Received: 14 July 2021 | Accepted: 15 November 2021

DOI: 10.1002/aaai.12062

ARTICLE

The AI field needs translational Ethical AI research

Jana Schaich Borg

Gross Hall for Interdisciplinary Innovation, Social Science Research Institute, Duke University, Durham, North Carolina, USA

Correspondence
Jana Schaich Borg, Gross Hall for Interdisciplinary Innovation, Social Science Research Institute, Duke University, Box 90989, Durham, NC, USA.
Email: js524@duke.edu

Abstract

Calls for Ethical AI have become urgent and pervasive, especially as ethical issues surrounding AI products at tech companies are increasingly scrutinized by the public. Yet even after a first wave of responses to these calls coalesced around Ethical AI principles to guide decision-making and a second wave generated technical tools to mitigate specific ethical issues, multiple lines of evidence indicate that these Ethical AI principles and technical tools have only a limited impact on the daily practices of AI users and producers. In other words, there is a big gap between what we publish in academic papers and what AI creators need to generate AI products that reflect society's values. Ethical AI is by no means the only field to have this problem. However, when medical and ecology fields documented similar gaps between their fields' scientific discoveries and the practices and products that people actually use, they invested tremendous resources into subfields that developed evidence about how to translate what was done in the lab to adopted solutions. I argue in this commentary that it is our research community's moral duty to invest in our own subfield of "Translational Ethical AI" that will determine how best to ensure AI practitioners can implement the Ethical AI technical tools we publish in academic venues in production settings. Further, I offer concrete steps for doing that, drawing on insights gleaned from other translational fields. Closing the "Ethical AI Publication-to-Practice gap" will be a considerable transdisciplinary challenge, but one of the AI research community has the unique expertise, political leverage, and moral responsibility to tackle.

INTRODUCTION

Many in the AI research community are invested in making sure that AI technology is used ethically. Members demonstrate their social commitments through adhering to community Ethical AI principles and generating technical strategies to help implement individual objectives set by those principles. The community, as a whole, has also taken specific steps to help researchers think about ethical issues related to AI more thoroughly, including holding top conferences dedicated to AI's social impact (like the Association for the Advancement of Artificial Intelligence (AAAI)/Association for Computing Machinery (ACM) Conference on AI, Ethics, and Society and the ACM Conference for Fairness, Accountability, and Transparency) and requiring or encouraging broader impact statements at other top conferences (like the Conference on Neural Information Processing Systems or AAAI's annual conference). Members of the AI community have even put their careers on the line by publicly advocating for or against specific AI uses and practices (Belfield 2020). Other technology-heavy fields have their own

ethical dilemmas but have not made as much demonstrable progress toward addressing their social impact, so the efforts of the AI community to improve its overall positive influence on society are commendable and encouraging.[1]

At the same time, it continues to be clear that the AI field's published Ethical AI principles and technical tools have negligible impact on the AI products that are sold, bought, and used in daily life (Schiff et al. 2020; Vakkuri et al. 2020). I have grappled with this documented disconnect as a researcher creating technical Ethical AI methods through our Moral AI lab at Duke University[2] and as an educator doing my best to prepare the next generation of data scientists to use AI responsibly. The disconnect is due to many interdependent factors that are not always obvious. Some of these factors, like the current financial ecosystem that prioritizes shareholders over citizens or the power dynamics caused by the underrepresentation of vulnerable populations in the AI technology industry, have already been discussed extensively elsewhere (Washington and Kuo 2020; Kalluri 2020; Battersby 2021). My motivation for writing this commentary is that there is an equally important contributing factor that has received comparatively little attention. Unlike solutions to the previously mentioned issues that require input and action from AI practitioners but are probably best analyzed through social science and ethics, the solutions to the issue I aim to bring to the fore require collaboration with social scientists and ethicists, but must be led by those who are responsible for the day-to-day creation and scaling of AI technology. The issue I want to focus on here is that technical tools that help mitigate AI-related ethical challenges are not used or accessible by the people who create AI products and put those products in the hands of consumers (Schiff et al. 2020; Vakkuri et al. 2020; Rakova et al. 2021).

Before continuing, let me be clear: technical tools are not sufficient to close the Ethical AI Publication-to-Practice gap, because they will not address all of the economic, social, and psychological phenomena that contribute to unethical uses of AI. However, I do assert they are a necessary part of any global Ethical AI solution. They are also one of the most scalable and pragmatic mechanisms for narrowing the Ethical AI Publication-to-Practice gap, especially given the dramatic mismatch between the rapid speed at which AI products are being made and slow pace at which public policy mechanisms can be implemented. Those who already dedicate much of their work to the systemic investigation of AI and AI applications are in privileged positions to put processes and incentives in place that will make technical Ethical AI tools more accessible and effective. Therefore, my goals in this reflection are to make this AI research community aware of some of the hurdles that prevent published Ethical AI technical tools from being "translated" to practice, offer concrete steps the AI research community can take to measurably reduce the Ethical AI technical publication-to-practice gap, and urge the community to take those steps.

WHAT DO I MEAN BY ETHICAL AI TECHNICAL TOOLS?

