work, the decision-making process through which systems are constructed and integrated within society, and participants' own perceived place within the production process. Following a semi-structured approach, we asked participants about their work, the systems that they create, and the projects in which they have been involved. Particular attention was placed on their routine throughout the lifecycle of each project. Following this, we asked participants to identify how considerations of fairness and ethics figure into their own design processes. We also asked participants about ethical issues faced by the AI community more broadly. These broader discussions had two effects: they allowed for the inclusion of topics not previously mentioned in the specific case-based inquiries, and provided participants with space to speak freely about their concerns and what they had observed from within the industry itself. Because we are interested in understanding the norms, beliefs, and practices of industry professionals rather than imposing any particular ethical standard or agenda, we intentionally did not define ethics for participants during the interviews. Instead, we relied on participants' tacit ethical principles and the values and behaviors entailed therein.

Following Kitchin's (2017) methodological assessment of critical algorithmic studies, our interviews sought to uncover the relational construction of AI systems (Burrell, 2016; Pasquale, 2015). Consequently, we do not focus on the specificities of a system that may be literally and metaphorically locked within the laboratory, nor the precise practices that occur within the creation of a particular system. Instead, the personal mediation of these details and events by each participant provides an entry point into the cultural world of AI design.

Each interview was audio-recorded, transcribed verbatim, and coded by the co-authors using abductive analysis, which moves between theory and data through a mix of deductive and inductive strategies (Tavory & Timmermans, 2014). An abductive approach assumes that existing theoretical frames can productively organize empirical data, while data will necessarily speak back – and often challenge – established theoretical perspectives.[4] Guided by critical theories of STS and allowing for novel insights from participant narratives, we identified all places in which practitioners spoke about ethics, morals, and crises of conscience. Using thematic analysis, we organized these passages into an initial set of broad themes. We then combined themes where appropriate and created sub-categories as they emerged. We replicated this process until all relevant passages fit into an existing (sub)category and no further categories could be identified (Braun & Clarke, 2006; Charmaz, 2014).

Coding was an iterative process in which the authors moved between collaborative and individual engagement with the data. First, both authors collaboratively identified general points of relevance. Each author then independently coded the data to generate broad preliminary themes, attached to exemplar passages from participant interviews. The authors then collaboratively solidified a set of thematics, with which the first author coded the entire data corpus. The second author 'spot checked' the coding with sample passages from each theme. These spot-checks showed full consensus between both coders. These processes of triangulation, peer-examination, and reflexivity bolstered analytic rigor and 'trustworthiness' of the findings (Krefting, 1991).

Findings

All practitioners we interviewed acknowledged ethics as a critical element of their work. They were keenly aware of social concerns about the effects of AI and familiar with cases of ethical failure within the sector (e.g., racial bias in the COMPAS parole system, Uber's self-driving car fatality, robo-debt calls in Australia). The problem of ethics in AI was clear to participants. Less clear for them – and for us – was how ethical responsibility does, and *should*, distribute. Certainly, practitioners are involved in the process of design and implementation. Indeed, they arguably represent the most obvious site of accountability – they are the ones who make the products. However, practitioners are but single nodes in a multiplex of actors and factors. We thus sought to understand how practitioners fit themselves within broader webs of relations, and to what effect upon accountability norms. What emerged from participants was a pattern of ethical dispersion: powerful bodies set the parameters, practitioners translate these parameters into tangible hardware and software, and then relinquish control to users and machines, which together foster myriad and unknowable outcomes.

Interviews reveal a relationship between those that set parameters (legislators, organizations, clients) and those who implement them (practitioners) as defined by two imbalances: one of power and the other of technical expertise. Participants felt bound by the expectations, mandates, interests, and goals of more powerful bodies. At the same time, practitioners have technical knowledge which those who commission (and often oversee) their work, do not. Thus, practitioners *cannot* act with full discretion, yet *must* exhibit independent efficacy. In turn, participants recognized the continued life of a product once it leaves the workbench. In the hands of agentic users and mutable by design, AI products remain always in process. Rather than singular attributions, practitioners evoked and moved between, a web of responsible parties.

