Effective Adoption and Implementation of AI Principles

Stephanie Kelley, Smith School of Business, Queen's University

Abstract

The study examines employee perceptions on the adoption and implementation of artificial intelligence (AI) principles in organizations. 49 interviews were conducted with employees of 24 organizations, across 11 countries. Participants all worked directly with AI and represented a range of positions from junior data scientist to Chief Analytics Officer. The study found that there are eleven components that could impact the effective adoption and implementation of AI principles in organizations: communication, management support, training, ethics office(r), reporting mechanism, enforcement, measurement, accompanying technical processes, sufficient technical infrastructure, organizational structure, and an interdisciplinary approach. The components are discussed in the context of the research on business code effectiveness. The findings offer a first step in understanding the impact of adoption and implementation on the effectiveness of AI principles in organizations.

Key words: AI, artificial intelligence, adoption, implementation, principles, ethics

Introduction

Reports of organizations proliferating bias and discrimination, jeopardizing customer privacy, jobs lost due to workforce automation, and making many other ethical dilemmas in their use of artificial intelligence (AI) are increasing (Whittaker et al. 2018). For example, the Apple Card, a joint venture between Apple and Goldman Sachs, was recently accused of discriminating against women in their use of an AI-based credit approval model ("Apple's 'sexist' credit card investigated by US regulator" 2019)[1]. Not long before that, IBM, Microsoft, and Megvii were found to be proliferating racial and gender discrimination in their AI-based facial recognition technologies (Buolamwini and Gebru 2018), whilst a similar technology from Amazon was also found to be racially discriminating. The use of AI in organizations is only projected to grow and with it, so too will the number of ethical issues (Kaplan and Haenlein 2019a). It is clear from these, and other reports, that adhering to the existing laws is not enough to prevent unethical AI outcomes; a concern shared by legal and management scholars (e.g., Barocas and Selbst 2016, Martin 2018).

Increasingly, organizations have accepted responsibility for the outcomes of their AI (Martin et al. 2019), and in an attempt to reduce the ethical issues many have turned to self-regulation initiatives (Cath et al. 2018). Preliminary regulatory documents, including the European Commission's White Paper

on Artificial Intelligence[2], and Denmark's data ethics amendment to the Danish Financial Statements Act[3] also point towards the use of organization-lead initiatives in the future. The most common self-regulation initiative, and the focus of this study, are AI principles. Several examples of these AI principles have been gathered by the OECD.AI Policy Observatory, available here: https://oecd.ai/countries-and-initiatives/stakeholder-initiatives.

Before proceeding to the research question, a definition of 'AI principles' is proposed; it is important to define a technology as its study is shaped by its definition (Martin and Freeman 2004). 'AI' refers to 'artificial intelligence,' "a system's ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation" (Kaplan and Haenlein 2019, p.17). 'Principles' is describes the contents of the documents which primarily are focused on stakeholder principles (Kaptein 2004) including transparency, fairness, and accountability (Fjeld et al. 2020; Jobin et al. 2019). The word 'principles' is also used in the title of most documents (e.g. Google[4], Microsoft[5], Telefonica[6]), and the majority of the literature to date (e.g., Fjeld et al. 2020; Floridi 2019; Morley et al. 2020).

Artificial intelligence principles (AIPs) are defined as *a formal document developed* (Kaptein and Schwartz 2008) *or selected by an organization that states normative declarations* (Fjeld et al. 2020; Hagendorff 2020) *about how artificial intelligence ought to be used by its managers and employees.*

Although less common, AI principles are also referred to as guidelines (Jobin et al. 2019), tenets (Mittelstadt 2019), codes of conduct (ACM Ethics 2018), declarations, ground rules, strategies, and statements (Fjeld et al. 2020). As with business codes, an AIP must be a "formal document" (Kaptein and Schwartz 2008). It can be "developed or selected by an organization," as in practice, organizations have been found to generate their own AI principles (e.g., HSBC[7]) or declare their adherence to principles developed by another party, such as an industry consortium (e.g., Partnership on AI[8]), an intergovernmental organization (e.g., The Organisation for Economic Co-operation and Development[9]) or a governing body (e.g., Monetary Authority of Singapore[10]). This is one of two key differentiating factors of AIPs versus business codes (BCs), as a BC must be an internal document "developed by and for a company" (Kaptein and Schwartz 2008, p.113). AIPs also must include "normative declarations," (Fjeld et al. 2020; Hagendorff 2020) that refer to the way AI activities ought to be done based on a set of stakeholder-aligned values (Spiekermann 2016). Values may differ across an organization, but the assumption is that the AI principles present a set of agreed-upon fundamental values (Brusoni and Vaccaro 2017). AIPs must address "artificial intelligence," the second key differentiating factor of AIPs

versus BCs, which instead cover multiple issues (Kaptein and Schwartz 2008); as AIPs cover a single issue, they could also be referred to as 'sub-codes' (Kaptein and Schwartz 2008) and could be one part of a larger ethics program (Kaptein 2009).

Although the study of AIPs is nascent, the study of AI ethics in organizations has been of interest since the technology was first developed and was first highlighted as a management concern by Khalil (1993), who argued that managers must remain legally and ethically responsible when using AI in decision making due to the technology's lack of human intelligence, lack of emotions and values, and possible incorporation of intentional or accidental bias. Since that time only a handful of scholars have studied AI ethics in organizations (e.g., Huang and Rust 2018; Kaplan and Haenlein 2019b; Martin 2019; Martin et al. 2019; Morley et al. 2020), but Khalil's (1993) notion that AI ethics is a management concern is still echoed by scholars today (Martin 2019). Within the broader study of AI ethics in organizations this study is focused 'ethics for design', the study of principles, codes of conduct, standards and other similar initiatives that ensure the integrity of AI systems developers and users (Dignum 2018), which assigns responsibility to the organization using the AI. Within ethics for design, this study investigates 'moving from principles to practice' (Morley et al. 2020) and specifically, the role of adoption and implementation in AI principle effectiveness.

The effectiveness of AI principles has been empirically studied by a handful of scholars because companies only recently started adopting AI principles, with the first thought to have been developed in 2016 by the Partnership on AI (Fjeld et al. 2020; Jobin et al. 2019). To examine the studies to date the nomenclature from the business code of ethics (or corporate code of ethics) literature is borrowed, which categorizes the empirical studies into three types: "c*ontent* oriented studies" (what is in the actual principles), "*output* oriented studies" (what effects the principles have), and "*transformation* oriented studies" (how the principles are adopted or not in an organizations)" (Helin and Sandström 2007, p.254).

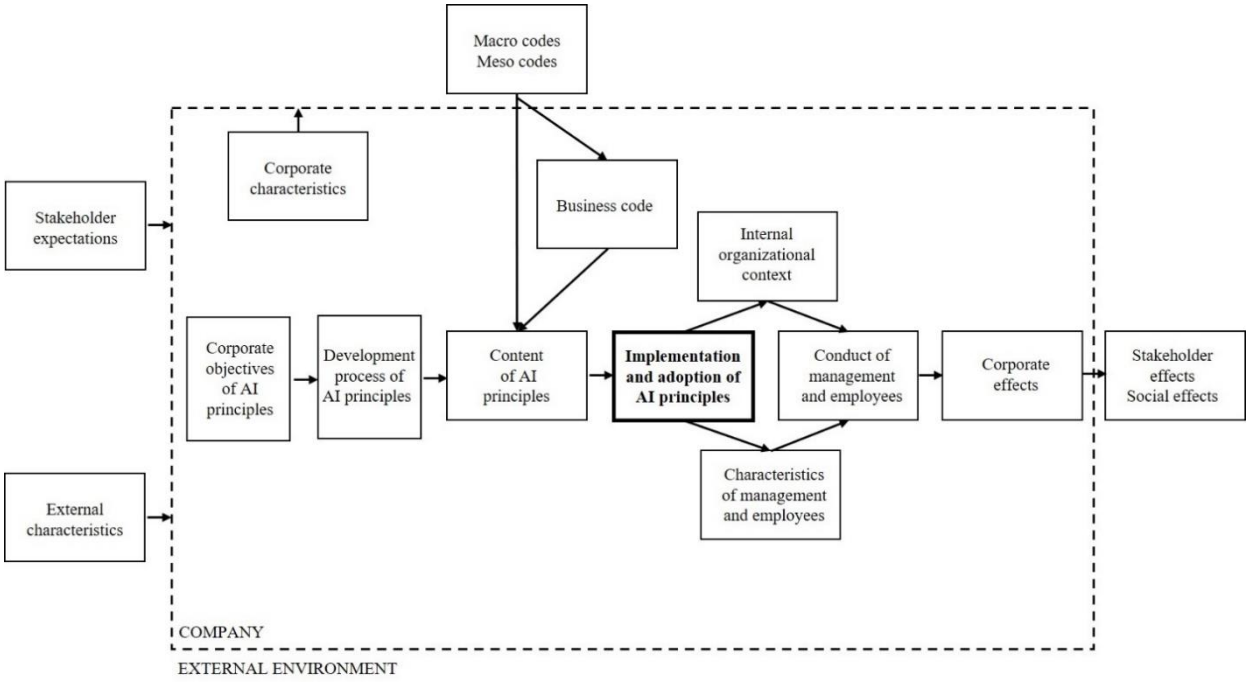
Most studies on AI principles have been *content* oriented. Jobin et al (2019) review 84 ethical principles and guidelines for AI and conclude that there exists a degree of convergence around five ethics principles: transparency, justice and fairness, non-maleficence, responsibility, and privacy (Jobin et al. 2019). Fjeld et al (2020) review 36 AI ethics principles and find there are eight key themes, suggesting a level of convergence in topics: privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values (Fjeld et al. 2020). Schiff et al (2020) review 88 AI principles and discuss the overarching challenges of the existing principles, motivations behind their creation, their potential