The overall goal of Ethical AI is to create AI that has impacts on individuals and societies that are consistent with our moral values. Ethical AI technical tools, then, are resources that can be leveraged during or after the engineering process of building an AI system to make it more likely that the system will perform in morally acceptable ways. These tools come in many different flavors, are intended to be used at different parts of the AI engineering process, and try to address different individual ethical issues. "Explainable AI" tools, for example, use combinations of algorithms, visualizations, and automated narratives to help AI teams understand and explain what types of factors their AIs use to make their predictions. Since many AIs function as "black boxes" that occlude what characteristics of the data are being used to make predictions, "explainable AI" tools aim to aid AI teams and other stakeholders identify instances when AI models use ethically questionable features like race, gender, or socioeconomic status to achieve their defined statistical objectives. Ethical AI checklists take a different approach. They aim to make it easier for AI technical teams to know how and when to build discussions of ethical issues into their work timelines by highlighting the ethical issues most relevant to each part of the AI development process and by suggesting discussion topics that can help AI teams identify ways those issues manifest in their particular use case. A third type of technical tool provides software to modify AI algorithms so that they either directly incorporate moral considerations or are easier to audit for moral principle violations. Other genres of Ethical AI tools aim to preserve privacy, prevent users with nefarious intentions from accessing AI systems, or provide decision aids to people who make decisions that affect life and death. Some Ethical AI tools are meant to be used before the technology is created, some are applied during the technology development process, and some are designed to monitor the impacts of technology after it is deployed. The tools have a wide range of production readiness, with some existing only at the initial theoretical stages and others already implemented into freely available software packages. No matter what level of development they are in, no single technical Ethical AI tool is designed to address all the ethical concerns AI poses, and the tools often address very narrow issues. Nonetheless, if some kind of technical Ethical AI tool were integrated into all AI systems, the net

overall impact of AI on society would indubitably be more ethically acceptable than it is now. More fundamentally, as I will explain, all Ethical AI principles will have to be translated into Ethical AI technical tools at some point if they are going to achieve their goal of influencing the way AI impacts society in real life[3].

TECHNICAL ETHICAL AI TOOLS ARE NECESSARY, EVEN IF INSUFFICIENT, FOR ADDRESSING THE ETHICAL ISSUES AI POSES

Some people are wary of endeavors to develop Ethical AI technical tools because they fear investing in such tools will, at best, divert resources and attention away from grappling with systemic societal issues that underlie some of AI's negative impacts, and at worst, incorrectly imply that all of AI's ethical implications can be managed by technological tools which, in turn, empowers organizations to "ethics whitewash," or misleadingly claim their practices are ethical solely by virtue of using Ethical AI technical tools. These fears are justified, and it is essential that the social issues and poor organizational leadership that contribute to AI's negative impacts be addressed through public policy, regulations, and research in tandem with the proposals I am making here. At the same time, it is imperative that we don't let such concerns cause us to neglect the types of technical changes that are simultaneously required to make Ethical AI a reality. Any instantiation of the goal to create machines that act like humans will require technology. All AI systems are built with software and hardware. All AI products, or systems of AI software and hardware functioning together to achieve a certain goal, are generated using the practical skills and processes used to create other types of technological goods and services (in addition to some AI-specific processes). For Ethical AI concepts or regulations to impact AI in practice, they must be reflected at the nitty-gritty level of these types of engineering skills and practices. Given the lateral, team-driven decision-making frequently championed by some technology companies, especially small companies, even questions about what kinds of AI products *should* be created need to be addressed through tools instantiated into everyday AI product-development practices, in addition to through ethics and public policy.

To appreciate this fundamental point in a different way, consider a country whose citizens need transportation to get to work and who rely on gas-fueled transportation because no other type of transportation is available to them. Imagine that the country passes a law that makes all gas-fueled cars illegal in an effort to control climate change. As important as addressing climate change is, it seems obvious that the new law could not be followed without socio-economic catastrophe or rebellion unless quality gas-independent technical advances, like electric cars, are simultaneously available to all of the society's members and organizations without endangering their livelihood or ability to exist. AI is predicted to increase the world economy by 13 trillion dollars by 2030 and one survey found that 84% of worldwide respondents could be confirmed to already use an AI-powered product or service on a regular basis (Bughin et al. 2018; Pegasystems 2017). Given the prominence of AI in the world economy and our daily lives, hopefully, it is apparent that—like transportation changes—society will not give up the promise and convenience of AI products for moral ends unless alternative, more ethical versions of the AI products are widely available. That's why technical tools and methods need to be part of any successful Ethical AI strategy. Until technical Ethical AI tools are accessible and usable to everybody who creates AI products society buys, the gap between Ethical AI principles and AI in practice will persist, no matter what regulations or social policies are put in place.

CONTRIBUTORS TO THE ETHICAL AI TECHNICAL PUBLICATION-TO-PRACTICE GAP

A few overarching issues drive the separation between published Ethical AI principles and technical tools from the resources available to AI practitioners. First, many technical Ethical AI tools are published as proofs of concept or prototypes without thorough instructions about how to apply the tools to individual use cases (the next section will illustrate what I mean by this). Most organizations that use or create AI products do not have staff members with adequate expertise to translate these general proofs of concept to the specific AI models in production (Hupfer 2020). AI teams may lack adequate technical skills, have insufficient practice with ethical reasoning, be unaware of social science research, or any combination thereof. Second, even when AI product teams know how to implement Ethical AI technical approaches, organizational challenges can get in their way. Managers may not approve the time or financial resources necessary to bring technical Ethical AI tools to fruition (Rakova et al. 2021), teams may not know how to make their agile work processes compatible with an ethical review (Strenge and Schack 2019), or organizational cultures may make contributors fear that calling attention to ethical issues will put their jobs in jeopardy (Rakova et al. 2021), especially given highly publicized dismissals of Ethical AI researchers, such as when Google's Ethical AI co-leads Timnit Gebru and Margaret Mitchell were fired because of controversy surrounding their resistance

to Google's censorship of their research on the possible biases in AI-based language models (Schiffer 2021).