Pre-set parameters

Participants understood their practice as extrinsically bound. They articulated their work as translational, both driven and constrained by the proclivities and prerogatives of more powerful actors. Participants placed the locus of responsibility for parameter setting in three main sources: legislative regulations, organizational norms, and clients who commission practitioners' work.

Legislative regulations

Practitioners cited legal regulations as a base stricture on design processes. For many practitioners, this was due to the perceived objectivity of the law and its operation beyond organizational and project specificities. Focusing on regulation, practitioners sidestepped subjective questions about ethics in favour of objective matters of compliance, thus shifting responsibility to policy makers and legislative authorities. As one participant (R13) highlighted:

> Usually as an organization using data, your number one priority is, 'is this legal?' And then 'is it ethical?' is so nebulous that it's nice to think about, but you are always focused on, 'am I going to break the law?'

Giving primacy to legal mandates renders ethical considerations a relative luxury – something 'nice to think about,' but ultimately subservient to the formal codes and regulations in place.

Even when practitioners merged ethical considerations with legal ones, the ethics remained a secondary concern. As participant R16 explained:

> The very minimum that you have to adhere to is the law. So, we start by ensuring that everything that we do, or our clients do is legal. Then we have to decide whether or not it is appropriate, which could be considered ethical or fair.

Relying first and foremost on legal codifications offloads responsibility from individuals and from organizations. Legal mandates unburden those in charge from the delicate task of ethical justification, as articulated by participant R19:

> Having these regulations in place gives company CEOs the impetus and authority to say, 'ok I will take that moral philosophy on-board and I don't need to defend it to my shareholders, it is the law'.

Legal codes thus become primary sites of responsibility, with ethics (i.e., what is 'fair', 'appropriate', and 'right') confined within existing legal boundaries.

Organizational norms

Although organizations and individual practitioners are compelled to comply with regulatory mandates, the regulations in place are often relatively loose, ill-defined, and uncomprehensive. This is because the field of AI (and technology development in general) has moved with a speed and multiplicity that regulatory codes have struggled to match while at the same time, organizations and industry professionals have been careful to set their own standards to avoid control at the hands of non-expert forces. For these reasons, the institutional context maintains an essential role in the development of AI ethics. Participants readily pointed to official organizational policies and informal norms as ethical benchmarks for individual practices.

Participant R6, charged with building autonomous weapons, referenced *Just War Theory* as a broad philosophical framework that forms the parameters of practice. The participant explained how this philosophy underpins norms within the field of autonomous weaponry and is entrenched as part of the practitioner's organizational philosophy. This participant's work is thus guided by principles that value human life and avoid life loss, charge states with defence over their citizenry, and justify violence for the protection of innocent subjects (Moseley, 2011). *Just War Theory* embeds weapons construction within an ethical framework and generates parameters within which participant R6, and others in the field, operate.

Participants also brought up company mottos and treatises about ethics. Interestingly, practitioners were often unable to recite these mottos or recall specific language from company documents, but still felt driven and protected by organizational values. For example, participant R18 reported reliance on organizational standards which, during the interview, remained elusive.

> Our company is a highly, highly ethical company … The company's motto … I can't remember what all the acronyms stand for, but that drives the operation … If the business operates in a fair and ethical way, then you need to test that the technology operates within that.

The organization's specific ethics thus emerged less important than an overarching ethical orientation. The organization's presumed ethics instilled confidence in the processes and products produced within it, ostensibly guiding practitioners down ethical pathways.

Clients

None of the practitioners we interviewed operate on their own accord. Rather, they are part of organizations, which are part of market relationships in which labor and expertise are exchanged for various forms of capital. Practitioners thus answer to clients, broadly conceived. In practice, 'client' represents an array of actors. We refer to clients as those who commission and oversee AI projects, but do not do the technical work themselves. Within our dataset, the category of 'client' is populated by three kinds of actors: customers of a business, senior executives of organizations in which a practitioner works, and financial benefactors of academic research. In each of these contexts, client needs and values create ethical parameters and serve as ethical barometers. Participant R16 explained the client's centralized role and related ethical responsibility:

> The clients are in charge, they retain us. We advise them to consider A, B and C and we won't do something illegal on their behalf, but ultimately, it's up to the client.

