From Ethical Artificial Intelligence Principles to Practice: A Case Study of University-Industry Collaboration

Keeley Crockett SMIEE[1], Sean Brophy[2], Samuel Attwood[3], Peter Monks[3], David Webb[3]
[1]Department of Computing and Mathematics, [2]Business School, [3]GM AI Foundry,
Manchester Metropolitan University, Manchester, M1 5GD, UK, K.Crockett@mmu.ac.uk

***Abstract*— The ethical artificial intelligence principal to practice gap is a significant challenge for micro and small medium businesses (SME). The policy and legal landscape is very dynamic and whilst there are limited toolkits, designed to help such businesses to embed the ethical design of responsible technology, there is generally a lack of skills, knowledge and resources on how to apply them within individual businesses. In this paper we present a small case study of practical examples that has led to the introduction of ethical AI practices into SMEs. Through a European funded university-industry collaboration, to date 102 SMEs within Greater Manchester, UK have been exposed to data and ethical AI workshops, with a subset of these choosing to deeper dive and apply tools such as consequence scanning and harms modelling within their business with support from university technical analysts and academics. The case study presents initial evaluations on embedding ethical principles, challenges faced by the SMEs and the reflections of technical analysts.**

Keywords—ethical artificial intelligence, toolkits, responsible technology, industry

I. INTRODUCTION

Artificial Intelligence (AI) is a field of study focussed on the synthesis and analysis of computational agents that act intelligently [1]. In 1943, McCulloch and Pitts created one of these programs and demonstrated that a simple artificial neuron could be the basis of a Turing-complete machine [2]. Technologies developed within the field of AI have the potential to be a truly transformative. In healthcare, AI technologies have been applied to better engage patients, streamline administrative processes, and to better diagnose and treat patients [3]. In finance, AI has been applied to credit evaluation, portfolio management, and general prediction and planning [4]. In transport, AI concepts have been used to better plan, design, and control transportation networks [5]. Unfortunately, AI also has the power to be transformative in a negative sense as well. Numerous news stories have highlighted unethical applications of AI and the damage such applications can do [6..8]. For example, [9] highlights the problems with large language models and is notable—not only for its insights—but also the backlash it received from large technology organisations [10].

It is therefore vital that stakeholders identify and understand the implications of unintended harms and consequences of AI, in their research, or development of new ideas, products and services. New forthcoming legislation in the use of AI (such as the proposed EU Regulatory framework [11]) will be one of the most disruptive forces that the technology industry has faced. Governments are prepared to sign up to ethical principles and recommendations (UNESCO [12]) but participation is voluntary. The monumental problem is how to not only implement these principles in practice, but how to physically monitor and audit business and organisations that develop AI technologies, where employees may not have the required skill set, resources are low, and ethical approaches may not be incorporated into the business plans. For small to medium size business (SME) (up to 250 employees) these problems are magnified.

The Greater Manchester AI Foundry (GMAIF) project is a £6M three-year research and innovation project, which began in July 2020 and is part funded by the European Regional Development Fund (ERDF). Manchester Metropolitan University, University of Manchester, University of Salford, and Lancaster University are the four partner universities delivering the project, which sees beneficiaries taken through two phases. A beneficiary is an SME who agrees to join the project. The project aims to support over 170 micro and SMEs in total over 3 years and support the development of over 65 new products and services through technical assists. In phase 1 of the project, beneficiaries participate in a series of educational workshops. Each workshop is 3 hours long and there are 8 workshops in total: (1) programme induction, (2) market opportunities, (3) exploring AI technology, (4) innovation lifecycle, (5) rapid innovation techniques, (6) prototyping, (7) ethics, and (8) project review. After phase 1, eligible beneficiaries progress to phase 2 where they receive technical assistance towards an innovative product or service linked to AI. The exact form of this technical assistance can vary and is highly dependent on the needs of individual beneficiaries. A key part of the technical transfer is specialised knowledge transfer between academics, a university based technical assist team working in partnership with the SME. The GMAIF project is therefore in an influential position, and it will play a key role in determining whether the potential of AI—both promising and perilous—is realised (at-least in the Greater Manchester region). The aim of this paper is to firstly describe how a university-industry collaboration has led to the introduction of ethical AI practices, through the use of toolkits and workshops, into SMEs as part of the Greater Manchester AI Foundry project (henceforth referred to as the 'GMAIF Project') [13]; this is the first step in trying to embed ethical principles into SMEs. Secondly to provide initial evaluations of these practices from both the technical analyst and academic perspective. The contributions resulting from our research are as follows:

1. We highlight and explain shortcomings of traditional University Institutional Review Boards) in managing the unqiue ethical risks posed by AI, in relation to knowledge transfer projects with industry. Such ethical risks often manifest ‘downstream’ within the project development lifecycle when AI is being applied.

2. We describe the approach of the GMAIF project—a university-industry research collaboration—giving examples of how ethical best practices can be introduced and embedded within SMEs. We evaluate these practices from a technical analyst and academic persepctive.

This paper is organized as follows; Section II covers related work in ethics and provides a brief review of the principles, and frameworks of 'Ethical AI' relevant to this work. Section III describes the GMAIF project approach and briefly describes and justifies a number of ethical toolkits that were introduced to SMEs. An exploratory evaluation of the application of these toolkits is given in section IV and conclusions and next steps are summarised in section V.