To some extent, these issues can be addressed through educational efforts, investment in politico-economic strategies that give organizational leaders committed to Ethical AI a competitive advantage, and rigorous data collection about how organizations and leaders who are genuinely committed to Ethical AI can proactively ensure their culture, processes, and structures facilitate AI's ethical uses (Schaich Borg 2021). The focus of this commentary, though, is a third reason, which can be related to the first two, but should be considered separately. This third reason is that there is a true *knowledge* gap between what is currently published and what is needed to deploy and test technical Ethical AI solutions, even when you have ideal staffing and organizational circumstances. Crossing this gap requires research, systematic experimentation, and scientific analysis specific to each case. The US Census team learned this lesson when they tried to deploy differential privacy to protect the data publishing and mining of the US Census, but encountered a host of unanticipated challenges even though they were aided by the world's differential privacy experts (Garfinkel, Abowd, and Powazek 2018). The goal of differential privacy is to allow valid statistical analysis of a data set while preserving anonymity through the strategic injection of statistical noise into the data. The challenges the US Census team encountered when pursuing this goal included the fact that differential privacy methods had not yet been developed for the census team's type of population sampling (a scientific problem), there was no established process for determining how a parameter that determined the trade-off between privacy loss and accuracy should be chosen (a conceptual and operational problem), and census data users did not know how to adjust their analyses to account for the noise differential privacy injects into the data (a user problem), just to name a few. Differential privacy is one of the more established approaches within the Ethical AI umbrella, and yet as researcher Ashwin Machanavajjhala said at an AI and Ethics conference, "[Applying differential privacy] is more of an art than a science. I think it can be made into a science, but it just needs more work" (Machanavajjhala 2020). AI teams trying to implement Fair AI tools, which aim to mitigate discrimination and bias in AI, report similar gaps in knowledge and similar needs for human judgment and experimentation, despite efforts by Ethical AI contributors to make Fair AI tools universally accessible (Holstein et al. 2019).

Those of us in teaching roles see the impact of these knowledge gaps in our classes. Few (if any) Ethical AI tools have complete outlines or procedures describing how to implement them in real-life settings. Productive efforts to make the procedures more straightforward are underway, including some of the ethical checklists I described earlier, but even with these tools, there remains a considerable knowledge gap between what the tools offer and what needs to be understood and accomplished in order to implement them. The only way to fill this gap without additional information from the technology creators is to use what I will call "applied data science problem-solving skills" to figure out what to do on a case-by-case basis. These include communication skills that allow the data scientist or AI engineer to learn from others enough to understand the application problem deeply and get advice from appropriate experts, creativity and thoughtfulness in how to design informative model inputs under the practical and social constraints of the specific application context, and technical prowess in choosing the right sequence of models and inputs to solve real-life problems, which are very different from the idealized problems often used in problem sets. It is widely acknowledged that these applied data science skills are not adequately developed in most statistics, math, or computer science curricula (Börner et al. 2018; Gilliland et al. 2019). Trainees also struggle to accept that these applied computational proficiencies are only moderately related to traditional math and coding and are actually quite difficult to acquire. Some trainees even resist investing energy in honing these skills, because they view them as "soft." The result of the combined lack of technical tool development and underdeveloped applied data science skills is that Ethical AI technical tools leave even mathematically advanced trainees feeling lost and frustrated, especially if they have not previously had the opportunity to think deeply about the ethical concepts the tools are trying to address. I have seen this outcome time and time again with my students, and similar reports are documented in the industry.

AN ILLUSTRATIVE EXAMPLE: FAIR AI

We can use the "Fair AI" field to illustrate some of the types of challenges users of Ethical AI tools must overcome. As I mentioned earlier, the goal of "Fair AI" is to assess and mitigate discrimination and bias in AI (usually machine learning) models. A tremendous amount of energy from inside and outside academia has been dedicated to figuring out how to achieve this goal in a scalable fashion. The fruits of this investment are multiple open-source technical software packages—including scikit-fairness (Vincent et al. 2019), IBM Fairness 360 (Bellamy et al. 2019), Aequitas (Saleiro et al. 2018), Google What-if (Wexler et al. 2019), and Fairlearn (Bird et al. 2020)—that were created to help AI teams make their AI products fair. Despite this encouraging and tangible progress, recent investigations confirm

that even AI experts are unsure of how to implement Fair AI tools in real use cases (Andrus et al. 2021; Richardson et al. 2021). So why do the AI teams struggle?

One of the most fundamental questions an AI product team must answer when applying Fair AI tools is what definition of fairness to focus on. Since this is a challenging conceptual issue to address even without having to get the underlying math right, the goal of Fair AI software packages is to reduce the mathematical sophistication teams who want to create "fair" AI products require by making mathematically vetted algorithms representing different types of fairness available to others who do not have a top AI researcher on staff. Over twenty different mathematical definitions spread across separate Fair AI software packages are available, and each definition-tool combination has its own technical or statistical limitations (for example, some software packages can only be used for binary predictions; Verma and Rubin 2018). One definition assumes that an algorithm is fair if it is trained on data that omits features tightly connected to groups to whom you want to be fair. Using this approach, a model could be considered fair if its predictions do *not* leverage race, gender, or sexual orientation labels or information. Another definition assumes that an algorithm is fair if its prediction accuracy is the same across all measured groups. By this definition, a model could be considered fair if its error rate is the same (by some chosen measure) across all tested racial, gender, or sexual orientation groups. At this point, there is no clear method for choosing between available fairness definitions, so the choice becomes part of the "art" of implementing Ethical AI. AI teams report needing weeks to months to understand the conceptual differences and implications of the myriad of fairness definitions offered (Lee and Singh 2021). A recent study concluded, "Practitioners without a thorough understanding of fairness debates are unlikely to decipher which toolkit aligns with their goals and the significance of the design choices on their [use] scenario" (Lee and Singh 2021). In a concerning twist, nonexperts often misinterpret AI fairness metrics (Saha et al. 2020), so teams without deep technical expertise may have even more trouble choosing an appropriate fairness metric for their specific use case.