An imbalance of knowledge (practitioner) and power (client) emerged as practitioners described priorities set by those who commission their work, and the apparent irrelevance for clients of the technical processes by which those priorities materialize. As participant R19 explained:

> … [S]enior executives don't understand machine learning models that their data scientists are producing … here are the parameters and here is what is actually, here is what matters. You have told me to maximize profits so, it really just comes down to [maximizing profit].

The imbalance of knowledge and power created strife for some practitioners, who felt constrained in their own moral and ethical autonomy. Positioning themselves as subservient to clients, practitioners deferred ethical judgement, even as they critiqued the result. Participant R19 continued:

> Data scientists take on what they think is the ethical framework that they are supposed to be operating under, and so it is usually, how do we maximize profits? How do we maximize user engagement, user retention? How do we make this game the most addictive, how do we get more people to sign up? Most clicks. And they are given the task … with no moral framework or support behind that … . So unless there is an absolutely crystal clear direction from senior management in companies about what they actually want to maximize, then we end up with this situation where, nobody really wanted to be supporting the baby munching neo-Nazi

> site, but now all of our ads are there and they maximize our profit, we get amazing clicks … and they are our most loyal members.

Even academics, who are granted greater relative autonomy compared with private and public sector practitioners, spoke of ethical deference in the face of extrinsic research priorities. Not only do academics move between private and public sectors, but they are also dependent on research funding from benefactors with specific agendas. Often, these benefactors are private companies or industry organizations, further blurring the lines between basic academic research and capitalist market relations. Participant R7 articulated these concerns:

> I am very happy to see more and more investment in research, but it is a double-sided sword. Some of the funding comes from [corporate bodies]. So, one has to be very careful about the terms and conditions in these funding projects … what is the purpose of the research and where will the research be used?

As a field, AI practitioners engage in work that serves others' agendas. Participants thus distributed significant ethical responsibility to clients, broadly conceived. Balancing client requests, organizational standards, and legal regulations, practitioners play the part of (highly skilled) technicians, rather than morally autonomous agents.

Practitioners as mediating experts

Despite a range of extrinsic bodies that set ethical parameters for practitioners, participants also recognized their own role in the production and implementation of AI systems. Indeed, practitioners hold technical knowledge that others do not, and their fingerprints are encoded in the systems they produce. While practitioners work within ethical boundaries set by legislative regulations, organizational contexts, and client prerogatives, implementation remains the practitioner's literal job. Stated simply by participant R5:

> There is always someone there who is telling you what to do, not how to do it.

For participants, this often meant implementing a set of directives with autonomy, actively excluding non-technical experts from on-the-ground decisions. Participant R21 explained:

> Accountability doesn't come up in any of our client discussions. It doesn't come up as you would think. It is because they don't understand what they don't understand. How many people will know … in detail how AI algorithms work, and who has actually practiced it to understand the nuances of an AI algorithm?

It also meant that practitioners were guided by technical and practical matters when enacting client and organizational directives, rather than, necessarily, the *best* and most thorough methods. Participant R1 stated this as a matter of time and resources:

> I don't want to spend a week on it, I have only got the afternoon, and I need to make it a bit better frankly.

Participants exercised discretion without client consultation, taking responsibility as technical experts. At the end of the day, practitioners maintain responsibility because their expert knowledge not only allows for this, but requires it. Articulating the burden of inevitable autonomy, practitioners worried over their ethical choices in the design process. Participant R13 expressed this self-critical lens:

> Quite often we will make … trade-offs naively and in line with our own experiences and expectations and fail to understand the implications of those trade-offs for others … We can assess all of the trade-offs, but we still don't weigh them in impartial ways.

Participants perceived themselves as necessary mediators between extrinsic mandates and the products and systems that result. Their expert status makes them indispensable and at the same time, makes a degree of personal control and autonomy unavoidable. For these reasons, participants cast themselves as constrained autonomous agents, encoding others' agendas while owning the ethical responsibility that accompanies technical processes of making.

A life of its own: AI as sociotechnical systems

Participants articulated a dynamic design process with a diverse range of stakeholders, each of which leave an imprint on the final product. However, participants also recognized that production is not the endpoint in the lifecycle of AI. The uses and manifestations of AI take shape through complex interrelations between users and machines (Christin, 2017; Crawford & Joler, 2018). Practitioners, like participant R5, articulated this in terms of AI as sociotechnical systems:

> There is not really such thing as an autonomous agent … it has kind of become important to say … it is now a sociotechnical system, not just a technical system.