governance impact. Although primarily *content* focused, Schiff et al (2020) propose five factors that could impact AI principle effectiveness, which further extends the work into the *transformation* oriented realm: engagement with law and governance, specificity of the document, document reach, enforceability and monitoring, and iteration and follow-up. Mittelstadt (2019) reviews a list of 84 principles, and notes a similar convergence around four principles from medical ethics; he suggests that AI principles lack commons aim and fiduciary duties, professional history and norms, proven methods to translate principles into practice, and robust accountability mechanisms.

To date there has been a single *output* oriented study by McNamara et al (2018), which used ethical vignettes to measure the response of software engineering students and professional software developers to the newly updated (2018) Association of Computer Machinery's Code of Ethics. They concluded that the AI code of ethics had no effect on ethical decision making when compared to a control group that did not read the code (McNamara et al. 2018). The finding of this *output* oriented study contrasts the large body of literature on the efficacy of business codes (BCs), where the majority of studies find BCs to be effective in reducing unethical behaviour (Babri et al. 2019). It has been suggested that the use of ethical vignettes may not provide a strong enough manipulation in ethics behavioural research (Kaptein and Schwartz 2008), which could explain the inefficacy found by McNamara et al (2018). To date, there have been no *transformation* oriented studies on AI principles, those that investigate adoption. The question "are AI principles effective?" remains unanswered because there is only on study on the efficacy of AI principles which contrasts with the research on business codes and may have methodological weaknesses.

Given the limited scope of past studies on AI principles, past studies on business code effectiveness are used as a starting place; Kaptein and Schwartz's (2008) integrated research model for the effectiveness of BC is used as a foundation for the study of the effectiveness of AI principles, and an adapted version of the model is proposed for the study of AI principle effectiveness in Figure 1. The hope is that by using components of an existing framework, this study will help lay the foundation for a consistent body of research on AI principles (Babri et al. 2019). This study focuses on one factor from the integrated research model, the 'implementation and adoption of AI principles' in organizations, herein referred to as 'adoption' of 'AIPs' for conciseness. Specifically, it asks "how are AIPs effectively adopted?", an important step in answering the broader question of interest: "are AIPs effective?".

Figure 1. An integrated research model for the effectiveness of AI principles, adapted from Kaptein and Schwartz (2008)



Several adoption factors have been empirically found to impact BC effectiveness; they act as a starting point for the study of effective AIP adoption, and are summarized below in Table 1.

Table 1. Summary of components the impact business code effectiveness

Adoption components	References
(1) Communication	
Distribution	(Weaver et al. 1999)
Sign-off process	(Schwartz 2004; Singh et al. 2011; Weaver et al. 1999)
Ongoing communication	(Kaptein 2011; Schwartz 2004)
Communication quality	(Adam and Rachman-Moore 2004; Kaptein 2011)
External communication	(Singh 2011)
(2) Management support	
Local management	(Adam and Rachman-Moore 2004; Kaptein 2011; Petersen and Krings 2009)
Senior management	(Kaptein 2011; Schwartz 2004; Singh et al. 2011; Trevino et al. 1999)
(3) Training	
Existence of training	(Adam and Rachman-Moore 2004; Schwartz 2004; Singh 2011; Weaver et al. 1999)
Preferred trainers	(Schwartz 2004; Trevino et al. 1999)
(4) Ethics Office(r)	(Kaptein 2015; Singh 2011; Trevino et al. 1999; Weaver et al. 1999)
(5) Reporting Mechanism	
Existence of a reporting mechanism	(Kaptein 2015; Schwartz 2004; Singh 2011; Trevino et al. 1999; Weaver et al. 1999)
Standardized procedures	(Weaver et al. 1999)
(6) Enforcement	
Audits	(Kaptein 2015; Singh et al. 2011)

Penalties	(Adam and Rachman-Moore 2004; Schwartz 2004; Singh 2011; Singh et al. 2011; Trevino et al. 1999)
Communicating violations	(Schwartz 2004)
Incentive policies	(Kaptein 2015; Trevino et al. 1999)
(7) Measurement	(Weaver et al. 1999)

It is unclear what the impact of these components will be on AIP adoption effectiveness, a gap this study addresses, in addition to determining the components unique to AIP adoption. It is assumed that the views of the employees working with AI are relevant in assessing the potential components that could impact the adoption of AIPs. Therefore, the ultimate research question explored in this study is: "according to the perceptions of employees who work with AI, what aspects of adoption and implementation might relate to AI principle effectiveness?"

After discussing the methodology in the next section, the paper presents evidence based on qualitative interviews for the existence of eleven components that could impact the effective adoption of AIPs. The paper concludes with a discussion of the practical implications, limitations, and future avenues of study.

Methodology

To explore the research question, 49 in-depth, semi-structured interviews were conducted with individuals employed in financial services organizations who work with AI. Financial services was chosen due to its wide-spread adoption of AIPs (DeutscheBank et al. 2019), and its high potential for AI growth, second only to the technology industry (Bughin et al. 2017). The interviews spanned 24 organizations, with an average of two employees interviewed at each organization. Participants were in the United States, Singapore, Canada, the United Kingdom, Australia, Sweden, China, Thailand, Mexico, Brazil, and South Africa.

A snow-ball technique was used for respondent selection due to the newness of AI ethics initiatives within organizations. The researcher contacted AI ethics leaders in financial services as it was assumed, they would have the best knowledge of the AI ethics initiatives are their organizations, and their adoption process. The researcher then asked these leaders to suggest other employees in their organizations, or other organizations who they thought would also be aware of their organization's AI ethics initiatives and their adoption. Of those contacted, 35% agreed to be interviewed, of which ~69% (34/49) were male, and ~31% (15/49) were female. With regards to their position in the organization, ~39% (19/49) were executives or vice presidents, ~55% (27/49) were managers, and ~6% (3/49) were

non-managers. Participants worked at their company from 1 to 17 years, with an average of 4.5 years at their organization.

Interviews took place between December 2019 and July 2020. 16 of the interviews were conducted in person, and the remaining 33 were conducted over videoconference technology or the phone due to COVID-19 travel restrictions. Interviews were conducted in English, the primary language of business for all participants and were on average 40 minutes long. 41 of the interviews were recorded, with the remaining eight, at the request of the participants, not recorded, although extensive notes were taken during and after these interviews. Recordings were first transcribed using a natural language processing (NLP) transcription technology (Otter.ai) and reviewed and edited by the researcher for accuracy. The transcriptions were then coded using qualitative research software (NVIVO).

The interview-based methodology was selected because it has been suggested as a better fit than quantitative methods to understand "the relationship between codes and behaviour," (Schwartz 2001), and "how codes work," (Babri et al. 2019). Furthermore, it allows for the investigation of AIPs in “an actual corporate setting involving actual users,” (Schwartz 2001), extending beyond the exclusive laboratory study of AIPs to date (e.g., Fjeld et al. 2020; McNamara et al. 2018).

Findings and discussion

The researcher initiated the interviews with questions on the components of AIP adoption from the BC literature; the findings of which are presented first. Participants were then asked a several open-ended questions on AIP adoption, not rooted in the BC research, which are presented second.