Choosing appropriate metrics is not the only problem. How should teams proceed when two or more suitable fairness metrics conflict, as they often do? One of the most publicized examples of this phenomenon comes from the recidivism-prediction tool COMPAS (Correctional Offender Management Profiling for Alternative Sanctions). A set of investigative journalists reported that COMPAS was *un*fair to Black defendants according to a false positive rate fairness definition (among defendants who did not get rearrested, Black defendants were twice as likely to be misclassified as high risk). However, the company that created COMPAS retorted that the technology was *fair* to Black defendants according to a positive predicted value fairness definition (among those called higher risk, the proportion of defendants who got rearrested is approximately the same regardless of race). Both claims were true (even if you believe that AI-driven recidivism prediction tools are unethical for other reasons). If an AI team is held accountable for the results of their AI products being fair, how do they navigate these kinds of debates from their technical positions—rather than political or legal positions—within an organization? Little guidance is available, and when advice is provided, it tends to be vague and open-ended like "explore a number of different options through different choices of models and parameters and use these options to motivate a conversation about the program's goals, philosophy, and constraints" (Foster et al. 2020). This type of advice is correct as far as it goes, but many (or, I might argue, most) technical experts do not have adequate training in how to initiate and structure these conversations, how to figure out who should be involved in them, or how to make the discussions culminate in a choice about which model to move forward with, so this kind of advice isn't very actionable. Further, even if such conversations happen, most managers and directors will be even less prepared than their AI teams to understand the differences, implications, and conflicts of fairness definitions, and will struggle even more with making decisions that successfully align with their intended ethical or organizational fairness goals.

Another choice that AI teams need to make when implementing Fair AI tools is what groups of people they are going to ensure are treated fairly. These groups are usually referred to as "protected" groups that are defined by "sensitive" features, and many Fair AI tools require their delineations to be explicit. Gender and race are common sensitive features that legal teams ask AI teams to track, but of course, many other groups—like people with a particular health condition, education, or income level—could be affected unfairly by an AI as well. Without any tools or procedures for identifying which groups are likely to be treated unfairly by a particular AI instantiation, defining protected groups and their attributes becomes another part of the "art" of implementing Ethical AI. Unfortunately, it's an art that is new to most AI experts. AI teams report feeling deeply unqualified to identify at-risk groups, and therefore often resort to trial and error. In one interviewee's words, "How do you know the unknowns that you're being unfair towards? [. . .] You just have to put your model out there, and then you'll know if there's fairness issues if someone raises hell online" (Holstein et al. 2019). A related complication is that AIs can be fair to very broad categories, like race or gender, while simultaneously being unfair to combinations or intersections of those same

groups, like Black women. This tempts teams and the entities that fund them to "fairness gerrymander," or choose whichever group definitions are most consistent with their business goals as opposed to which definitions have the best outcomes for society (Kearns et al. 2018).

Even if groups of people who need to be protected have been identified, figuring out how to track those groups typically requires technical ingenuity that is not provided by Fair AI software packages. For instance, Fair AI software usually assumes that your data already have demographic labels that distinguish the groups that need to be protected. In other words, if you want to make sure your AI is fair to different genders, the packages assume that you know which gender(s) is associated with each of your data points. However, this is frequently not the case, sometimes because the appropriate demographic information was not collected, sometimes because the appropriate demographic information cannot be shared (due to privacy regulations or sharing agreements), and sometimes because the request for relevant demographic information is not legal (in order to prevent discrimination in credit-based decisions, for example, Andrus et al. 2021). In such cases, AI teams must either find ways to collect demographic information before and after an AI model is deployed (first to choose the preferred model and then to track whether it has the intended fair outcomes) or develop other ways of inferring which demographic category each datum is associated with, a process that may be plagued by its own bias and types of unfairness. The technical implementation of identifying and tracking sensitive groups is rarely straightforward, and this is yet another type of engineering process that has to be developed in order to translate Fair AI software packages into real implementations.

It is important to know that the issues I have discussed so far represent just a small subset of the challenges that create the gap between what Fair AI technical tools offer and what is needed to implement them. For readers who are interested in more thorough discussions, I strongly recommend the research papers I cited earlier by Holstein et al. (2019) and Lee and Singh (2021). Even armed with only this subset, though, I hope the depth of issues Fair AI technical tools leave unresolved is now more evident. Rest assured that similar collections of issues plague other Ethical AI technical approaches as well.

LEVERAGING TRANSLATIONAL RESEARCH AND PRACTICE TO REDUCE THE ETHICAL AI PUBLICATION-TO-PRACTICE GAP

The frustrating end result of the Ethical AI publication-to-practice gap is that Ethical AI technology, despite good intentions and clever approaches, is not integrated into the AI products society interacts with. Most experts who were willing to provide their input to a survey collected by the Pew Research Center and Elon University in 2020 agreed. When 602 technology experts were asked, "By 2030, will most of the AI systems being used by organizations of all sorts employ ethical principles focused primarily on the public good?," 68% answered no. So, at least to some extent, the Ethical AI publication-to-practice gap is an acknowledged reality.

I propose that the gap does not need to be as wide as it currently is. There are technical and nontechnical ways to address Ethical AI's translation issues. Our mandate now is to cultivate the type of problem-solving and research that will make translation more successful. In other words, we need not only the type of Ethical AI research that leads to the generation of tools like Fair AI software packages, but also translational Ethical AI research that figures out what needs to happen in order for those tools to be implemented in the production-level settings that generate the AIs actively interacting with society.