II. RELATED WORK

A. Ethics in context

This section contextualizes our paper by providing a critical overview of the ethical context within which the ethical policies and procedures for the GMAIF project have been developed. There is a growing body of literature that recognizes the unique ethical challenges AI poses to individuals and society [14, 15]. What is less clear, however, is the role that universities play in mitigating the perceived risks of AI in scientific and industrial applications. Of particular concern to academia is the ethical challenges that arise in university-industry research collaboration, where conflicts often arise between parties based on competing values, goals, processes, and timeframes [16..17]. This indicates a need to understand the intersection of AI ethics and research ethics in this context of university-industry research collaboration. It is necessary here to clarify some key concepts that will inform the rest of this paper. To paraphrase Isaiah Berlin [18], ethics are ultimately concerned with *the proper relations between people*, which is also to say that AI is somewhat incidental to a human-centered conception of ethics. In an AI context, relations can be said to include those between developers of algorithmic systems, those responsible for their deployment, the end-users, as well as broader constituencies like the academy, firms, industry sectors, communities, governments, and society at large. The human-centered conception of ethics has led some to suggest that the very idea of 'moral machines' is conceptually flawed from the outset and that the priority for the practice of AI ethics ought to be ensuring humans actually understand ethical concepts before considering whether their mechanical creations can [19]. In the field of ethics, a distinction is often made between morality, moral theory, meta-ethics and applied ethics [20], all of which are contested terms, and the debates about which will not be discussed in detail here. This paper is largely concerned with *applied ethics* – i.e., the application of moral theory in the form of AI ethics and research ethics – in the context of a single project, the GMAIF.

The likely effects of any AI system – e.g., whether a particular AI system is harmful or beneficial to individuals or society – depends on both scientific and ethical considerations. Take, for example, the claim that an algorithm used in recruitment software is biased because it results in an outcome where fewer women are selected for job interviews than men. This claim can be judged partly on the basis of facts and partly on the basis of values. In this example, a common technical solution to identifying and correcting sample selection bias in machine learning is to reweight the cost of training point errors to better approximate that of the test distribution [21]. Therefore, the incidence of bias – when taken as a matter of fact – can be discovered scientifically and solved technically. However, the technical solution of ensuring that the number of females selected for interviews matches the distribution of females in the training data is a statistical principle and not an ethical principle. To illustrate how an ethical approach to bias is different from a technical approach, take, for example, the *IEEE P7003 Standard for Algorithmic Bias Considerations* [22], which states that "*Unjustified bias refers to differential treatment of individuals based on criteria for which no operational justification is given*." Whether bias is justified or not – which is another way of saying whether it is right or wrong – is a question of values that cannot be answered by appealing to the scientific method. Instead, questions of values must rely on some form of moral theory – such as deontology, utilitarianism, or virtue ethics – to determine whether they are right or wrong, justified, or unjustified [23]. An example of a utilitarian argument against the use of biased algorithms in hiring decisions is that gender discrimination does harm (or has the potential to do harm) to half of the working population and is, therefore, immoral. It is immoral because it is less likely than its alternative (equality of opportunity) to provide the greatest good for the greatest number. Similar arguments can be devised on the basis of deontological or virtue ethics. The distinction between *questions of fact* and *questions of values* is a restatement of the *is-ought problem* first described by David Hume (see [24] for a discussion), which is a reminder that it is problematic (or impossible) to base ethical claims (like what we *ought* to do related to AI) on factual arguments (like what *is* profitable, fashionable, or technically feasible).

A further distinction can be made between *normative morality* – which is ultimately concerned with moral theory – and *descriptive morality* – which is moral theory in a codified form [25]. Descriptive morality – which is more the focus of this research – is ultimately empirical since the moral principles in question can be observed and described in specific contexts. Examples of descriptive morality include observables like research codes of conduct such as the Universities UK concordat for research integrity [26], regularity frameworks like [1] and the practices of university Research Ethics Committees (RECs) (also known as Institutional Review Boards or IRBs). RECs evolved out of the medical model of ethical review, and their primary function is to review research proposals and render a judgment as to whether the research project in question is ethical on the basis of descriptive standards. Both research ethics and ethical AI frameworks have adopted common principles from medical ethics that are primarily concerned with the welfare of individual human research participants [27], foremost among them are what Diener and Crandall referred to as the 'Big Four' values of non-maleficence, beneficence, autonomy, and justice [28].

The development of research ethics in the second half of the 20th century should be understood as a response to post-war debates about the unethical conduct of scientists during the Second World War that harmed wider society, including

the participation of scientists in genocide and the development (and use of) weapons of mass destruction. RECs originated within the National Health Service (NHS) in the 1960s, largely in reaction to a series of developments over the preceding 20 years like the Nuremberg Code of 1947, the World Medical Association's Helsinki Declaration of 1964, and the US Surgeon General's 1966 memo on research ethics to recipients of grants from the US Public Health Service, which included several leading UK institutions at the time [29]. It was also in the 1960s that the medical model of ethical review became common in the social sciences, partly in reaction to the Milgram experiments at Stanford University that used controversial methods to explore obedience to authority and the psychology of genocide in the aftermath of the holocaust [30]. The experience of the Second World War was also important for the development of ethical practices in engineering and the physical sciences. The leading role of scientists in the Manhattan Project, which ultimately resulted in the use of nuclear weapons on Japanese civilians, led to a debate in the scientific community about the proper role of scientists in wider society. The Pugwash Conferences on Science and World Affairs were first organized in 1955 in response to the *Russell–Einstein Manifesto* of 1955, affirming that scientists have a special obligation to society because of their specialized expertise [31].