Components impacting AI principle adoption effectiveness from the business code literature

Communication

It is assumed, per BCs, that simply having an AIP is not enough to ensure their effectiveness in preventing unethical AI outcomes. Communication is likely an important first step in their adoption to create awareness. Participants were therefore asked questions regarding AIP: (a) *distribution*; (b) *sign-off process*; (c) *ongoing communication* internally; (d) the *communication quality*, and (e) *external communication*.

(a) *Distribution.* The extent of BC distribution amongst employees has been found to impact effectiveness (Weaver et al. 1999), and has also been proposed to impact AIP adoption (Schiff et al. 2020). When asked about the extent of distribution amongst employees, there appeared to be two opposing views; the first

group of participants suggested that it would be a positive for all employees working with AI to be aware of the principles, but more important for them to know how to apply them in practice:

> …there's a difference between knowing this is an area we should be worrying about, and that work is underway, versus 'do I know all the principles?'…

The second group of participants questioned whether it was necessary for there to be widespread awareness of the principles in AI employees.

> … I want to leave my data scientists free to find any possible pairings that they might find interesting or relevant and then we (the participant and their boss) decide whether this is something that is safe to put out in the open.

Several respondents did however suggest with some cynicism that distribution by itself may not be enough to create effective adoption, leaving the importance of it unclear.

> …yes, they have read it, I'm kind of laughing because I think they read it – the data science leaders – I think they read it because they wanted to look good – it wasn't genuine – there's a lot of superficiality.

> …I think everyone now even at a junior business level is aware of those principles and the need for fairness, ethics, accountability, and transparency in model building – even though they don't always know what those things mean, they know it's important…

With respect to the timing of distribution, the idea of sharing the AI principles prior to employment (Schwartz 2004) was not mentioned by participants, suggesting it may also not be important for AIP adoption; however, several participants discussed the potential importance of discussing AI ethics in the hiring process:

> …we asked all the potential candidates to complete a case study – build a model…tell us do you believe the model discriminates against age, or gender? We were trying to get people thinking about that…one way to contain that issue is to discuss it in the hiring process as an evaluation.

> …it has been discussed right from the beginning…so, it has always been very clear from the start that we shouldn't do anything unethical with out customer's data, anything outside of the framework.

The importance of how the AIPs were distributed emerged as potentially important though. There appeared to be four ways participants found out about them: (1) they were asked to participate in the AI principle design; (2) they were provided the principles directly by a manager; (3) they found out them through internal marketing or; (4) they had to actively seek out the AI principles. There appeared to be no

obvious preference between having the AI principles provided by a manager (2) or through internal marketing (3), but participants spoke positively about being asked to participate in the principle design, and suggested that a lack of participatory design could hinder adoption.

> … it isn't conducive to a cultural mindset shift…there hasn't been a concerted effort to co-create with the practitioners or the developers that need to change their thinking.

Participants who had to actively seek out the AI principles worked in organizations that are currently piloting the principles (a factor that will be discussed in greater detail later) in select business units, and it was confirmed that their teams were not selected as part of the pilot (e.g., human resources, International banking). They did not speak positively or negatively about the lack of communication to their team, simply that it was a reality of their organizational structure.

(b) *Sign-off process.* Mandatory sign-off practices are common practice in several countries (Singh et al. 2011; Weaver et al. 1999), but are known to receive pushback from employees (Schwartz 2004) which could negate some of the practice's effectiveness. When asked about AIP sign-off, most participants simply noted that it was too early in the existence of the AI principles, or that is just was not a priority at this point. Some participants suggested that they might look at sign-off in the future, but it appears not to be an important factor for AIP adoption at this time.

(c) *Ongoing communication*. BC effectiveness is known to be impacted by communication frequency (Kaptein 2011; Schwartz 2004; Weaver et al. 1999), with about one third of employees receiving ethics communication more than once a year (Weaver et al. 1999). Multiple channels are used by companies to increase communication frequency including policies, memos (including email), poster, newsletters, videos, and the company intranet (Kaptein 2011). When it comes to AIPs organizations appear to be heavily engaged in ongoing communication, and participants specifically discussed the importance of frequency to adoption of the principles:

> …not just communicate once and say 'this, is out there,' but really have our practitioners and people that support our practitioners understand what it means.

As with BCs, companies use multiple formal channels to increase AIP communication frequency including company intranets, emails, conferences, lunch and learns, executive speeches, annual reports, academic conferences, and the media. The individual impact of these channels on AIP adoption was not explored. Informally, AIP ongoing communication occurs through employee communities, team meetings, and peer-to-peer discussions, which could be linked to communication quality:

> …our bosses will send out an email to remind us that, you know, when you are querying data you have to do this if you have any sensitive information, don't leave them unattended, you're computers need to locked when you're not at your desk. So, this is always a reminder and always communicated to us on a very periodic basis.

(d) *Communication quality.* The use of informal communication methods, such as managers openly discussing the principles with their employees, or through social norms (Adam and Rachman-Moore 2004), as opposed to formal training, directives, or classes have been found to be effective in communicating business codes. Accessibility, understandability, and usefulness of the communication of a business code have also been found to impact effectiveness (Kaptein 2011).

Respondents suggested that formal communication may improve adoption. Specifically, creating a clear message with internal marketing, potentially with mnemonics or marketing phrases to easily describe the AIP was suggested to improve understandability; having clear definitions of terms included in the principles (e.g., AI, fairness, explainability); and ensuring cultural and contextual relevance of the messaging around the principles for multinational firms were both suggestions to increase understandability to improve adoption. In terms of informal communication, discussion of the principles in team meetings, employee communities, and peer-to-peer chats was noted, but when questioned about their effectiveness, participants did not suggest these methods were more effective than the formal channels. The use of informal systems in addition to formal systems have, however, been found to impact business code effectiveness (Smith-Crowe et al. 2015), so it may be that a combination of the two could be most effective for AI principle adoption.

An additional communication quality factor affecting AIP adoption, not explored in the BC literature discussed by participants was scale, or impact, particularly of the AIP initial distribution.

> …marketing 101 – we overwhelmed people over a period of a few months: stickers on tables, we have coffee carts where you'll get a free coffee and on it there will be [the AI principles], emails, quizzes, giveaways, we designed a set of videos that tells a story with cartoons – so it becomes real with them – stickers on elevators…we did a whole bunch of stuff in every location…

(e) *External communication*. Singh (2011) found that communicating a BC externally impacts code effectiveness; for example, sharing the code with customers, suppliers, or displaying it externally.

When it comes to sharing AI principles, there appears to be four degrees of sharing that occur. First, some organizations do not share their principles externally, often citing reputational risk as rationale.

> ….nothing published, which is interesting but deliberate because [the bank] is a privately held company, and overall very, very secretive.

Second, some have created a white paper, or summary of their principles to be shared, whilst keeping the AIPs confidential internally. Participants again cited reputational risk as a key barrier of sharing, a concern also shared by the third group of organizations who ended up sharing their principles. Organizations share their AI principles through one or more external channels: an annual report, in a sustainability/corporate social responsibility report, on the organization's website, or through the media. Those that did share suggested it would bring an additional degree of importance to the principles, likely impacting their effectiveness.

> We went backwards and forwards on how we declare this to the world – at one point we were going to go big and write an op-ed in a [national newspaper] but in the end we just thought there was a lot of downside and not a lot of upside, a bit too brash. So instead we've done a few things in the comms, we did a formal publication of the principles – falls under enterprise sustainability report and there's a link to the principles there.

However, some participants suggested that sharing the principles externally was nothing more than 'whitewashing', and that it ultimately may not be effective in changing behaviour.

> …you can find it, it's open source….there is a lot of emphases on it – I don't know if it's helpful…it's brought up a lot as an excuse not to do something rather than for legitimate reasons…

> …we're signalling that we're a bank that's very conscious of this, right – we pushed out the training, we advertise it, but it's a signal and it's not effective.

The fourth group of organizations are those that adopted publicly available AIPs, the Monetary Authority of Singapore has a set of AI principles [vi] (not formal regulation), which several banks in Singapore have adopted. These principles are already publicly available, but the firms may publicly promote their commitment to these principles in varying degrees.

Management Support

Management support, both from (a) *local* (Kaptein 2011; Petersen and Krings 2009) and (b) *senior management* (Kaptein 2011; Schwartz 2004; Singh et al. 2011; Trevino et al. 1999) has been found to impact BC effectiveness. Not surprisingly, support from both levels of management also appear to impact AIP adoption.