Translational Ethical AI research will require collaboration with AI technology users, social scientists, and ethicists, but it will also require sustained momentum from within the AI technical community. Unfortunately, right now such momentum is severely lacking because the incentives in the AI Research community are not well-aligned with translational work. Even though AI researchers are actively encouraged to create new technical Ethical AI tools or concepts, they are simultaneously discouraged from doing the research necessary to figure out what feature, design, preprocessing, mathematical, and political decisions must be made to implement their technical Ethical AI tools in production settings, or to monitor them to ensure they are having the intended outcome once they have been deployed. The "publish or perish" environment academic researchers function in disincentivizes projects that do not result in traditional disciplinary publications in quantities proportionate to the time they require to complete, and academic researchers are rewarded for theory and proof-of-concept implementations more than full-scale deployments, so there is little reward for studying Ethical AI problems past the prototype stage (Black et al. 2020). Industry AI researchers who create Ethical AI technical tools, on the other hand, have few, if any, professional enticements or sources of support to create tools that can be used by other groups (Rakova et al. 2021). Despite growing recognition of the importance of Ethical AI for organizations' success (McCormick 2021), disseminating advice about how to implement Ethical AI tools is not likely to be a strategic priority. The result is a dearth of knowledge and methodology about how to streamline the application of Ethical AI technical tools in

production settings, especially when experts are not available to implement them.

This is a moral problem. If we as a research community truly want to make sure AI is used ethically, we must invest in its implementation in production settings, not just in generating ideas that we hope others will implement to moral ends.

AI researchers may find some solace in learning that Ethical AI is not alone in its battle with "translational" work. Other fields, such as medicine and environmental science, have recognized the difficulty of bringing work from the lab to practice and community health (and back again to refine the lab research being pursued) for well over a decade. Despite much progress and investment, these other fields still struggle with avoiding the infamous "valley of death" between publication (or discovery, in the case of industrial research) and practice, and with ensuring the benefits of basic research are fully reflected in people's daily lives (Butler 2008; Schlesinger 2010). Those familiar with this literature would find it unsurprising that Ethical AI tools struggle to enter real implementation settings. Translational work is notoriously hard, time-consuming, unappreciated, and requires transdisciplinary teamwork between people with a wide range of skill sets. The silver lining, though, is that the Ethical AI field can leverage the lessons other translational disciplines have learned to make progress more successfully. It is also worth pointing out that the AI field is different from the natural sciences in many critical ways, including some that will make its translational work both easier and faster. The key to achieving tangible results will be to have humility about how difficult translational problems are to solve.

WHAT WOULD TRANSLATIONAL ETHICAL AI LOOK LIKE?

In the medical field, translational research is described as a subset of applied research that requires a potential solution to a specific medicine-related problem to be evaluated in, and optimized for, real-life settings. Translational research consists of multiple phases. The most simple model of these phases depicts the first phase as "lab to bedside" and the second phase as "bedside to community." The lab-to-bedside phase might lead to the development of a potential intervention, like a medication for dementia or vaccine for COVID-19. The bedside-to-community phase would not only test whether the medication or vaccine works on people, but would also examine whether patients and doctors want the medication or vaccine in the first place, and identify barriers to patients and doctors using the medication or vaccine in a way that achieves the intended health outcome. Full "translation" of a potential intervention from laboratory to fruitful everyday practice requires many iterative cycles of modifying and testing the intervention and how it is administered until the desired impact on medical practice or patient health is achieved at the population level (National Center for Advancing Translational Sciences 2021b). Some modifications may require changes to the intervention's underlying technology, such as when researchers adjust a medication to reduce side effects or redesign a vaccine so that it does not need to be refrigerated, while other modifications require changes to how the technology is made available and scaled, such as when medications or vaccines are made available in community barbershops so that they are accessible to people who are wary of medical settings.

Borrowing these concepts, I propose that translational Ethical AI is the research and work done to determine whether Ethical AI tools are being used in practice with their intended ethical impact and to determine how to maximize their adoption. It represents the space between technical Ethical AI proofs of concept and the AI products that get used by society ethically. It emphasizes AI creators and users over (but not to the exclusion of) general basic scientific principles. My motivation for defining translational Ethical AI is not to distinguish it as its own field. In fact, I am indifferent towards that prospect. Rather, I offer a definition of translational Ethical AI to help catalyze bold enterprise in the research space the definition describes. In doing so, I want to call attention to the fact that progress in the later stages of translational AI requires a different set of resources, skills, and knowledge than the creation of research prototypes. I also think it is strategically important to acknowledge that the processes required for translational AI to succeed are different than those needed for more basic research. Traditional research often prioritizes the cultivation of new theories and principles and is evaluated according to experimental and statistical design ideals. It may indeed be helpful to uncover and describe general principles that govern the adoption of AI technology offerings, but pursuit of these principles should not detract from the real-life implementation and evaluation translational Ethical AI must be grounded in, which due to the complexities of real-life, require concessions in how well experimental variables can be controlled. The most effective ways to describe and analyze the principles that arise from these experiences and observations will likely transcend traditional disciplinary boundaries and should not be expected to advance theories from any particular field. Further, since some of our assumptions about how general principles apply to specific use cases will always turn out to be wrong, translational Ethical AI research will require iterative cycles that incorporate feedback from different stakeholders, and should use concepts like adoption rates and performance in practice as

primary evaluation metrics rather than academic publications or grant awards (which can also be used as metrics, but should be considered secondary priorities).