Participants referenced two interrelated ethical players that shape a system's ongoing post-production development: users and machines.

Users

User ethics figured into practitioners' accounts in two ways: market forces and behavioral autonomy. Participants assumed that users have ethics and values which will inform their assessments of and engagement with, AI systems. Users would reject systems that they deem unethical and support those which meet cultural ethical standards. In turn, users may engage technologies in myriad, sometimes unexpected and creative ways.

Participants discussed the relevance of user ethics as a market force shaping the industry. They reported that AI systems which do not attend to users' ethical concerns will likely dissipate, while those which adhere to users' values will proliferate. Framing the issue in terms of trust, participant R17 summarized the market perspective:

> [Companies] that aren't transparent or ethical, eventually, or you would hope, end up being prosecuted or sued or you know, all citizens as a whole would choose not to engage with them because they've been identified as an untrustworthy organization … Because, trust becomes the currency on which we trade upon. And will be more so as AI embeds itself in everything that we do.

Practitioners also recognized users as autonomous agents. Encounters between users and AI systems cannot be entirely controlled through design. Responsibility for the outcomes of AI systems are thus partially placed with the users who engage them. Evoking a classic example, participant R20 analogized the ethics of AI to the ethics of gun violence:

> We were a technology provider, so we didn't make those decisions … It is the same as someone who builds guns for a living. You provide the gun to the guy who shoots it and kills someone in the army, but you just did your job and you made the tool.

While the radical constructionism of 'guns don't kill people' has long been critiqued by STS scholars (and is antithetical to Schraube's materialized action approach), it maintains purchase in the professional imagination as a mechanism of dispersion for ethical accountability.

Machines

Participants recognized machines as responsible parties, changing through unexpected (and unknowable) progressions. 'Intelligent' technologies are mutable by design, defined by these technologies' capacity to learn and grow. While humans remain 'in the loop,' AI has unpredictability in its programming.

Attributing responsibility to AI systems, practitioners compared machine errors to those made by human actors. If both human and machine are intelligent, participants argued, each maintain the potential for miscalculation. Referencing a hypothetical (and extreme) case of medical error resulting in death, participant R5 provided an imagined justification:

> I can say, yeah ok that was a fault, but this is how we did the safety analysis. And I can see that this was missed, not because we were negligent, but just because it is so complicated. In this case, somebody died, but we did have the right ethical framework … But sometimes accidents happen. I think that is the kind of argument that you are going to have to make.

Practitioners also recognized the limits of their own foresight and intentionality. They may design a system to serve a particular set of standards and outcomes, but that system can result in various latent effects for which the designer did not account. Of note, practitioners were not blasé about this reality but instead, fretted about unintended effects even as they relinquished personal responsibility and control. Participant R13 highlighted this tension:

> We are developing systems that are better than human … only to discover as time goes on, that maybe they make things worse. And I don't think that that is a cynical thing to say. I think it is just a reflection of how every technology innovation has unfolded so far. What we need to do, as designers, is be aware that we could be designing the system that works and changes people's lives, or you could be designing the system that makes people's lives worse.

Practitioners' fraught position in relation to the systems they produce is compounded by design decisions that serve client needs (e.g., profit maximization, attention optimization) omitting consideration for other, unintended outcomes. Elaborating on the previous point, participant R13 cited Facebook's privacy violations as exemplary of this myopic focus:

> [I]t's not that we thought what we were doing was safe, it's just that, certain inbuilt desires … to increase clicks, to increase attention, to maximise advertising … was our primary motivation. You didn't have to think about any other consequences.

In short, participants recognized the autonomy of both users and machines, placing ethical responsibility in each. As part of sociotechnical systems, AI are (re)molded and shaped through engagements with agentic user-publics. Further, as these systems are autonomous by design, they necessarily do more than reflect encoded priorities, but also learn, grow, and develop with indeterminable results.