(a) *Local management.* Employees have been found to follow BC advice from local management (i.e., supervisors and manager within the business unit) even when it contradicts the BC (Petersen and Krings 2009). BC effectiveness is also positively impacted when local managers model appropriate ethical behaviour (Adam and Rachman-Moore 2004; Kaptein 2011; Petersen and Krings 2009). When it comes to AIPs, local management support is important because they are seen to be more knowledgeable about AI ethics than senior managers or data scientists.

> …the people that are the most aware about AI ethics is the mid-level managers because they are managing products directly and they take full aspect for the project more carefully thank the actual data scientists themselves.

(b) *Senior Management.* Seeing senior management model appropriate ethical behaviour (Kaptein 2011; Schwartz 2004), hearing them talk about the code (Schwartz 2004; Singh et al. 2011), discovering they know about and understand the code (Schwartz 2004), or generally take the code and ethics seriously (Trevino et al. 1999) are all factors impacting BC effectiveness. Senior management support is also seen as an important factor in AI principle adoption; however, respondents noted that given the technical nature of AI ethics senior managers did not actually work with AI, they expected them to know about the principles, and to talk about them, but not to model specific behaviour. Senior management support appeared to have greater gravity to impact adoption across the organization compared to local management support.

> One thing I've been really pleased about it the level of interest, commitment at an executive management level… think that level of engagement throughout and the application of it – without that it wouldn't be where it is – it would just be a reputational risk.

> … so it's not just our direct managers, it's the CEO. He immediately picks up on how to incorporate ethics which is a good sign of how seriously we think about it.

A lack of senior management support could also be detrimental to AIP adoption:

> …we don't have somebody who frequently talks about this from the executive level, management committee level, no – I wish that would happen.

Training

Offering BC training has been found to impact its effectiveness (Adam and Rachman-Moore 2004; Schwartz 2004; Weaver et al. 1999), especially when it is offered to all employees (Singh 2011). Participants were asked their thoughts on the importance of the (a) *existence of training* for the AIP, as well as their (b) *preferred trainers*, and highlighted the importance of (c) *general AI training*.

(a) *Existence of training.* Training has been found to improve employee awareness of BCs, and help indicate the importance an organization puts on BCs (Schwartz 2004). Although several participants spoke about the importance of having training on the AIPs, just under two-thirds of the 24 participant organizations have formal training programs on them. Training is thought to be beneficial because most data scientists are not trained in AI ethics, making it difficult for them to implement the AIPs:

> …the junior data scientist is coming out of school right now – are they learning anything about AI ethics? I am not sure about that – I think probably like 50% or less are, because you're still focusing on trying to develop the best model, learn different languages…

Some organizations are mandating training for anyone working with AI, which could help with awareness and ultimately effectiveness of the principles. Training seems to be done online, as well as at annual meetings, between once and twice a year, given the scale.

> …it's an interactive education series type thing that's just on our internal [system] and it get blasted out….to every analytics practitioners.

> …we've set up a training programs…a technology academy…with training courses aimed at 3 different audiences…we're educating all the people who hold accountability for the model, who build the stuff – anchored on the principles…you do it every year.

Participants did however mention the potential for pushback on forced training, potentially decreasing its efficacy.

> …that training…now, candidly, do I think it's going to work? No, because it's like every other corporate training…

Many organizations without a formal training program have some type of informal training run by an internal party (or parties). The training is often not well communicated or widespread, with presenters relying on word of mouth to book training sessions. This likely means it is not as effective as it could be at driving adoption. In addition to this informal internal training, funds are often made available for AI employees to go to external conferences on AI ethics; however, these are not specific to AIPs, and likely do not impact their effectiveness.

(b) *Preferred trainers.* Employees most often prefer training to be done by someone internal to the organization as opposed to an external consultant, preferably a senior manager (Schwartz 2004; Trevino et al. 1999) when it comes to business codes.

Most training programs are run by internal representatives; it is usually not a direct manager who runs the training program because it requires more specialized knowledge on AI ethics and the AI principles than most managers have. Often the trainer is someone who was involved in the development of the principles as they are likely to be one of the most knowledgeable employees on AI ethics. A couple of organizations do appear to have a "train the trainer" program though to expand the training roster, which could increase the number of direct managers performing training.

> … train the trainer…run workshops for practitioners – so those are people that they're just trainers, they would show them how, and create a coach.

One participant did highlight the importance of co-branding a training session as both internal and external; meaning there is a potential that external training, or co-branded training could be more impactful than internal training alone.

> …our certification program…we wanted to have it externally branded because it gives us that level of importance and say we're learning from international best practices, academic best practices – we wanted to make it real for people.

(c) *General AI training.* Close to three quarters of respondents spoke about the lack of general education on AI as a barrier to AIP adoption. A lack of general AI understanding amongst non-AI employees apparently makes it difficult to discuss the AIPs which could reduce their effectiveness.

> …the lack of understanding and low levels of comfort with data science is still a big blocker….so then it means that their involvement is at arm's length, their decision making is impaired….for the few people that I've been able to take through journey into machine learning…for example, I sent my boss on a two day

> training on machine learning…and that was literally night and day when it came to support, discussions about things like ethics…
>
> …I think there's this unfamiliarity with algorithms, and because they don't understand them, they are fighting back on the ethics stuff – that's not helpful...

Some organizations have addressed this concern by developing additional training programs on AI in general for non-AI practitioners, each with varying degrees of focus on ethics and the AI. In every case, these trainings are tailored to specific audiences (i.e. one for people who don't work with AI, one for people who work indirectly with AI, one for executives) likely to increase their relevance, and ultimate effectiveness.

Ethics Office(r)

Having an ethics office (Weaver et al. 1999), an ethics officer (Singh 2011), or an ethics committee (Kaptein 2015; Singh 2011) have all been found to impact BC adoption. Employees are thought to use these resources to ask questions on the BC (Singh et al. 2011), and the individual or group is often assigned responsibility for the BC (Weaver et al. 1999).

When asked about the ethics office or ethics officer, respondents suggested that AIPs would likely not fall under the ethics officer, highlighting the AI-specific knowledge that was required for managing AIPs which the ethics officer would not have. Although the central ethics office does not appear to be involved in AIPs, most organizations have assigned responsibility of the AI principles to a group or individual in the organization. These responsible parties ranged drastically across organizations though, and in most cases, responsibility was assigned to someone who was involved in the development of the AIP like an AI ethics committee, the head of analytics and/or AI, or another senior manager (e.g., risk, legal, sustainability). Several reasons were given for having a panel, but almost all respondents spoke about the importance of a committee or panel to discuss the "hard decisions," as a group, which may suggest they are more effective than a single AI ethics officer.

> …we wanted a panel to review the hard decisions, to safeguard the principles, be the governing body if we find the principles don't work, or cases that challenge them…the panel is a subset of executives from the management committee…across all lines of the business…
>
> …we have a committee of senior leaders…they make decisions on use case and strategic decision on what will be on the framework – critical that is was initially top down…

Reporting Mechanism

With respect to BC, the (a) *existence of a reporting mechanism* has been found to impact effectiveness (Kaptein 2015; Schwartz 2004; Singh 2011; Trevino et al. 1999; Weaver et al. 1999), as well as the existence of a (b) *standardized procedure* for dealing with ethics issues (Weaver et al. 1999);

(a) *Existence of a reporting mechanism.* Several forms of reporting mechanisms have been found to impact BC effectiveness including phone lines, websites, ethics officers, and mail boxes (Weaver et al. 1999). There is strong evidence that employees don't report all BC non-compliance instances despite being mandated to do so (Schwartz 2004), which puts into question their effectiveness. Anonymity could potentially increase their effectiveness (Sims 1991), but lack of awareness, the negative perception of the line as a snitch line, and employee concerns about abuse and actual protection of their anonymity are all factors that could be detrimental to effectiveness (Schwartz 2004); which may help explain why they have been found to only indirectly relate to BC effectiveness (Kaptein 2015).

When asked about reporting AI ethics concerns respondents suggested that the complexity of AI ethics issues makes them more difficult to handle than traditional ethics issue. Participants differentiated between malicious AI ethics acts and non-malicious acts (accidents), the latter of which were by far the most common concern.

> …there's a spectrum – purposely versus accidentally…

The idea that most AI ethics issues happen by accident, perhaps because of a lack of knowledge, or the newness of AI technology, meant that most respondents were focused on getting AI ethics concerns to the right people above all else (e.g., a formal process, or anonymity). The vast majority of respondents noted that they were not aware of any formal reporting mechanism for non-malicious AI ethics issues, but that they would bring the issue up with someone more senior who they thought could address the issue.