The topics, processes, and goals of translational Ethical AI research will differ from other types of technical Ethical AI research (or other types of philosophical or social science research related to AI's implications), and so will the people implementing the research. Translational research requires iterative feedback from, and partnership with, nonresearch stakeholders. Neither the skills used to identify and collaborate with these stakeholders nor the technology adoption metrics that reflect the result of that collaboration are typically taught or rewarded in technical research environments. On the other hand, the technical innovation and modification required to move iteration cycles forward are not typically within the expertise of disciplines with more experience gathering qualitative information or navigating social and ethical issues. Thus, unless—or until—a generation of researchers is trained with all of the requisite background and skills to implement Translational Ethical AI research independently, Translational Ethical AI research will require highly functioning teams of researchers with diverse backgrounds and training in both highly technical and nontechnical domains. Of note, productive multiple disciplinary teamwork is notoriously difficult to manifest, so managing such teams represents a challenge unto itself that must draw on yet another unique set of skills (Begerowski et al. 2021).

It has recently been acknowledged that there is very little translational research being done in the field of computer science as a whole (Abramson and Parashar 2019). I want to bring attention to the fact that there is even less translational research being done in the realm of Ethical AI, specifically. My claim is that this is one of the causes, not just one of the symptoms, of the Ethical AI publication-to-practice gap. Fortunately, it is also a cause we can do something about.

WHAT SHOULD WE DO?

What are the first steps the AI research community should take towards translational research that closes the Ethical AI publication-to-practice gap? Here are my suggestions.

Incentivize translational Ethical AI publications and reflection

One of the most straight-forward ways to incite motivation for production-level Ethical AI research is to cultivate incentives for publishing translational technical work. To begin, the AI community can organize more high-quality archival conferences and journal issues dedicated to the translation of Ethical AI technology, such as the Policy and Practice Track at The Association for Computing Machinery's Conference on Equity and Access in Algorithms, Mechanisms, and Optimization or the brand new Translational Computer Science track in the journal *Computing in Science & Engineering*. Top AI conferences could also consider recommending that submissions include a section that outlines the steps required to apply the work described in the paper to production settings (potentially with page limit extensions; Abuhamad and Rheault 2020), and provide awards for the best translation contributions and open-source tools. AI journals could be called upon to encourage Publication-to-Practice submissions. Even easier, top AI conferences could dedicate more panels and training workshops to production-level translation of Ethical AI research that include participants from AI product teams.

Create infrastructure to facilitate tri-directional feedback between AI researchers, AI practitioners, and community members

Of course, researchers will struggle to take advantage of these opportunities if there isn't a simultaneous campaign to expose AI researchers to how "AI in the wild" manifests and who implements it. The primary audience for AI research publications are other AI researchers with PhDs. However, the primary audience for Ethical AI applications are AI product teams who are comprised of people who often do not have PhDs in computer science and who are functioning under powerful practical and social constraints that are typically not discussed in traditional computer science educational settings. The academic research community needs widespread access to information about what AI product teams' processes look like and which aspects of those processes cause challenges in implementing published Ethical AI research if we are going to help overcome those challenges.

That said, AI practitioners are not the only people impacted by the Ethical AI tools AI researchers create. General community members are usually the ones who have to live with the consequences of the AI technology put into production, but typically only have opportunities to give feedback after an AI system is deployed (Black et al. 2020). If community feedback was solicited much earlier in the AI creation process, the chances that ethically problematic AI products could be intercepted before they are implemented would greatly increase (as some people argue should have been the case for AI-powered criminal sentencing or military tools). Therefore, the next

critical contribution the AI research community should make is to cultivate a resource of AI practitioners and general community members who are willing to give Ethical AI researchers high-quality feedback about their specific research contributions. To avoid putting all the onus on already overburdened researchers, we must also recruit "cultural brokers" who can mentor AI researchers in how to solicit and integrate that feedback (Matthews et al. 2018; Pinsoneault et al. 2019). Other translational fields have found this infrastructure step to be critical, despite obviously being resource-intensive (Jasny et al. 2021; Rose, Evans, and Jarvis 2020; LeClair et al. 2020). Most notably, many academic medical research institutions now have dedicated translational groups who help individual basic researchers connect with clinicians and central community-engagement groups who help researchers and practitioners build working relationships with community members impacted by their research (Heller and de Melo-Martín 2009). The Translational Ecology and Sustainability fields have built translational infrastructure as well, though they rely more heavily on "boundary-spanning" organizations who facilitate channels of communication between researchers and nonresearchers (Safford et al. 2017). Sometimes these organizations are attached to academic institutions, like the University of California Cooperative Extension that "serves as the bridge between local issues and the power of UC research" (Division of Agriculture and Natural Resources), but more often they are run as independent nonprofits or governmental agencies.

The AI research community could aim to create similar opportunities for feedback and collaboration with practitioners and community members, and design the opportunities as a central resource, drawing on lessons from previous translational programs from other fields (Nease et al. 2018; Fisher et al. 2019; Towfighi et al. 2020; Lawson et al. 2017). Some aspects of the supporting infrastructure could reside within universities and piggyback off of already-established mechanisms some academic institutions fund like service-based learning programs, education "practitioner–researcher partnerships," or translational medicine initiatives. Other aspects of the infrastructure could reside outside of universities in nonprofit or for-profit consulting settings, inspired by translational ecology models. Initially these services would facilitate the process of AI researchers learning what AI practitioners and users need. After researchers have engaged in enough facilitated feedback interactions or built their own relationships with AI practitioners, they may be able to initiate and maintain such efforts independently.

When creating this infrastructure, collaboration should be solicited from disciplines and practitioners who incorporate and teach the suite of skills necessary to understand different types of users and stakeholders. Design thinking, empathy for other perspectives, usability testing, and qualitative interviews are all critical for understanding AI practitioners' pain points, but very few technical AI researchers develop these skills during their training. Therefore, efforts to communicate what is needed "in the AI wild" would be well-served by recruiting help from people who have deep training in these skillsets.