B. *Principles, Polices and Frameworks of 'Ethical AI'*

The emergence of publications, policies, and guidelines relating to ethical AI over the last 5 years [32] can be understood within this context. One such publication, of particular importance to this research, is the proposed EU Regulatory Framework, which outlines a risk-based approach to AI and was issued Apr-21 [11]. Although this proposal may take several years to become law, its impact will be at least as significant as the GDPR 2018 to business, with severe fines being imposed on organisations which fail to comply. Under the regulation, organisations with high-risk AI systems will need to have a number of procedures and processes in place, relating to: data and AI governance, risk management of AI, the transparency and explainability of AI decision making, human oversight, and conformity assessment against legalisation. It is also likely that some existing systems will be categorised as unacceptable risk and no longer be permitted. Mckinsey have highlighted the need for organisations to prepare for future legislation now, by undertaking an inventory of existing AI systems, developing risk classification and migration systems, and undertaking conformity assessments which should be fully documented [33]. Large organizations with extensive resources, such as Microsoft and IBM, have welcomed the proposed regulation but also suggested it is too prescriptive and suggested several changes [34, 35]. Unfortunately, the voices of smaller organizations are in danger of going unheard, which means they could be disadvantaged and unprepared when regulations like the EU Regulatory Framework are written into law. Projects like the GMAIF, which work closely with these organizations, are therefore obliged to help them prepare for these upcoming changes.

There is abundance of ethical toolkits that could be applied within the context of the GMAIF project to help organizations and businesses develop responsible and trustworthy AI [36]. Crockett et al [37] evaluated and ranked 77 published toolkits based on 33 evaluation criteria which included: toolkit scope in addressing identified SME barriers to adoption; human-centric AI principles, relevance to specific AI lifecycle stages, and key themes around responsible AI and practical usability [37]. The results of this study showed that there is no one-size-fits-all toolkit that addresses all criteria and is suitable for all organizations, including SMEs. This research therefore selects some of the most promising toolkits and practices identified by Crockett et al [37] and applies them within the context of the GMAIF project. Additionally, it explores some approaches that this prior work fails to consider and evaluate. By doing this, this research aims to elicit valuable feedback to inform the remainder of the GMAIF project and help researchers, scientists, and practitioners that are tasked with delivering similar projects.

III. GM AI FOUNDRY APPROACH

In this section the exercises and practices that have been applied with respect to AI and research ethics, within the context of the GMAIF project, are described. These include the application of several tools: micro-futures, consequence scanning, and the Microsoft Responsible Innovation Toolkit are all described. Additionally, an explanation as to how (and why) the standard use of Institutional Review Boards (IRBs) has been supplemented to account for AI ethics is provided. The simplified flowchart seen in Figure 1, which excludes extraneous procedures unrelated to research and AI ethics, shows when each of the exercises and practices is performed with respect to one another within phase 1 and 2 of the project. Throughout this section the SMEs involved within the GMAIF project are referred to as *beneficiaries* and the researchers and technical analysts working on the project are referred to as the *project team*.

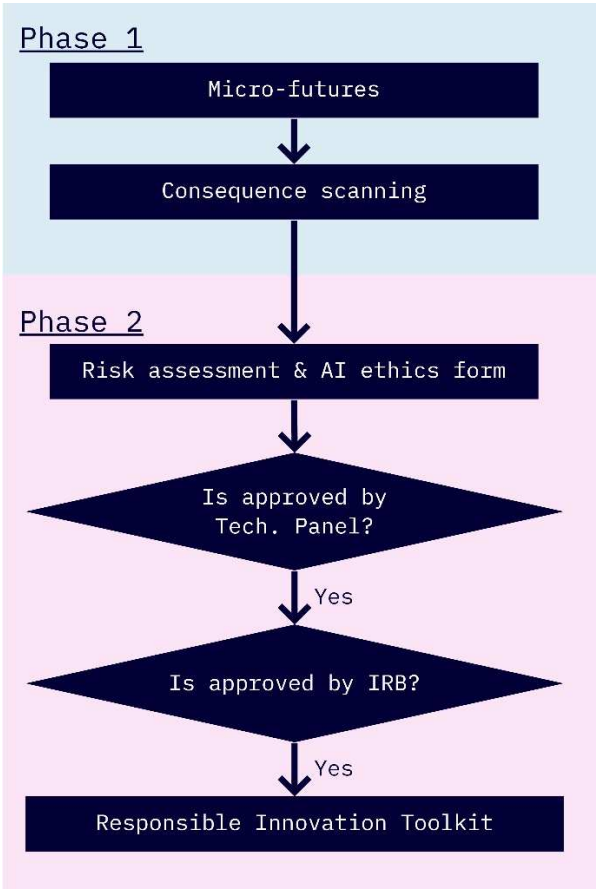

Figure 1: A simplified flowchart showing the exercises and practices relating to AI ethics that have so applied within the context of the Greater Manchester AI Foundry project.

A. *Micro-Futu*res

Micro-futures are a kind of science-fiction prototyping (SFP), which is itself a technique that sees narrative techniques used to explore potential future scenarios [39]. It has been said SFP allows for immersive design experiences that might otherwise be impossible [39]. The approach is also very low cost and requires less resource than many other prototyping approaches. Micro-futures take the science-fiction prototyping idea to its logical conclusion and involve the creation of a very short science-fiction story, anywhere between 6 to 1000 words in length [40]. In one prior study, Burnam-Fink [41] reflects on experiences using science-fiction prototyping at a workshop and provides three critiques: (1) writing is slow and science-fiction prototyping is time consuming, at-least within the context of a workshop; (2) there was insufficient reflection and consideration of alternatives; (3) if materials are not presented in a neutral way participants are less likely to develop their own interpretations of issues / consequences [41]. Micro-futures have the potential to naturally address these critiques.

Another study proposes and evaluates a card game--Judgement Call—that similarly uses design fiction to explore ethical considerations relating to technology [42]. The game sees players: (1) choose a scenario; (2) identify stakeholders; (3) draw a hand containing a rating card, stakeholder card, and ethical consideration card; (4) write a review based on their hand; (5) share reviews and discuss with other players [42]. It is a lightweight exercise, at-least compared to one that involves the creation of a full science-fiction prototype, and clearly has the potential to address the critiques raised by Burnam-Fink. However, despite Judgement Call being a lightweight exercise it was not possible to use it during in the GMAIF phase 1 workshops due to time constraints. It was however possible to conduct a less time-consuming exercise in which beneficiaries were tasked with creating micro-futures.