> I would report them to compliance and model risk management – from my team's perspective, if people would know the process as it pertains to AI…for front-line employees there's lots of ways to escalate – their manager, executives, whistleblower hotlines…

Having a formal mechanism was discussed as potentially important for more junior employees though.

> …I'm not aware of any formal ways, I would know how to get it into the right hands just because of where I'm positioned…probably our Chief Privacy Officer and [the head of analytics] and that would state the ball rolling, but I'm not sure that other folks would know those people or be comfortable doing it.

Anonymity seemed detrimental to adoption as participants assumed the person on the other end of the phone would not be able to help given the complexity of AI ethics issues.

> …you have a code of conduct that says don't steal, it's pretty obvious what stealing is – you say like most employees understand what that is….what is unethical as it refers or applies to analytics and AI maybe isn't as understood...I suspect no, I don't think a whistleblower line will work because I think fairness is such a hard concept to grasp.

Although not a major concern, in instances where there is an AI issue due to the malicious action of an employee respondents suggested that the general ethics whistleblower line would be the appropriate reporting tool to use to report the concern.

> … the whistleblower hotline for conduct issues – if you look at something and its wrong with ethics in the question around AI because of misconduct – if it's a conduct issue you can call the hotline.

(b) *Standardized procedure*. Having clear procedures for dealing with ethical issues or complaints, even if employees don't agree with them creates a sense of procedural justice in an organization which improves BC effectiveness (Weaver et al. 1999).

Preference for a formal operating procedure or board approved policy was found to be extremely important to ensure the AIP are taken seriously, suggesting it would improve adoption. However, several organizations noted this was the ultimate plan, and that they were either in the process of creating a standardized procedure for AIPs or hoped to do so. Only a handful of organizations have formal procedures in place today.

> ….this will be a guideline for everybody initially, later a formal standard would be developed from this which will actually get enforced – initially a guideline will probably be taken a little more lightly…

> …we are now finalizing ethics and AI policies – so these are risk policies that we as an organization have to adhere to, and now we're going to be building out over 2020 the framework and the guidelines and the standards that will feed into those.

The lack of standardized procedure for reporting AI ethics issues, amongst other things, appears to be felt by the more technical respondents implementing the AI. As one respondent put it:

> So, like topics regarding anti-money laundering, or due diligence, or you know, bribery – those things, there are specific trainings and specific training courses and channels for you to report, but there hasn't been a specific training for [AI] ethics and a specific channel for reporting [AI] ethics issues. I can see that we still have a long way to go.

Enforcement

The use of enforcement mechanisms for BC effectiveness has been extensively studied, with four types of mechanisms found to impact behaviour: (a) *audits* (Kaptein 2015; Singh et al. 2011); (b) *penalties for breaching the code* (Adam and Rachman-Moore 2004; Schwartz 2004; Singh 2011; Singh et al. 2011; Trevino et al. 1999); (c) *communicating violations* (Schwartz 2004); and (d) *incentive policies* (Kaptein 2015; Trevino et al. 1999).

(a) *Audits.* Both internal and external auditors are used by organizations to enforce BCs (Singh et al. 2011), and the use of monitoring and auditing has been found to have a direct negative relationship with unethical behaviour (Kaptein 2015). With respect to AIPs, respondents suggested the potential effectiveness of both internal and external auditors. Internal audits are a common first step, used for auditing process and behaviour compliance to the AIPs. External auditors are then brought in or are being considered for more technical auditing of algorithms, as most participants noted a concern that their internal auditors were not knowledgeable enough about AI to conduct the audits.

> …our audit group have been talking to third parties about how they will audit machine learning…

> …we're working with [a start-up] on the explainability side…it will cover bias and explainability and will help transparency…

Several respondents did note that they thought regulators would start to mandate the use of third party technology for audits; for example, the Australian regulator has already approved the use of DataRobot, a cloud-based machine learning platform[11], to aid in data sharing for consumer loan reporting.

(b) *Penalties.* Penalties for breaching as BC such as reprimand, fines, demotion, dismissal, or legal prosecution are used by organizations globally (Singh et al. 2011), but have been found to be effective in varying degrees (Adam and Rachman-Moore 2004; Singh 2011). Consistency could impact penalty effectiveness (Trevino et al. 1999), and could help prevent the generation of cynicism which reduces a BC's effectiveness (Schwartz 2004). Schiff et al (2020) propose follow-up and enforceability of penalties could impact AIP adoption.

When asked about AIP penalties respondents again differentiated between malicious and non-malicious AI ethics issues and suggested that existing penalties used for breaching a business code (e.g., termination, legal prosecution) could be applied to malicious acts.

> …we have a thing internally where if you break bank principles – effectively employees get conduct points – and if you get negative conduct points then it impacts your bonus a the end of the year, it can impact your ability to get promoted…
>
> There are clear ethical policies that go beyond data that would envelope that and there's a clear process by which you determine the level of punishment – how wrong the thing you did is, which dictates the punishment – being fired, being reported to regulators.

However, the same penalties would not be applied to non-malicious breaches of the AIP; participants suggested there would be repercussions, such as investigations to understand why the non-malicious breach occurred, but there would not be direct penalties.

> …I don't know that there'll be penalties…in my mind penalties is like a financial or job implication – I think there will be repercussions...

Some respondents did mention the use of a "system," used to remind employees to act in line with AIPs and suggested it could be effective.

> …we have a system in our team where if you leave your computer unattended and somebody catches you, you have to buy the whole team drinks…if it happens to you once, then you'll be like 'it actually costs a lot right?

(c) *Communicating violations*. Schwartz (2004) found that communicating violations of a BC is effective way in deterring employees from future breaches; however, the technique has limitations, should remain anonymous, and reserved for serious violations.

Respondents did not suggest that communicating malicious breaches of AIPs would be effective; however, several suggested that post-mortem reports and sharing of accidental breaches would be effective.

> …I think they'll be like 'no this is totally the wrong thing to do, why did we ever go down this route?' It would be like a kind of post-mortem if you screw up a project.

(d) *Incentive policies.* Rewarding good ethical behaviour is consider an effective way to reduce unethical behaviour in organizations (Trevino et al. 1999). Increased ethical behaviour does not result in a decrease in unethical behaviour, but incentive policies have been found to decrease unethical behaviour (Kaptein 2015), but should be excluded from performance reviews (Schwartz 2004).

With respect to AIPs, not a single participant knew of an incentive policy at their organization. Incentive policies for ethics in general were viewed in two very different lights; some participants, particularly those in Canada and Europe seemed to think the use of incentives was unnecessary:

> …like a positive or performance-based reward for that ethical behaviour?...I go back to the notion of a code of conduct that all humans that work here sign – we don't walk around and high five each other every year, I think we tick a box on year end to say, I didn't do bad things…I can see a world where machine learning scientists every year make a declaration or something…
>
> … I think people would be punished for unethical behaviour, but you wouldn't be rewarded for ethical behaviour.

Participants from other countries noted the use of ethics incentives, and although not originally designed to include AIPs, they thought malicious behaviour could fall under these incentives.

> Nothing specific to AI at this point, more broadly there's strong incentives on whistleblowing – there's an annual competition if you report fraudulent or incorrect behaviour you could win [money] as a positive reinforcement.

This means that incentive policies for AIPs could be effective; however, the effect would likely be moderated by cultural dimensions as observed in BC research (Singh et al. 2011), or perhaps by ethical climate (Victor and Cullen 1988; Wimbush and Shepard 1994), or ethical culture (Key 1999).

Measurement

The use of measurement, such as external evaluation may suggest an organization is serious about the BC, and that it is not just a symbolic gesture, increasing its effectiveness (Weaver et al. 1999). Measurement of employee understanding of the code through testing, however, has been found to be ineffective and potentially patronizing (Schwartz 2004).

Only a handful of organizations are currently measuring adherence to AIPs, although several are interested in measuring effectiveness in the future, suggesting its potential importance. Measurement appeared to be broken into two types: adherence to the procedures and policies, and technical adherence (i.e., the algorithmic outcomes).

> …once you have an ethics checklist and guidelines, how do we ensure practitioners or the key groups are using it effectively, are adhering to the principles?...what percentage of people are following it? How many people have it as part of their formalized procedures?

> …once we have a policy, it's great that people know about it but then how do you know its working? What kind of reporting do you have on it so you can actually assess the efficacy of the policy?