Conclusions

Our project began with a simple question: how do AI practitioners attribute and distribute ethical responsibility for AI systems? Through interviews with twenty-one practitioners in diverse fields, we mapped a trajectory of ethics within the professional milieu. Broadly, practitioners distribute responsibility among varied actors and factors. Practitioners are subject to powerful forces that establish the initial parameters, including legal codes, organizational norms, and client mandates. Because of practitioners' unique technical expertise, they are charged with actual production within these pre-set constraints. Here, practitioners cannot escape some degree of responsibility, and thus name themselves as partially accountable parties. Following production, practitioners relinquish responsibility to users and to the machines themselves, both of which maintain an autonomy of their own.

Practitioner accounts of responsibility were largely defined by two interrelated imbalances: that of decision-making power, on the one hand, and technical knowledge, on the other. Participants placed responsibility with powerful players – such as policy makers and organizational bodies – whose mandates for how AI systems should operate create clear strictures upon design and implementation. At the same time, practitioners have a unique skillset and thus maintain an autonomy of necessity, unable to seek authorization for technical decisions that only the technician fully comprehends.

The flow-on effects of technologies after they leave the proverbial workbench once again remove responsibility from practitioners and locate it in the nebulous currents of user practices and machine evolutions. As sociotechnical systems, AI are not just encoded with human values, but continuously develop through encounters with active user-subjects. The effects of AI are never entirely knowable, and participants recognized the ontological limits of design and production.

Findings from this research contribute back to the AI ethics literature in two main ways. First, by attending to practitioners, we can better understand the nature of ethics in sociotechnical practice – how ethics figure into material products, from the perspective of those who make them (Holstein et al., 2019; Veale et al., 2018). Second, we depict ethical processes as dynamic, evolving, and interdependent. This shifts the ethical lens from individual actors to broader networked relations and highlights the need for continued vigilance towards ethical parameters and outcomes. Together, these contributions bolster Schraube's (2009) materialized action approach and map it onto the field of AI: autonomous technologies are imbued with human values, their analyses require a critical lens, and because their effects are unknowable and changing, ethical problems cannot be solved, but only continuously resolved.

Understanding the way AI practitioners treat ethics and perceive norms of accountability provides insight into the ethical character of current and future AI systems. We show that AI ethics cannot be individually located, absolute, or finite. Rather, AI ethics are networked, dynamic, and in flux. Moving away from paternal statements of proscription and prescription, it is more useful to understand ethics as a collaborative process, developed and iteratively (re)configured through material practices and continued negotiations. This flexibility in AI ethics is useful, as it allows for a protean approach, fitted with mechanisms of adjustment for when AI systems move in directions unimagined. Indeed, our data reveal a set of accountability norms which preclude simple calculus about blameworthiness should systems go awry – and given recent history with AI, 'going awry' is near inevitable. Thus, while technological artifacts push back, advocates and industry insiders must ensure human autonomy and accountability are central in the future of technological innovation.

Limitations

This research is not without limitations, and the findings should be read with the following limitations in mind. First, our sample is confined to those who agreed to participate, likely presenting a distinct image from those who refused. Furthermore, practitioners are a single-entry point into the complex world of AI design, and our exclusive focus on this population excluded other equally relevant players (e.g., clients, policy makers, users). Finally, as discussed in the Methods section, our sample was disproportionately white and male. This not only affected the diversity of perspectives our data represent, but also prevented us from contextualizing quotes with participants' demographic characteristics. Excluding demographics was necessary to ensure confidentiality and anonymity, especially for marginalized participants.

These limitations offer a jumping off point for future work. For example, subsequent studies can systematically interrogate who opts-in to ethics-based research and who opts-out. Research will also benefit from engaging the full ecology of actors involved in the production and implementation of AI systems. Finally, recruiting demographically diverse samples will help centralize marginal voices and contextualize practitioner narratives as they intersect with race, class, gender, and other markers of social status. Indeed, a diversified set of accounts will be critical for future research into the ethics of AI.