Advocate for translational infrastructure funding

How would we help pay for this proposed infrastructure? That's where our AI community's advocacy would come in. We should petition federal agencies, private partners, and foundations to sponsor these types of infrastructure initiatives, as well as sponsor grants and awards that will support the research and the administrative investment necessary to make the infrastructure successful.

Although that may initially sound prohibitively daunting, the National Institutes of Health have already set a precedent for making this kind of substantial infrastructure investment. For over a decade, the National Center for Advancing Translational Sciences (NCATS) has funded multimillion dollar Clinical and Translational Sciences Awards (CTSA) that require (and support) awarded institutions to have community engagement cores that nourish long-term bidirectional relationships between CTSA institutions and affected community members (Holzer and Kass 2014; National Institutes of Health 2007; National Institutes of Health 2021). The NCATS budget has been between $400 and over $700 million a year (Liverman et al. 2013; National Center for Advancing Translational Sciences 2021a), and has led to the creation of core facilities and consulting services in over 60 academic institutions that support researcher–clinician engagement and researcher–community engagement (Pelfrey et al. 2017; National Center for Advancing Translational Sciences 2021a). Not only do these cores accelerate the translation of basic science research to treatments clinicians use and patients accept (Berg 2020), they also increase high-impact scholarly output as measured by publications (Llewellyn et al. 2018; Llewellyn et al. 2020).

It would be reasonable to advocate for the NIH to allocate some of these translational infrastructure funds to creating cores that support practitioner and community engagement with medical uses of AI. In fact, their recent announcement of the Bridge to Artificial Intelligence (Bridge2AI) program signals that federal institutes may be receptive to calls to do so. The NIH could also advise and work with the National Science Foundation (NSF) to add a community–engagement track within the NSF Computer and Information Science and Engineering

(CISE) Community Research Infrastructure (CCRI) awards that currently have a budget of around 25 million dollars.

Beyond federal funding, the AI research community should urge foundations and corporations to sponsor the expansion of "boundary-spanning" organizations that facilitate tri-directional feedback between researchers, practitioners, and community members. Organizations like OpenAI, the AI Now Institute, or the Future of Life Institute have taken the first step towards encouraging these types of tri-directional interactions, but much more funding and investment will be required for feedback to be solicited about individual Ethical AI research projects or to disseminate results of these discussions to AI researchers more broadly.

Improve professional recognition

Advocating for translational funding is one of the research community's imperatives, but it cannot be pursued in isolation. Another one of our essential tasks will be to insist that research institutions and departments value translational Ethical AI work in professional evaluations, including promotion packages and press coverage (Marrero et al. 2013). As a field, we must also recognize that engagement work takes longer to implement and requires more transdisciplinary teamwork than many other types of AI research (Black et al. 2020; Teufel-Shone 2011; Fam et al. 2020). That means publications may be more infrequent, have longer author lists, and may even need to appear in different venues than other types of AI research. Fundamentally, we must make sure our peers within the AI research field assign appropriate value to this kind of translational Ethical AI work and appreciate the unique kinds of skills and investments it requires. Previous translational efforts make clear that these cultural and administrative changes are absolutely essential for empowering AI academic researchers, especially junior researchers, to dedicate time and energy into closing the Ethical AI Publication-to-Practice gap in a way that is compatible with their career goals (Littell, Terando, and Morelli 2017; Raynor 2019; Vogel et al. 2019). Notably, it will be easier to convince academic institutions to value translational work if we follow-through with providing more high-quality opportunities for researchers to publish it, as I suggested earlier.

Prepare future AI contributors

Moving forward, we should also do better at preparing technical trainees to implement Ethical AI tools. Figuring out how to do this in already over-packed computer science and data science curriculums will not be easy, but it is a challenge we have an obligation to tackle. Rather than focusing on ethics in general or technical Ethical AI tools in particular throughout training programs, a more manageable and sustainable strategy might be to focus on cultivating applied data science skills more broadly, and then incorporate opportunities to practice applying those skills to technical problems that have ethical considerations that can be mitigated with Ethical AI tools. More specifically, we should revise AI and data science curriculums to include training in identifying and talking to stakeholders, and to offer classes and experiences dedicated to teaching students how AI is developed and implemented in practice. We must make sure that students gain appreciation for how scientific judgment and experimentation is often needed to apply mathematical concepts to real problems, and how nonmathematical factors learned about through listening to stakeholders influence what types of technical solutions are possible or justifiable. We must also instill in trainees that part of their job as AI technology producers is to monitor the impact their work has on others, and care whether their research, especially their Ethical AI research, can be used by AI practitioners. These applied skills and experiences will make the process of navigating Ethical AI publication-to-practice gaps much easier and more familiar, even as available tools change. In addition, we should continue testing the best way to expose technical students to the types of social science evidence and ethical thinking that motivate and impact implementations of Ethical AI technical tools. The recent enthusiasm for embedding ethical modules throughout technical training programs seems very promising (Grosz et al. 2019), but evidence is still lacking for whether such efforts succeed at preparing students adequately for the types of ethical issues they encounter in their applied work or that are relevant to implementing Ethical AI tools.

It is important to acknowledge again that the steps I have proposed here are neither definitive nor comprehensive. For example, they will not address all of the systemic or organizational issues that prevent Ethical AI tools from being implemented, nor will they make all organizations that use AI prioritize ethics. A successful Ethical AI strategy will require many parallel initiatives, including regulation, policy, organizational guidance, education, and community engagement. The steps I have proposed here, though, need to be part of the strategy. Ethical AI requires effective technical tools as well as effective social and ethical tools, and we must make sure that Ethical AI technology is available and usable to the majority of AI creators. The reason I have made this plea in *AI Magazine* is because the AI

research community is in the best position to make this happen.