As of today, the ability of organizations to measure the technical adherence to the AI principles remains "a pipe dream," as one respondent put it. Another respondent summarized what it could look like:

> …on the technology, which we're further behind on right now…maybe automated tools to detect bias and data as an example, or it could be how we set the operational guidelines on actual AI in production. I think there will be KPIs around measurement or around incidents and on-compliance…

Novel components impacting AI principle adoption effectiveness

In addition to the above components from the BC literature found to impact AIP adoption effectiveness, participants were asked 'whether there were additional adoption and implementation factors that would impact the effectiveness of AI principles?'. Several additional components were suggested including *accompanying technical processes, sufficient technical infrastructure, organizational structure,* and using an *interdisciplinary approach.* The novel components, discussed below, offer a first look into the unique factors an organization should consider when adopting AIP, in addition to those explored from the BC literature.

Accompanying Technical Processes

The lack of technical guidance in AIPs is considered a key gap by respondents when it comes to their potential effectiveness; a concern echoed by AI developers (Peters 2019) and researchers (Mittelstadt 2019). Along similar lines, clarifying the roles of AI developers and AI deployers in AIPs has also been suggested as important by an AI practitioner consortium (DeutscheBank et al. 2019). For example, an AIP might state that "AI models should not discriminate," but what does "discrimination" mean in a technical sense? Those in the machine learning field will know that there is no single definition of discrimination, fairness (Berk et al. 2017), or any other ethics terms, making it difficult for data scientists to implement AIPs in their technical practice. To remedy this, many respondents noted the creation (or ongoing development) of accompanying technical processes to provide detailed technical guidance on the AIPs. These processes were referred to as checklists, frameworks, assessments, and guidelines.

> …we're looking at ethical checklist or data ethics impact assessments…how to embed that within our existing processes and not create new processes for things.

> …when the [AI principles] were announced and released to the bank internally we started looking at, you know, how that would change what we currently have in place…what we needed to do was to supplement it with a bit more guidance …it's not replacing policies that the bank already has, it's just supplementing it.

Respondents suggested several considerations for the technical processes that could potentially improve effectiveness: it should be an adaption of existing processes, such as model validation; it should be integrated into all product and service development, such that any AI model that goes to market goes through the process; it should have different processes for different levels of risk, so as not to stifle innovation; and it should be automated for ease of distribution and use. Adapting an existing process, and using different risk levels were both suggested as recommendations in a recent industry white paper on AI principle implementation (DeutscheBank et al. 2019), echoing their potential importance. The most comprehensive example of such an accompanying technical process was explained in detail by one of the respondents:

> …in terms of a more granular process – you are a data scientist, you've been through legal, and quality, and then you are asked – [does this adhere to our AI principles]? There's an [AI principles] assessment website – you would first look at and search a whitelist of different data use cases and AI use cases that have already been reviewed by senior leaders in the bank – if what you are doing is identical…you just go ahead…for the remaining 5% not on the whitelist you do a self-assessment and very detailed set of questions – depending on how you have answered the questions there is meta-data…certain red flags are triggers - from the way you answer the question it will give you a risk rating on your particular product and if there is enough of a potential risk of this being sensitive or surprising to people then you'll submit the assessment and the head of the department will make a decision whether to take responsibility to deploy the product themselves or escalate the decision to a committee…there is always a balance where we want to be agile and not be a bottleneck and not stand in the way…at the same time we don't want to miss things that will hurt out customers and end up in the newspapers.

Sufficient Technical Infrastructure

Having sufficient technical infrastructure to implement AIPs in models was as an important factor for AIP adoption, specifically: (a) having a *complete AI inventory* of projects, and (b) ensuring *data and system compatibility.*

(a) *Complete AI inventory.* This helps ensure complete distribution of the AIPs and allows for measurement. Some organizations appeared to already have an AI inventory, but many were working to get them completed to aid in adoption and noted significant challenges associated with the task.

> ……trying to get an inventory of where AI is being used in the bank. So again, that is very, very, tricky, and it's really been proven very difficult to get because it's not contained in any particular part of the organization.

(b) *Data and system compatibility.* Respondents noted that legacy systems and data issues also posed a threat to effectiveness, as the technical practitioners are not able to comprehensively implement the principles if data is missing or they cannot access all the systems.

> …[the organization's] data is not even that well managed and organized – so that's a task in itself – so then to talk about how you're using that, if it's ethical, yeah it is quite difficult.

Organizational Structure

The importance of organizational structure specifically centralized versus decentralized structures may impact AIP adoption. Respondents with centralized analytics and AI teams noted it was potentially easier for them to gather a complete AI inventory, highlight the importance of AI, and distribute AIP.

> …we've come together as one team…there are people outside of us that do analytics…however there is a dotted line for those team from the analytics hub, so that makes it a little bit easier to implement something like this…I think it's a very positive move that's it's consolidated under one area…but I think that has delayed us, like the internal trying to figure out who's doing what and what has to be done…

Those with decentralized teams suggested the structure makes AIP distribution, communication, enforcement, and gathering an AI inventory more difficult, and responsibility for AIPs may be less clear which could hinder adoption.

> …right now the structure we still have, for example small data science teams scattered across the bank – while we as the biggest team can set guidelines we can definitely not enforce and take responsibility for other teams and the way they do AI…I think this is the biggest ethics risk in a sense that in the smaller teams – that are probably more like cowboys and can do things that we wouldn't be able to do.

> …a federated set up – in the current structure it's a bit more difficult to constrain and to make sure the communications are sent to the right people.

Interdisciplinary Approach

Throughout the interviews there surfaced a common sentiment from participants across all levels of seniority: AI ethics is a highly complex topic which no one has come close to solving. In response to this concern organizations everywhere have suggested the use of an interdisciplinary approach to the creation, adoption, and use of AIPs. This includes (a) *interdisciplinary teams*, (b) *combining AI ethics with data ethics*, (c) *hiring the right people,* and (d) *engaging with third party experts.*

(a) Interdisciplinary teams. The more people involved in the discussion, the better according to participants; they suggested a wide range of potential team members including data practitioners, analytics and AI practitioners, privacy, legal, compliance, risk, sales, marketing, business strategy, HR, and ethics practitioners. These conclusions echo the suggested importance of diversity and culture for AIP implementation from practitioners (DeutscheBank et al. 2019).

> So we are leveraging cross functional people across the different teams, which includes legal and includes privacy…so I would say that we are including everyone we can simply because this is not something that's very clear that I can say 'okay, this is a regulation, you have to do it.'.

> …an enterprise working group – so from across the enterprise, data and analytics practitioners, legal, privacy…I think there's a better opportunity to have it be successful if there's more people involved in the decision-making and ideation around it, and then actually implemented...

(b) *Combining AI ethics with data ethics.* Another interdisciplinary approach several organizations appear to have adopted is a joint AI and data ethics program; several respondents suggested that AI is highly reliant on data, and therefore AIP adoption should be aligned with data ethics initiatives.

> …we intentionally are doing this under the data management heading because we also see lots of interesting connections between this and privacy…we're approaching this holistically. We don't have a separate stream on AI governance, but it's part of one data risk management framework.

> …these principles – use them both on AI, machine learning, privacy and data use – so it's not just about AI anymore…without data there is no machine learning or artificial intelligence that can be done…so I think those were kind of eye-opening moments for us as an organization when we started to look more holistically about how to apply ethics.

(c) *Hiring the right people*. Several respondents noted that today they simply don't have the right expertise in their organization to implement AI principles properly, and to rectify this have worked to hire people from outside the organization with the proper skills, including the use of third party vendors. In line with this finding, it has been noted that technology vendors can be an important factor in implementing AI principles (DeutscheBank et al. 2019).

> …on the compliance side what they have done is a lot of self-reflection…and have identified that they don't know enough and started hiring accordingly people that have that skillset.

Some organizations felt they already had the skills internally though so hiring more people was not a priority, suggesting it is potentially important for effective adoption, depending on an organization's current talent pool.

(d) Engaging with third party experts. Regardless of whether an organization has hired people to help with AIPs, most respondents noted the importance of engaging with third party AI experts. Organizations appear to be primarily engaged with technology companies who are heavily invested in AI and AI ethics, AI vendors, as well as academia.

> …we've convened a group of experts; we don't believe we can solve this in our own. No one [organization] can solve this thing on their own…so a key part of this is Microsoft. Microsoft have an ethics board, and [one member of the board] graciously helped and supported a bunch of the stuff we did…

> …asking ourselves all the way through, 'do we need external advice?' – mostly on the practicalities of the setting up and running this. We are already Google and Microsoft customers, so we reached out to speak to the people who run their equivalent panels and principles and brought some of those lessons in.