Notes

1. What is (or is not) ethical is the topic of an entire subfield of philosophy; those debates are outside of our concern here. Rather, we are interested in how AI practitioners understand the ethical landscape and their own role within it.
2. On this point see the extensive literature on technological affordances (e.g. Davis & Chouinard, 2016; Evans, Pearce, Vitak, & Treem, 2017; Gibson, 2014; Nagy & Neff, 2015; Norman, 1988).
3. Although our sample is disproportionately male and white, it surpasses the 16% of women that constitute Australia's STEM-qualified population, and coincides with the ratio of women employed in STEM fields (27%) (Office of the Chief Scientist, 2016). We found no available race data for STEM in Australia, but the predominant whiteness of the field has been well documented in the United States (Funk & Parker, 2018).
4. Using existing theories as a starting point necessarily creates interpretive bounds, thus potentially limiting the full spectrum of readings that researchers might apply to a set of data. We believe this limitation of the abductive approach is outweighed by the rigorous grounding provided by established and well-tested theoretical frameworks.

Acknowledgement

We would like to thank Siobhan Dodds and Hannah Gregory for their contributions to this work.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Will Orr is a postgraduate student at the Australian National University. He studies sociology and data analytics.

Jenny L. Davis is a Senior Lecturer in the School of Sociology at the Australian National University.

References

Ananny, M. (2016). Toward an ethics of algorithms: Convening, observation, probability, and timeliness. *Science, Technology, & Human Values*, *41*(1), 93–117. doi:10.1177/0162243915606523.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989. doi:10.1177/1461444816676645. Retrieved from <Go to ISI>://WOS:000429899100008.

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

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., … Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64. doi:10.1038/s41586-018-0637-6.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101.

Bruun Jensen, C. (2010). Asymmetries of knowledge: Mediated ethnography and ICT for development. *Methodological Innovations Online*, *5*(1), 72–85. doi:10.4256/mio.2010.0011. Retrieved from https://journals.sagepub.com/doi/abs/10.4256/mio.2010.0011

Buolamwini, J. A. (2017). *Gender shades: Intersectional phenotypic and demographic evaluation of face datasets and gender classifiers* (Massachusetts Institute of Technology).

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 1–12. doi:10.1177/2053951715622512.

Callon, M., & Law, J. (1997). After the individual in society: Lessons on collectivity from science, technology and society. *The Canadian Journal of Sociology / Cahiers canadiens de sociologie*, *22*(2), 165–182. doi:10.2307/3341747. Retrieved from http://www.jstor.org/stable/3341747

Charmaz, K. (2014). *Constructing grounded theory*. Thousand Oaks, CA: Sage.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163. doi:10.1089/big.2016.0047. Retrieved from https://www.liebertpub.com/doi/pdf/10.1089/big.2016.0047

Christin, A. (2017). Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society*, *4*(2), 2053951717718855.

Crawford, K. (2016). *Artificial intelligence's white guy problem*. New York: New York Times Company.

Crawford, K., & Joler, V. (2018). Anatomy of an AI system: The Amazon Echo as an anatomical map of human labor, data and planetary resources. *AI Now Institute and Share Lab*.

Davis, J. L., & Chouinard, J. B. (2016). Theorizing affordances: From request to refuse. *Bulletin of Science, Technology & Society*, *36*(4), 241–248. doi:10.1177/0270467617714944. Retrieved from http://journals.sagepub.com/doi/abs/10.1177/0270467617714944

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). *Fairness through awareness*.

Eubanks, V. (2017). *Automating inequality: How high-tech tools profile, police, and punish the poor* (1st ed.). New York, NY: St. Martin's Press.

Evans, S. K., Pearce, K. E., Vitak, J., & Treem, J. W. (2017). Explicating affordances: A conceptual framework for understanding affordances in communication research. *Journal of Computer-Mediated Communication*, *22*(1), 35–52. doi:10.1111/jcc4.12180

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). *Certifying and removing disparate impact*.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, *14*(3), 349–379. doi:10.1023/B:MIND.0000035461.63578.9d. Retrieved from <Go to ISI>://WOS:000222799900004.

Funk, C., & Parker, K. (2018). *Women and men in STEM often at odds over workplace equity*. Retrieved from https://www.pewsocialtrends.org/2018/01/09/diversity-in-the-stem-workforce-varies-widely-across-jobs/

Gibson, J. (2014). *The ecological approach to visual perception: Classic edition*. New York: Psychology Press.

Greene, D., Hoffmann, A. L., & Stark, L. (2019). *Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning*. Proceedings of the 52nd Hawaii International Conference on System Sciences.

Harding, S. G. (2004). *The feminist standpoint theory reader: Intellectual and political controversies*. New York: Psychology Press.