The general form of the micro-futures exercise beneficiaries were asked to complete is as follows: (1) beneficiaries are presented with a summary of science-fiction prototyping; (2) beneficiaries are presented with a summary of micro-futures; (3) beneficiaries are presented with an overview of the key steps to creating a micro-future; (4) beneficiaries are presented with an example micro-future, with the key components highlighted; (5) beneficiaries are tasked with creating their own micro-future. For the final part of the exercise beneficiaries are split into groups and given a short amount of time, typically 15 to 25 minutes.

The micro-futures exercise was delivered as a part of the sixth workshop in the phase 1 component of the GMAIF project. By this point, beneficiaries had already attended several workshops and been given a comprehensive overview of AI, from both a business and technical perspective. Materials were presented in a balanced and neutral way throughout the workshops. Beneficiaries were therefore well equipped to form their own opinions and perspectives when completing the exercise.

B. *Consequence Scanning*

Consequence scanning is a technique developed by Doteveryone [38] and curated by the Open Data Institute. It is designed to help organizations think about the potential impact of their AI/data driven products and services on individuals and society. Although the technique can be applied to existing products, it's intended usage is at the conceptualization stage and then throughout the product lifecycle. One of the key benefits of using this technique is that it can be facilitated by the business, requires little training (comes with a handbook), is not resource intensive and is free. Consequence scanning involves asking three key questions which are asked in facilitated sessions [38].

1. What are the intended and unintended consequences of this product or feature?
2. What are the positive consequences we want to focus on?
3. What are the consequences we want to mitigate?

After each session all consequences are recorded in a log, mitigation plans for unintended consequences can be prioritized and established, and measurement criteria can be set. Consequences may be positive or negative and apply to individuals (users/bystanders who may be direct or indirect stakeholders) and the wider world. Unintended consequences are what could happen because of a business's actions. For example, a business develops a new facial mood detection system for the purposes of advertising, where the aim is to provide customers with a personalized advertising service to improve the conversation ratio of ads to sales. This could be seen as a positive consequence for the business, however what would be unintended consequences of showing an offensive add to a customer, based on a high false positive rate? Consequence scanning was first introduced to beneficiaries enrolled onto phase 1 of the GMAIF as a part of an interactive (and virtual) ethics and data governance workshop. Beneficiaries were asked to consider their AI product/idea or service that they were looking to explore and develop. The sessions were run under Chatham House rules. Pre-set pad lets were used to record individual business answers to the three questions. Answers were at a high level to protect intellectual property. To date, three cohorts (52) of beneficiaries have taken part in consequence scanning.

C. *Adopting a Risk based apprach to AI*

All beneficiaries who wished to progress to phase 2 of the GMAIF project were required to undertake ethical approval. All partner universities are signatories of the *Universities UK concordat for research integrity,* which ensures the minimum ethical review standards across the partnership are consistent with high standards of research integrity and governance. However, unique ethical challenges presented by artificial intelligence mean that important issues are not covered in standard IRB review procedures. Therefore, there was a need to go beyond current compliance and introduce beneficiaries to the concept of a risk-based approach to AI. Using the initial EU proposed AI regulation framework [11], a *Research Ethics and Research Governance Code of Practice* was agreed between partner universities on 20th July 2021. The existing partner ethical approval processes and procedures adequately address the GMAIF project's impact on the wider research community, but AI presents several risks to society that may conflict with the mission of higher education institutions as forces for good. This concern stems from the fact that many of the ethical challenges related to AI manifest themselves downstream in the application of the technology in the marketplace and not upstream in the research phase of the project. The GMAIF project has therefore adopted additional

ethical and risk governance structures. A supplementary AI risk and ethical assessment form was created, which is to be completed by the beneficiary and members of the project team (at the partner university that delivered the phase 1 workshops to the beneficiary). This form asks for the potential positive and negative impacts the AI product/service could have on society to be listed, before requiring the AI product/service to be classified in accordance with the EU Framework [11]. If the AI product/service is deemed 'High' risk, the beneficiary and project team members are required to answer additional questions around who the AI system will affect and how it will affect them, risk mitigation strategies, quality and of datasets, human oversight, and explainability. The completed form is ultimately presented to the GMAIF Technical Panel special Cross-Institutional Review Board that comprises of members of the four partner universities. Members determine whether the risk classification is appropriate and whether the GMAIF project should provide technical assistance to the beneficiary as a part of phase 2. Table 1 presents the GMAIF project interpretation of the risk classifications outlined in the EU Framework [11] alongside the actions, the project team members are expected to take if the phase 2 technical assistance is approved.

Table 1: Risk classifications derived from the Proposed EU Regulation Framework on AI [11] with the corresponding actions for the project teams

Risk	**Description**	**Req. Actions**
Unacceptable	AI systems that violate fundamental rights.	None. Activities relating to AI systems of this kind are banned by the GMAIF project.
High	AI systems that present a high risk to health, safety, or fundamental rights. AI systems used in any of the following areas: biometric identification and classification, critical infrastructure, educational and vocational training, employment and workers management, access to essential private and public services and benefits, law enforcement, asylum or border control, and administration of justice and democratic processes.	The project team is expected to address concerns around who the AI system will affect and how it will affect them, risk mitigation strategies, quality and of datasets, human oversight, and explainability, to the satisfaction of the awarded partner's Institutional Review Board and throughout any phase 2 technical assistance.
Limited	AI systems that pose transparency issues, such as chatbots that interact with humans, systems that detect emotions or infer social categories based on biometric data, and systems that generate or manipulate content.	The project team is expected to address concerns around transparency issues to the satisfaction of the awarded partner's Institutional Review Board and throughout any phase 2 technical assistance.
Minimal	AI systems that do not present a high risk to health, safety, fundamental rights and do not pose transparency issues. AI systems used for spam filtering and in video games.	The project team is expected to satisfy the awarded partner's Institutional Review Board.