(e) *Engaging with regulators*. Several of the participants noted that they are proactively engaging with industry regulators, which they suggest has aided in AIP adoption, for example, through increased awareness. Activities include having regular meetings with the regulator, responding to calls for research or surveys, and inviting the regulator to external AI ethics events. Certain regulators are not concerned at this time with AI ethics, which may impact the importance for effectiveness overall, but there appears to be high levels of engagement in Singapore, Canada, the UK, and Australia.

> …the [regulator's AI principles], we were one of the parties that contributed to the development of that and I think everyone now even at a junior business level is aware of those principles…

> …the whole ethics of AI as well as the AI strategy is really grounded by [our regulator], right. So we were very keen with the survey – we don't normally say that with [our regulator], but we see it as an opportunity…we think's it's in our best interest to kind of forge the discussion around it.

A summary of the findings, the eleven components that could impact the effective adoption of AI principles is presented in Table 2.

Table 2. Summary of components that could impact the effective adoption of AI principles

Adoption components	Relationship to AI principle effectiveness	Summary of relationship importance
Components impacting AI principle adoption effectiveness from the business code literature		
(1) Communication		
Distribution	Potentially important	Two groups: "all AI employees should see them," and "only managers need to worry about them"; distribution is nothing without understanding.
Sign-off process	Potentially important	Not common practice today, but some organizations may look to implement in future.
Ongoing communication	Important	Formal (e.g., company intranet, reports) and informal communication (e.g., team meetings, employee communities) are both used for frequency, effective communication.
Communication quality	Potentially important	Aligning communications with marketing, having clear definitions, ensuring cultural relevance, and communication scale all could play a factor.
External communication	Potentially important	Four groups: not shared, white paper/summary shared, shared, already publicly available; could be considered whitewashing.
(2) Management Support		
Local management	Important	Seen to be more knowledgeable about AI then senior management; likely has greater impact on adoption in technical employees.
Senior management	Important	Greater gravity than local management; without their support adoption will likely fail.
(3) Training		
Existence of training	Potentially important	Training important to educate technical team on ethics; mandatory training may get pushback.
Preferred trainers	Potentially important	Likely important to have someone internal; external co-branding/partnerships for training may make training more impactful.
General AI training	Important	Basic training on AI for anyone involved with AI projects is seen as very important.
(4) Ethics Office(r)	Potentially important	Specific AI ethics officer not necessarily important, but responsibility assigned to an individual or ethics panel is vital.
(5) Reporting Mechanism		
Existence of a reporting mechanism	Potentially important	Malicious AI principles breaches would use existing ethics reporting mechanism; non-malicious acts may not need a reporting mechanism but could benefit junior employees.
Standardized procedures	Important	Increases seriousness of the principles, may be particularly important for technical employees.

(6) Enforcement		
Audits	Potentially important	Internal and external auditors are being used, but lack of knowledge of AI may be a barrier.
Penalties	Potentially important	Important for malicious AI principles breaches; existing penalties could be used. Softer penalties may be important for non-malicious acts.
Communicating violations	Potentially important	Likely not important for malicious breaches; important to share non-malicious breaches via post-mortem to prevent future issues.
Incentive policies	Potentially important	Highly dependent on country the organization is operating in.
(7) Measurement	Potentially important	Many want to measure the principles, but only a handful are currently doing so.
Novel components impacting AI principle adoption effectiveness		
(8) Accompanying Technical Processes	Important	Translating the AI principles into more technical guidelines is necessary to ensure their adoption.
(9) Sufficient Technical Infrastructure		
Complete AI inventory	Important	Aids in the distribution and tracking of principles.
Data and system compatibility	Important	Data issues and legacy systems can prevent adoption of principles at a technical level.
(10) Organizational Structure	Potentially important	A centralized AI team may make adoption easier.
(11) Interdisciplinary Approach		
Interdisciplinary teams	Important	Increased diversity of thought, especially from outside the AI team is important.
Combining AI ethics with data ethics	Potentially important	Some organizations have integrated the two, but operating structure usually remains separate.
Hiring the right people	Potentially important	Important if AI ethics talent is not available internally.
Engaging with third party experts	Important	Technology companies, AI vendors, and academia all play a role.
Engaging with regulators	Potentially important	Dependent on willingness of regulator to engage, and the regulator's AI ethics maturity.

Conclusion

This paper presents the first *transformation* oriented study of AI principles by exploring the views of AI practitioners. Eleven components that could impact AIP adoption effectiveness are uncovered, sever of which are from the BC effectiveness literature and have been found to apply to AIP adoption, albeit in varying degrees of importance. These similarities suggest that the integrated research model for the study of BC effectiveness from Kaptein and Schwartz (2008) can, in fact be adapted for the study of AIPs (per Fig. 1). However, this study also finds that there are four novel components unique to AIP adoption: accompanying technical processes, sufficient technical infrastructure, organizational structure, and interdisciplinary approach. These novel components highlight the uniqueness of AIPs compared to BCs, and the ongoing need for AIPs to be identified and studied as their own entity. Furthermore, they indicate that organizations should not treat AIPs as BCs without making adjustments for the nuances and differences in adoption to ensure AIP effectiveness.

There are, however, limitations to the study. First, the participants in the study are all employed by financial services organizations, which could limit the generalizability of the findings to other industries. Future research should explore AI principle adoption and implementation in other industries using AIPs such as technology (e.g., Microsoft 2018), and telecommunications (e.g., Telefonica 2019). Second, the qualitative nature of the study relies on self-reported data, which could be plagued by social desirability bias (Randall and Fernandes 1991), amongst other issues. Future research to address could use empirical tests to measure the effectiveness of the eleven adoption components uncovered here. To do so, a measure of unethical behaviour as it relates to AIPs must be developed, perhaps per the measure of unethical behaviour used in the study of BCs (Kaptein 2008). The development of such a measure represents an avenue for additional future research on the effectiveness of AI principles.

A measure of unethical AI behaviour could also be used to study the effectiveness of AIPs content, building upon the existing studies from Fjeld et al (2020), Schiff et al (2020), and Jobin et al (2019), as well as other factors suggested to impact effectiveness in the proposed integrated research model (Fig. 1), such as ethical culture and ethical climate (internal organizational conduct), and the development process. In addition to the empirical study, the theoretical study of AI principle effectiveness is likely warranted (Babri et al. 2019), as suggested by Kaptein (2011).

In practice, organizations across many industries are implementing AIPs. The findings from this study suggest that simply having AI principles will not be enough to prevent unethical AI outcomes, but that several adoption and implementation components could help address this. Further research on AI principle development, content, and implementation should be conducted to continue developing the understanding of AI principle effectiveness in practice.

Ethics Approval

This study was performed in line with the principles of the Declaration of Helsinki. Ethics approval was granted by the Queen's University General Research Ethics Board (GREB) (11/12/19, amended 03/09/2020, and 05/09/2020; TRAQ #6028134)

End Notes

[1] https://www.bbc.com/news/business-50365609

[2] https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

[3] https://kammeradvokaten.dk/nyheder-viden/nyheder/2020/06/nu-skal-virksomheder-redegoere-for-dataetik-i-aarsrapporten

[4] https://ai.google/principles/

[5] https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6

[6] https://www.telefonica.com/en/web/responsible-business/our-commitments/ai-principles

[7] https://www.hsbc.com/-/files/hsbc/our-approach/risk-and-responsibility/pdfs/200210-hsbc-principles-for-the-ethical-use-of-big-data-and-ai.pdf?download=1

[8] https://www.partnershiponai.org/partners/

[9] https://oecd.ai/ai-principles

[10] https://www.mas.gov.sg/~/media/MAS/News%20and%20Publications/Monographs%20and%20Information%20Papers/FEAT%20Principles%20Final.pdf

[11] https://www.datarobot.com/solutions/banking/

References

ACM Ethics. (2018). 2018 ACM Code of Ethics and Professional Conduct: Draft 2.