Harding, S. G., Grewal, I., Kaplan, C., & Wiegman, R. (2008). *Sciences from below: Feminisms, postcolonialities, and modernities*. Durham, NC: Duke University Press.

Holstein, K., Vaughan, J. W., Daume, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? *Chi 2019: Proceedings of the 2019 Chi Conference on Human Factors in Computing Systems*. doi:10.1145/3290605.3300830. Retrieved from <Go to ISI>://WOS:000474467907057.

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information Communication & Society*, *20*(1), 14–29. doi:10.1080/1369118x.2016.1154087. Retrieved from <Go to ISI>://WOS:000386299500002.

Krefting, L. (1991). Rigor in qualitative research: The assessment of trustworthiness. *American Journal of Occupational Therapy*, *45*(3), 214–222.

Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science*, *65*(7), 2966–2981. doi:10.1287/mnsc.2018.3093. Retrieved from <Go to ISI>://WOS:000475704700002.

Latour, B. (1988). *The pasteurization of France*. Cambridge, Mass.: Harvard University Press.

Latour, B. (2005). *Reassembling the social: An introduction to actor-network theory*. New York: Oxford University Press.

Latour, B., & Porter, C. (1993). *We have never been modern*.

Letter to Google C.E.O. (2018). Retrieved from https://static01.nyt.com/files/2018/technology/googleletter.pdf

Mackenzie, A. (2018). From API to AI: Platforms and their opacities. *Information, Communication & Society*, *22*(13), 1–18.

Moseley, A. (2011). Just war theory. *The Encyclopedia of Peace Psychology*.

Murphy, R. R., & Woods, D. D. (2009). Beyond Asimov: The three laws of responsible robotics. *IEEE Intelligent Systems*, *24*(4), 14–20. doi:10.1109/MIS.2009.69.

Nagy, P., & Neff, G. (2015). Imagined affordance: Reconstructing a keyword for communication theory. *Social Media + Society*, *1*(2), 2056305115603385. doi:10.1177/2056305115603385. Retrieved from http://journals.sagepub.com/doi/abs/10.1177/2056305115603385

Neyland, D. (2016). Bearing account-able witness to the ethical algorithmic system. *Science Technology & Human Values*, *41*(1), 50–76. doi:10.1177/0162243915598056. Retrieved from <Go to ISI>://WOS:000365740700003.

Norman, D. A. (1988). *The psychology of everyday things*. New York: Basic Books.

Office of the Chief Scientist, A. (2016). Australia's STEM workforce: science, technology, engineering and mathematics.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown Publishing.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Cambridge, MA: Harvard University Press.

Reddy, E., Cakici, B., & Ballestero, A. (2019). Beyond mystery: Putting algorithmic accountability in context. *Big Data & Society*, *6*(2), 1. doi:10.1177/2053951719863500. Retrieved from <Go to ISI>://WOS:000474228200001.

Russell, S. J., & Norvig, P. (2003). *Artificial intelligence: A modern approach* (2nd ed.). Upper Saddle River, N.J: Prentice Hall.

Sayes, E. (2014). Actor–network theory and methodology: Just what does it mean to say that non-humans have agency? *Social Studies of Science*, *44*(1), 134–149. doi:10.1177/0306312713511867. Retrieved from http://journals.sagepub.com/doi/abs/10.1177/0306312713511867

Schraube, E. (2009). Technology as materialized action and its ambivalences. *Theory & Psychology*, *19* (2), 296–312. doi:10.1177/0959354309103543. Retrieved from <Go to ISI>://WOS:000265235200010.

Seaver, N. (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, *4*(2), 205395171773810. doi:Unsp 205395171773810410.1177/2053951717738104. Retrieved from <Go to ISI>://WOS:000415052700001.

Shank, D. B., DeSanti, A., & Maninger, T. (2019). When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Information Communication & Society*, *22*(5), 648–663. doi:10.1080/1369118x.2019.1568515. Retrieved from <Go to ISI>://WOS:000462285600006.

Tavory, I., & Timmermans, S. (2014). *Abductive analysis: Theorizing qualitative research*. Chicago: University of Chicago Press.

Veale, M., Van Kleek, M., & Binns, R. (2018). *Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making.*

West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems: Gender, race and power in AI.