D. The Microsoft Responsible Innovation Toolkit

The Microsoft Responsible Innovation Toolkit aims to help developers consider the effects of technology and future science on society [43]. In 2021, it was ranked, the number one toolkit for SMEs out of 77 [37] due to its coverage of AI principals, suitability for SMEs, coverage of the AI lifecycle stages and addressing responsible AI and practical usability. Despite being in the early stages, the toolkit has been shared so that feedback can be provided. It includes 3 practices: judgement call, harms modeling, and community jury [37]. The first of these practices—judgement call—is a card game that has already been discussed in section *C* and is described fully by Ballard et al [42]. The second, harms modeling is similar to threat modeling and requires a developer to think through and document the harms a system or technology could feasibly inflict upon society, ideally in collaboration with other stakeholders [44]. To help identify harms Microsoft provide a template that outlines 10 types of harm: (1) physical or infrastructure damage; (2) emotional or psychological distress; (3) opportunity loss, which involves limiting access to resources or services; (4) economic loss, which is similar to opportunity loss but concerned with access to financial resources and services specifically; (5) dignity loss, which involves interfering with the exchange of honor and respect; (6) liberty loss, which involves infringing legal rights or amplifying existing biases in social systems; (7) privacy loss; (8) environmental impact; (9) manipulation, which involves creating highly personalized and manipulative experiences that ultimately undermine trust; (10) social detriment, which otherwise refers to ways a technology could impact communities and social structures [44]. They also suggest considering how acutely an individual or group would be impacted by each type of harm (severity), how broadly the impact would be experienced (scale), how likely it is that the harm would occur (probability), and how often the harm could arise (frequency), to help evaluate the overall landscape and plan accordingly [44]. The third and final practice is to hold a community, or citizen, jury [45]. A practice first described by Ned Crosby in 1971 [46] that has since been applied considerably [47]. At a high level, holding a jury involves sampling a representative group of participants from a community, providing them with background information, facilitating discussion amongst jury members, helping to create recommendations, and ultimately amplifying and sharing the findings of the jury. Community juries are similar to market research techniques such as surveys and focus groups in the sense that they provide a means of uncovering what a community thinks of a particular topic or technology. However, community juries are different from these approaches as they also provide a means of informing and educating a community, such that they can ultimately provide meaningful input and be involved in creating collaborative solutions. The Microsoft Responsible Innovation Toolkit provides valuable guidance for organizing community juries based around technical products and services. It recommends that product teams prepare and present artifacts to the jury, such as: a harms assessment,

storyboards, relevant reports from academia and media, documentation describing data flows, and prototypes [45]. It also recommends that someone outside the product team—a neutral user researcher—acts as a moderator. Before the session, the appointed moderator should plan a structure that facilitates learning and deliberation, they should also run a pilot jury with a smaller number of members to try and identify issues in advance [45]. During the session, they should attempt to ensure all perspectives are heard (including those of the product team) and reinforce the value of participation by explaining how juror feedback will be integrated into the product. After the session, they should ensure jurors are compensated accordingly and issue a report that explains the findings of the session [45]. Within the context of the GMAIF, we have sought to apply the Microsoft Responsible Innovation Toolkit when projects have been categorized as high risk with respect to the EU framework (see section III, C.). We also offer it to other beneficiaries accepted onto phase 2, but it is considered mandatory for those categorized as high risk. It has not yet been possible to organize a community jury through the GMAIF with assistance instead ending with the recommendation that the artifacts of the assistance are presented to such a jury. Typically, the artifacts of the technical assistance include: a harms assessment, an interactive notebook or dashboard, and a written report that describes the assistance provided by the GMAIF in detail.

IV. Exploratory Evaluation

In this section, the GMAIF approach (described in section III) is evaluated based on the experiences of the technical analyst team and the academics who have delivered the exercises and applied the practices with participating beneficiaries (preliminary survey results and quotes from participants are also reported where possible). To guide this evaluation, the authors considered how well each exercise/practice surfaced new ethical issues, helped with the identification of new stakeholders and promoted consideration of their perspectives, and how engaging and practical each exercise/practice was. Empirical results relating to the descriptive morality first outlined in the EU framework are also reported as a part of this evaluation.

Up until January 2021, 102 beneficiaries have participated in the GMAIF project and engaged in the phase 1 workshops whilst 26 have progressed onto phase 2. However, these numbers do not equate to the number of beneficiaries that have been taken through each of the exercises described in section III for two reasons: (1) each partner university delivers their own educational workshops in phase 1 and will not necessarily have applied the micro-futures and consequence scanning exercises, (2) not all of the exercises and practices have been in place since the inception of the project, which means participants who went through phase 1 in earlier cohorts may not have participated in some of the exercises. Exact numbers for each exercise/practice are therefore reported in the relevant subsections.

A. Micro-Futures

To assess the utility of micro-futures within the context of the GMAIF workshops, beneficiaries participating in the third round of educational workshops were given the option of taking a short survey after completing the exercise. This survey featured a series of statements that beneficiaries were able to agree or disagree with using a Likert Scale from 1-5, (where 1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly agree). The statements were as follows: (1) I understood the micro-futures exercise; (2) I had plenty of time to complete the micro-futures exercise; (3) I enjoyed the micro-futures exercise; (4) The micro-futures exercise helped me think of additional negative impacts my planned AI system / product could have on users or wider society; (5) The micro-futures exercise helped me think of additional groups / stakeholders that my planned AI system / product could affect. Ten of the beneficiaries participating in the third phase 1 cohort completed the survey. No beneficiary responded 'Strongly disagree' for any of the statements. Figure 3 shows the results and suggests that, in general, beneficiaries perceived the exercise as being worthwhile.