Adam, A. M., & Rachman-Moore, D. (2004). The methods used to implement an ethical code of conduct and employee attitudes. *Journal of Business Ethics*, *54*(3), 225–244. https://doi.org/10.1007/s10551-004-1774-4

Apple's "sexist" credit card investigated by US regulator. (2019). *BBC News*. https://www.bbc.com/news/business-50365609. Accessed 11 November 2019

Babri, M., Davidson, B., & Helin, S. (2019). An Updated Inquiry into the Study of Corporate Codes of Ethics: 2005–2016. *Journal of Business Ethics*. https://doi.org/10.1007/s10551-019-04192-x

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *104 California Law Review 671*.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). *Fairness in Criminal Justice Risk Assessments: The State of the Art*. https://doi.org/https://doi.org/10.1177/0049124118782533

Brusoni, S., & Vaccaro, A. (2017). Ethics, Technology and Organizational Innovation. *Journal of Business Ethics*, *143*(2), 223–226. https://doi.org/10.1007/s10551-016-3061-6

Bughin, J., Hazan, E., Ramaswamy, S., Chui, M., Allas, T., Dahlstrom, P., et al. (2017). *Artificial Intelligence - The Next Digital Frontier*. https://doi.org/10.1016/S1353-4858(17)30039-9

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commerical Gender Classification. In *Proceedings of Machine Learning Research* (Vol. 81, pp. 1–15). https://doi.org/10.2147/OTT.S126905

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. *Science and Engineering Ethics*, *24*(2), 505–528. https://doi.org/10.1007/s11948-017-9901-7

DeutscheBank, Linklaters, Microsoft, StandardChartered, & Visa. (2019). *From Principles to Practice: Use Cases for Implementing Responsible AI in Financial Services*.

Dignum, V. (2018). Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology*, *20*(1), 1–3. https://doi.org/10.1007/s10676-018-9450-z

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. C., & Srikumar, M. (2020). *Principled Artificial intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*. *The Berkman Klein Center for Internet & Society Research Publication Series*. https://doi.org/10.1109/MIM.2020.9082795

Floridi, L. (2019). Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. *Philosophy and Technology*, *32*, 185–193. https://doi.org/10.1007/s13347-019-00354-x

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, *30*, 99–120. https://doi.org/10.1007/s11023-020-09517-8

Helin, S., & Sandström, J. (2007). An inquiry into the study of corporate codes of ethics. *Journal of Business Ethics*, *75*(3), 253–271. https://doi.org/10.1007/s10551-006-9251-x

Huang, M. H., & Rust, R. T. (2018). Artificial Intelligence in Service. *Journal of Service Research*, *21*(2), 155–172. https://doi.org/10.1177/1094670517752459

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine*

Intelligence, *1*, 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kaplan, A., & Haenlein, M. (2019a). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*, 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

Kaplan, A., & Haenlein, M. (2019b). Rulers of the world, unite! The challenges and opportunities of artificial intelligence. *Business Horizons*. https://doi.org/10.1016/j.bushor.2019.09.003

Kaptein, M. (2004). Business codes of multinational firms: What do they say? *Journal of Business Ethics*, *50*(1), 13–31. https://doi.org/10.1007/978-94-007-4126-3_27

Kaptein, M. (2008). Developing a measure of unethical behavior in the Workplace: A stakeholder perspective. *Journal of Management*, *34*(5), 978–1008. https://doi.org/10.1177/0149206308318614

Kaptein, M. (2009). Ethics programs and Ethical culture: A next step in unraveling their multi-faceted relationship. *Journal of Business Ethics*, *89*(2), 261–281. https://doi.org/10.1007/s10551-008-9998-3

Kaptein, M. (2011). Toward Effective Codes: Testing the Relationship with Unethical Behavior. *Journal of Business Ethics*, *99*(2), 233–251.

Kaptein, M. (2015). The Effectiveness of Ethics Programs: The Role of Scope, Composition, and Sequence. *Journal of Business Ethics*, *132*(2), 415–431. https://doi.org/10.1007/s10551-014-2296-3

Kaptein, M., & Schwartz, M. S. (2008). The effectiveness of business codes: A critical examination of existing studies and the development of an integrated research model. *Journal of Business Ethics*, *77*(2), 111–127. https://doi.org/10.1007/s10551-006-9305-0

Key, S. (1999). Organizational ethical culture: Real or imagined? *Journal of Business Ethics*, *20*(3), 217–225. https://doi.org/10.1023/A:1006047421834

Khalil, O. E. M. (1993). Artificial Decision-Making and Artificial Ethics: A Management Concern. *Journal of Business Ethics*, *12*(4), 313–321.

Martin, K. E. (2019). Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics*, *160*, 835–950. https://doi.org/10.1007/s10551-018-3921-3

Martin, K. E., & Freeman, R. E. (2004). The Separation of Technology and Ethics in Business Ethics. *Journal of Business Ethics*, *53*, 353–364. https://ssrn.com/abstract=1410846

Martin, K. E., Shilton, K., & Smith, J. (2019). Business and the Ethical Implications of Technology: Introduction to the Symposium. *Journal of Business Ethics*, *160*, 307–317. https://doi.org/10.1007/s10551-019-04213-9

McNamara, A., Smith, J., & Murphy-Hill, E. (2018). Does ACM's code of ethics change ethical decision making in software development? In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering - ESEC/FSE 2018* (pp. 729–733). https://doi.org/10.1145/3236024.3264833

Microsoft. (2018). Microsoft AI Principles. http://nirn.fpg.unc.edu/about-nirn/our-approach. Accessed 9 September 2019

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices.

Science and Engineering Ethics, *26*, 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Peters, D. (2019, May). Beyond Principles: A Process for Responsible Tech. *Medium*. https://medium.com/ethics-of-digital-experience/beyond-principles-a-process-for-responsible-tech-aefc921f7317. Accessed 14 September 2020

Petersen, L. E., & Krings, F. (2009). Are ethical codes of conduct toothless tigers for dealing with employment discrimination? *Journal of Business Ethics*, *85*, 501–514. https://doi.org/10.1007/s10551-008-9785-1

Randall, D. M., & Fernandes, M. F. (1991). The Social Desirability Response Bias in Ethics Research. *Journal of Business Ethics*, *10*(11), 805–817. https://doi.org/10.1007/978-94-007-4126-3_9

Schiff, D., Biddle, J., Borenstein, J., & Laas, K. (2020). What's next for AI ethics, policy, and governance? A global overview. In *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 153–158). https://doi.org/10.1145/3375627.3375804

Schwartz, M. S. (2001). The nature of the relationship between corporate codes of ethics and behaviour. *Journal of Business Ethics*, *32*(3), 247–262. https://doi.org/10.1023/A:1010787607771

Schwartz, M. S. (2004). Effective corporate codes of ethics: Perceptions of code users. *Journal of Business Ethics*, *55*(4), 323–343.

Sims, R. R. (1991). The institutionalization of organizational ethics. *Journal of Business Ethics*, *10*(7), 493–506. https://doi.org/10.1007/BF00383348

Singh, J. B. (2011). Determinants of the Effectiveness of Corporate Codes of Ethics: An Empirical Study. *Journal of Business Ethics*, *101*(3), 385–395. https://doi.org/10.1007/s10551-010-0727-3

Singh, J. B., Svensson, G., Wood, G., & Callaghan, M. (2011). A longitudinal and cross-cultural study of the contents of codes of ethics of Australian, Canadian and Swedish corporations. *Business Ethics*, *20*(1), 103–119. https://doi.org/10.1111/j.1467-8608.2010.01612.x

Smith-Crowe, K., Tenbrunsel, A. E., Chan-Serafin, S., Brief, A. P., Umphress, E. E., & Joseph, J. (2015). The Ethics "Fix": When Formal Systems Make a Difference. *Journal of Business Ethics*, *131*(4), 791–801. https://doi.org/10.1007/s10551-013-2022-6

Spiekermann, S. (2016). *Ethical IT Innovation: A Value-Based System Design Approach*. (J. Cantella, Ed.). Boca Raton: Taylor & Francis Group, LLC.

Telefonica. (2019). *Our Artificial Intelligence Principles*.

Trevino, L. K., Weaver, G. R., Gibson, D. G., & Toffler, B. L. (1999). Managing Ethics and Legal Compliance: What Works and What Hurts. *California Management Review*, *41*(2), 131–151.

Victor, B., & Cullen, J. B. (1988). The Organizational Bases of Ethical Work Climates Author. *Administrative Science Quartely*, *33*(1), 101–125.

Weaver, G. R., Treviño, L. K., & Cochran, P. L. (1999). Corporate ethics practices in the mid-1990s: An empirical study of the fortune 1000. *Journal of Business Ethics*, *18*(3), 283–294. https://doi.org/10.1007/978-94-007-4126-3_31

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., et al. (2018). *AI Now Report 2018*. New York. www.ainowinstitute.org

Wimbush, J. C., & Shepard, J. M. (1994). Toward an understanding of ethical climate: Its relationship to ethical behavior and supervisory influence. *Journal of Business Ethics*, *13*(8), 637–647.

https://doi.org/10.1007/BF00871811