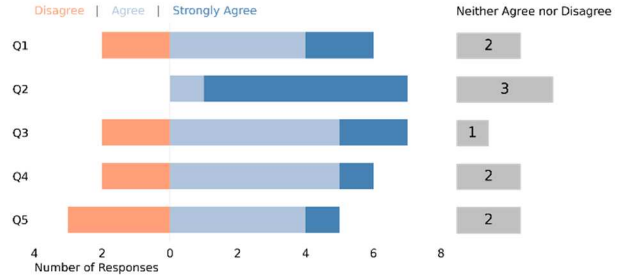

Figure 2: Responses from beneficiaries that completed the micro-futures survey.

Most of the beneficiaries found the exercise enjoyable and felt it helped them think of additional negative impacts their planned AI system could have on users or wider society. Interestingly, the statement beneficiaries disagreed with the most was 'The micro-futures exercise helped me think of additional groups / stakeholders that my planned AI system / product could affect', which reflects the fact that the exercise did not encourage participants to think about this; when presented with the key steps that should be followed to create a micro-future, beneficiaries were told only to 'imagine a character, someone that is using or affected by this product/service'. This is a shortcoming of the exercise that could be addressed in future workshops by having beneficiaries draw a stakeholder card that encourages them to think about the product/service from a different perspective (like in the judgement call game). However, making a change like this has the potential to make the exercise more time-consuming and would therefore need to be done with care.

After the beneficiaries had completed the exercise and survey, they were encouraged to share their micro-futures with the group, and it is notable that there was some reluctance to this. This could suggest that the beneficiaries had not understood the exercise and that it could be presented clearer in the future, however, it is the more likely a reflection of the attachment beneficiaries felt towards their planned products/services. If beneficiaries have already invested a significant amount of time, or money, in their planned product/service they may be hesitant to share a story that reflects on it poorly (even if it is fictional). Overall, these observations and the survey data suggest the micro-futures exercise is valuable as an engaging exercise that can surface new ethical issues but limited in its ability to help with perspective taking. Because of this, we suggest micro-futures

be applied as an early intervention and not later in the product development process, where more detailed analysis is possible.

B. Consequence Scanning

Beneficiaries took part in consequence scanning in the Data and AI ethical workshop in phase 1. In the majority of cases, the identification of positive consequences for their specific AI product/service was seen to be the easy task, aligning with their business planning, whereas thinking of negative consequences was more challenging, especially beyond the targeted consumer base and required facilitation from the moderator of the session. Feedback from a subset of beneficiaries suggested this was due to a new way of thinking, as the focus had been on development of a successful business plan (focused on the advancement of their specific new AI product/service), driven with a need to first breakeven. In most cases, beneficiaries did not have a chance to think about how negative consequences could be mitigated as this would have required more time. For example, the design of the tech was focused around a specific user, rather than general population where certain design choices may not be appropriate (a negative consequence). Citizen involvement in the conceptualization and design stages was not something a subset of beneficiaries had considered. In the 2021 UK inquiry Our Place Our Data [48] a key recommendation to build trust in AI was "…. *for citizens to be involved at the start of AI/machine-learning design, and at regular stages thereafte*r." The majority of feedback was positive, with SMEs requesting that the session be run again within their own specific businesses beyond the GMAIF project.

C. Adopting a Risk-based Approach

Table *2* reports the number of beneficiaries (in the first 12 months of the project) that were classified as 'Unacceptable', 'High', 'Limited', and 'Minimal' risk using the supplementary AI risk and ethical assessment form, which is based on the interpretation in Table 1. It shows that 18 of the 26 beneficiaries that have progressed to phase 2, since the introduction of the form in July 2021. This was due to the proposed EU regulation framework on AI only being published in April 2021, when some beneficiaries were already in phase 2 of the project. Each classification reported in Table 2 was proposed by the beneficiary, in collaboration with members of the project team, and approved by the GMAIF Technical Panel. The fact that 0 beneficiaries were classified as being of 'Unacceptable' risk is therefore to be expected. If the members of the project team working with a beneficiary considered them to pose an 'Unacceptable' risk they would not present their case for approval to the panel. The fact there have been 9 'High' risk classifications is a symptom of the broad areas outlined in Table 1, which are derived from Annex III in the EU Framework [11]. Considered alongside the number of 'Limited' and 'Minimal' risk classifications, the comparably high number of 'High' risk classifications hints at the far-reaching impact the EU Framework will likely have when written into law, unless there is very specific concrete guidance.

In general, beneficiaries found it difficult to report the potential negative impacts their AI product/service could have on society. With respect to the risk classification component of the form, the authors have found making risk classifications to be straightforward, except when beneficiaries intending to implement an AI-as-a-service business model. In these situations, the authors have found it difficult to properly formulate the intended use of the AI product/service and opted for the 'High' risk classification, given the AI product/service could feasibly be applied in any number of areas (a large language model exposed as a service could be used within a low-risk setting, like a video game, or a high-risk setting, like healthcare).

Table 2: The number of beneficiaries receiving unacceptable, high, limited, and minimal risk classifications (July 201 – January 2022).

Risk Classification	No. of classifications
Unacceptable	0
High	9
Limited	6
Minimal	3

D. The Microsoft Responsible Innovation Toolkit

To date, only the second practice in the Microsoft Responsible Innovation Toolkit—harms modeling—has been performed as a part of the GMAIF project. It has not been possible to play the judgement call game because of scheduling issues. The game could feasibly be played by an individual to work around these issues but is best played by several members of a product team. It has also not been possible to organize a community jury because of time constraints. Recruiting a representative group of participants is the main challenge in this regard, though we expect it would also take a significant amount of time to plan and organize an effective community jury session. To work around these issues, we intend to produce prototypes of a lower fidelity—so producing static storyboards as opposed to interactive dashboards—when providing technical assistance that is considered high risk (according to the supplementary forms based on the proposed EU AI regulation framework) in the future. This will allow for additional time to be spent organizing and applying the judgement call and community jury practices.

Harms modeling has so far been performed in collaboration with two beneficiaries as a part of phase 2 of the project. It has been performed primarily by technical analysts working on the GMAIF project with beneficiaries providing input and comments during regular milestone meetings. General feedback from this limited sample of beneficiaries has been positive, with the beneficiaries both commenting that the approach helped them think about their AI product/system in a different way, but work needs to be done to implement a more formal feedback mechanism.

The beneficiary feedback is supported by the experiences and observations of the technical analysts that performed the harms modeling. In general, the analysts considered the harms modeling a worthwhile exercise, as it highlighted issues they could consider and sometimes work towards mitigating in the remaining technical assistance. The analysts also found the guidance and templates provided by Microsoft to be useful. The 10 types of harm outlined by the template (see section III) encouraged a structured approach to thinking about the harms a system could inflict on society. Furthermore, the examples included in the template acted as a useful guide as to the level of detail that should be included

when reporting potential harms. However, the analysts also found the template limited in several ways. Analysts found assigning a risk level to each type of harm challenging as there was no examples of what constituted a high, medium, or low risk. Also, no space for potential mitigations that could be applied with respect to each harm is included in the template.

V. Conclusions and Further Work

Traditionally, standard institutional review board review procedures were conceived without considering the impact of AI and have not been developed to factor in these technologies and the unique ethical challenges they present. When universities are involved in knowledge transfer with industry there is collision between public and private sector ethics in terms of responsibility. This is why it is essential that such projects always go through a full ethical approval process and data processor and data controller relationships are well defined through data privacy impact assessments (when personal data is involved). However, to go beyond current legislation and get SME's (in particular) a head of the curve, the ability to risk assess new data driven and AI embedded technologies with regards to its impacts on individuals and society is critical. As recent history tells us, ethical risks often manifest 'downstream' within the project development lifecycle when AI is being applied. Encouraging the use of ethical toolkits such as consequence scanning and harms modelling can help change the thought process, contribute to more responsible and trustworthy AI applications and upskill SME's.

The activities introduced by the GMAIF approach have proved to be largely successful, for example participant feedback indicated the micro-futures exercise was engaging and provided value by encouraging beneficiaries to consider potential negative impacts their developments may have on wider society; however, as some beneficiaries felt reluctant to discuss potential downsides of their developments, this exercise could be adapted to encourage all to participate. Phase 1 activities have proven to be quick and effective, adding value to the programme and allowing completion within the time constraints of foundry workshops. Phase 2 activities encourage broader consideration of risks relating to specific types of development. This is appropriate for those beneficiaries who choose to continue onto this phase to further develop their own products/services, allowing the risk level of developments to be identified and managed.

Further work is required to expand upon the range of practices implemented within the GMAIF approach. The GMAIF project continues to utilize feedback and experience of SMEs, to improve the delivery of existing activities. Conducting detailed evaluations of all exercises and corresponding feedback responses would drive continuous improvement of all practices during the foundry phases. More funding is needed to embed ethical practices into SMEs, however the greatest challenge will be in shifting the business focus to include other success factors such as developing tech that does no harm.

Acknowledgment

The research is conducted as part of this paper was part of the Greater Manchester AI Foundry project which is funded by the European Regional Development Fund. The project website can be found here: https://gmaifoundry.ac.uk/

References

[1] Poole, D. L. and Mackworth, A. K. (2010) *Artificial Intelligence: foundations of computational agents.* Cambridge University Press.

[2] McCulloch, W. S. and Pitts, W. (1943) 'A logical calculus of the ideas immanent in nervous activity.' The bulletin of mathematical biophysics, 5(4) pp. 115-133.

[3] LeCun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning.' *nature*, 521(7553) pp. 436-444.

[4] Gogas, P. and Papadimitriou, T., (2021). Machine learning in economics and finance. Computational Economics, 57(1), pp.1-4.

[5] Abduljabbar, R., Dia, H., Liyanage, S. and Bagloee, S. A. (2019) 'Applications of artificial intelligence in transport: An overview.' *Sustainability*, 11(1) p. 189.

[6] Pan, Y., Froese, F., Liu, N., Hu, Y. and Ye, M., (2021). The adoption of artificial intelligence in employee recruitment: The influence of contextual factors. The International Journal of Human Resource Management, pp.1-23.

[7] von Struensee, S., (2021). Eye on Developments in Artificial Intelligence and Children's Rights: Artificial Intelligence in Education (AIEd), EdTech, Surveillance, and Harmful Content. EdTech, Surveillance, and Harmful Content (June 4, 2021), Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3882296

[8] Wakefield, J. (2021) 'Amazon faces spying claims over AI cameras in vans.' [online], Available: https://www.bbc.co.uk/news/technology-55938494

[9] Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S., (2021), March. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623).

[10] Hao, K. (2020), We read the paper that forced Timnit Gebru out of Google. Here's what it says. MIT Technology Review, [online], Available: https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

[11] European Union, (2021), Proposal for a Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Act, [online], Available: EUR-Lex - 52021PC0206 - EN - EUR-Lex (europa.eu)

[12] UNESCO, (2021), Draft text of the Recommendation on the Ethics of Artificial Intelligence, UNESCO Digital Library, 2021, [Online], Available: https://unesdoc.unesco.org/ark:/48223/pf0000377897

[13] The Greater Manchester AI Foundry, 2021, Available: https://gmaifoundry.ac.uk/about/

[14] Brady, H. E. (2019). The Challenge of Big Data and Data Science. *Annual Review of Political Science*, *22*(1), 297-323. https://doi.org/10.1146/annurev-polisci-090216-023229

[15] Kazim, E., & Koshiyama, A. S. (2021). A high-level overview of AI ethics. *Patterns*, *2*(9), 100314. https://www.sciencedirect.com/science/article/pii/S2666389921001574

[16] Hillerbrand, R., & Werker, C. (2019). Values in University–Industry Collaborations: The Case of Academics Working at Universities of Technology. *Science and Engineering Ethics*, *25*(6), 1633-1656. https://doi.org/10.1007/s11948-019-00144-w

[17] Kenney, M. (1987). The Ethical Dilemmas of University: Industry Collaborations. *Journal of Business Ethics*, *6*(2), 127-135. https://www.jstor.org/stable/pdf/25071641.pdf

[18] Berlin, I. (1990). *The crooked timber of humanity* (H. Hardy, Ed.). Princeton University Press.

[19] Sparrow, R. (2021). Why machines cannot be moral. *AI and Society*. https://doi.org/10.1007/s00146-020-01132-6

[20] Haldane, J. (2003). Applied Ethics. In N. Bunnin & E. P. Tsui-James (Eds.), *The Blackwell Companion to Philosophy* (Second Edition ed., pp. 490-498). Blackwell Publishers Ltd.

[21] Cortes, C., Mohri, M., Riley, M., & Rostamizadeh, A. (2008). *S*ample selection bias correction theory, [online], Available: https://arxiv.org/abs/0805.2775

[22] Koene, A., Dowthwaite, L., & Seth, S. (2018, 29-29 May 2018). IEEE P7003TM Standard for Algorithmic Bias Considerations. 2018 IEEE/ACM International Workshop on Software Fairness (FairWare),

[23] Kitchener, K. S., & Kitchener, R. F. (2009). Social Science Research Ethics: Historical and Philosophical Issues. In D. M. Mertens & P. E. Ginsberg (Eds.). SAGE Publications, Inc. https://doi.org/10.4135/9781483348971

[24] Cohon, R. (2018). Hume's Moral Philosophy. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2018 ed.): Metaphysics Research Lab, Stanford University.

[25] Gert, B., & Gert, J. (2020). The Definition of Morality. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Fall 2020 ed.)

[26] Universities UK. (2019). *The concordat for research integrity*. https://www.universitiesuk.ac.uk/sites/default/files/field/downloads/2021-08/Updated%20FINAL-the-concordat-to-support-research-integrity.pdf

[27] Lauer, D. (2021). You cannot have AI ethics without ethics. *AI and Ethics*, *1*, 21-25. https://doi.org/10.1007/s43681-020-00013-4

[28] Diener, E., & Crandall, R. (1978). *Ethics in social and behavioral research*. U Chicago Press.

[29] Hedgecoe, A. (2009). "A Form of Practical Machinery": The Origins of Research Ethics Committees in the UK, 1967–1972. *Medical History*, *53*(3), 331-350. https://doi.org/10.1017/S0025727300000211

[30] Milgram, S. (1997). Obedience to authority : an experimental view. Pinter & Martin.

[31] Butcher, S. I. (2005). The Origins of the Russell-Einstein Manifesto.

[32] Jobin, A. Ienca, M. Vayena, E. "The global landscape of AI ethics guidelines". Nat Machine Intelligence 1, 2019, pp.389–399

[33] McKinsey, (2021), [Online], Available: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/what-the-draft-european-union-ai-regulations-mean-for-business

[34] Micrsoft (2021), A Balancing Act : Regulating AI to boost responsible innovation in Europe [Online], Available: https://blogs.microsoft.com/eupolicy/2021/09/16/a-balancing-act-regulating-ai-to-boost-responsible-innovation-in-europe/

[35] IBM, Feedback from IBM, (2021), Available: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665615_en

[36] Morley, J. Floridi, L. Kinsey, L. Elhalal, L. (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, pp. 2141–2168.

[37] Crockett, K. Gerber, L. Latham. A. Colyer, E. "Building Trustworthy AI Solutions: A Case for Practical Solutions for Small Businesses," in IEEE Transactions on Artificial Intelligence, doi: 10.1109/TAI.2021.3137091.

[38] Brown S. (2019) Consequence Scanning Manual Version 1 . London: Doteveryone.

[39] Bell, F. a. (2013). Science fiction prototypes: Visionary technology narratives between futures. Futures, 5--14.

[40] Callaghan, V. (2014). *Micro-Futures.* IOS Press.

[41] Burnam-Fink, M. (2015). Creating narrative scenarios: Science fiction prototyping at Emerge. *Futures*, 48--55.

[42] Ballard, S. a. (2019). Judgment Call the Game: Using Value Sensitive Design and Design Fiction to Surface Ethical Concerns Related to Technology. Proceedings of the 2019 on Designing Interactive Systems Conference (pp. 421-433). San Diego: Association for Computing Machinery.

[43] Microsoft, "Responsible Innovation: A Best Practices Toolkit", 2020, [Online], Available: https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/

[44] Microsoft, Harms Modelling, (2021), [online], available: https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/

[45] Microsft, Community Jury, (2021), [online], Available: Community jury - Azure Application Architecture Guide | Microshttps://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/community-jury/oft Docs

[46] Crosby, N. and Hottinger, J.C., 2011. The citizens jury process. The Book of the States, 2011, pp.321-325.

[47] King, G., Heaney, D.J., Boddy, D., O' Donnell, C.A., Clark, J.S. and Mair, F.S., 2011. Exploring public perspectives on e - health: findings from two citizen juries. Health Expectations, 14(4), pp.351-360.

[48] Policy Connect, "Our Place Our Data: Involving Local People in Data and AI Based Recovery", 2021, [Online], Available: https://www.policyconnect.org.uk/research/our-place-our-data-involving-local-people-data-and-ai-based-recovery