THE ETHICAL AI PUBLICATION-TO-PRACTICE GAP IS OUR RESPONSIBILITY

It can be tempting to declare that the challenges that currently prevent Ethical AI tools from being used in production settings require organizational solutions rather than technical solutions, so the AI technical research community should not have to bear the pain of solving them. After all, as some have pleaded, "we are just engineers" or "just researchers" (Dignum 2018; Hutson 2018). But we must reject such reasoning, as relieving as it would be to be absolved of moral responsibility. First, it seems somewhat obvious to point out that such claims seem ironic and insincere when they come from a discipline whose raison d'être is to use technology to mimic human-level intelligent behavior and decision-making. The AI field nourishes subfields dedicated to group decisions, artificial social intelligence, and decisions made in organizational contexts. It will be increasingly difficult for society to trust AI researchers' intentions if we continue to encourage these subfields while simultaneously claiming that the particular set of organizational decisions that prevent our AI technology from being used happen to be outside our field's purview.

Second, and much more fundamentally, we are not absolved from moral responsibility just because we don't currently know how to solve a problem or just because our contributions to the morally unacceptable consequences of our work are indirect. The AI community has pledged to maximize AI's positive influences on society and to minimize its harms, as I believe it should. The AAAI Code of Professional Ethics and Conduct also clarifies that AI professionals are obliged to undo or mitigate the harms our technology causes, even if that harm is unintended, and that designing processes that lead to those harms through negligence is ethically objectionable (AAAI 2019). As a field, then, we have agreed that is unacceptable for us to ignore the mounting evidence that we are not yet succeeding at preventing the documented unethical consequences of much of our field's technology. If we know that there is a gap between Ethical AI research and what is needed for AI to be used in a way that does not harm society, it is our ethical and professional responsibility not just to document that gap, but to close it.

The author of a recent *New Yorker* article titled "Who Should Stop Unethical A.I.?" wrote, "For now, A.I. research is mostly self-regulated—a matter of norms, not rules" (Hutson 2021). Especially while this remains true, the AI research community is responsible for setting expectations that Ethical AI research not be considered complete until it is implemented in practice. Humans are motivated (unconsciously and consciously) to avoid information in moral situations that interferes with our convenience or self-interest (Rabin 1995). Let's not fall prey to that dynamic by ignoring the evidence that part of the reason AI technology is having negative impacts on society is because our Ethical AI technical efforts are not functionally accessible to AI practitioners. We have the means to make changes within the research community and to solicit financial and institutional support from outside of the research community to minimize the Ethical AI publication-to-practice gap. My entreaty to the AI research community is that we make those investments so that we don't look back and regret our inaction, and so that AI research realizes its potential to enrich human life in the way that inspired so many in this field to dedicate their scientific efforts to the promise of computer intelligence.

ACKNOWLEDGMENTS

The author would like to thank Vincent Conitzer, John Dickerson, and Duncan McElfresh for their helpful comments on this piece, Walter Sinnott-Armstrong for his comments and more extensive editing of this piece, and Templeton World Charity Foundation, Duke Bass Connections, and Duke Research Collaboratories for their support of the author's Moral AI research.

CONFLICT OF INTEREST

The author declares that there is no conflict.

ORCID

Jana Schaich Borg https://orcid.org/0000-0002-0066-761X

ENDNOTES

[1] For example, consider my original disciplinary field of neuroscience. It is widely acknowledged that neuroscience-based technologies can impact society in both profoundly positive and negative ways, but most neuroscience publishing venues do not have a requirement, or even a good mechanism, for requiring neuroscience researchers to consider the potential negative impacts of their work. Further, whereas the Association for the Advancement of Artificial Intelligence Code of Professional Ethics and Conduct states explicitly that "An essential aim of AI professionals is to minimize negative consequences of computing, including threats to health, safety, personal security, and privacy" and "AI professionals should consider whether the results of their efforts will respect diversity, will be used in socially responsible ways, will meet social needs, and will be broadly accessible," the Society for Neuroscience Ethics Policy makes no comparable commitment to minimizing harm and ensuring the benefits of neuroscience research are broadly accessible. The Society for Neuroscience Ethics Policy focuses only on the importance of reporting data accurately, attributing authorship truthfully, reporting conflicts of interest, treating humans

and animals humanely, and preventing harassment between colleagues.

[2]The Duke Moral AI lab draws on multiple disciplines to develop strategies and principles for building ethics into artificial intelligence. The lab is co-directed by Walter Sinnott-Armstrong (a philosopher), Vincent Conitzer (a computer scientist), and myself (a neuroscientist co-opted by data scientists).

[3]Unless you hold the uncommon opinion that Ethical AI principles dictate society should prohibit the use of all types of AI for any purpose.

REFERENCES

AAAI. 2019. "AAAI Code of Professional Ethics and Conduct." https://www.aaai.org/Conferences/code-of-ethics-and-conduct.php

Abramson, D., and M. Parashar. 2019. "Translational Research in Computer Science." *Computer* 52(9): 16–23.

Abuhamad, G., and C. Rheault. 2020. "Like a Researcher Stating Broader Impact For the Very First Time." In *Workshop on Navigating the Broader Impacts of AI Research Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*.

Andrus, M., E. Spitzer, J. Brown, and A. Xiang. 2021. "What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 249–60.

Battersby, S. 2021. "How Stakeholder Capitalism and AI Ethics Go Hand in Hand." Venture Beat.

Begerowski, S. R., A. M. Traylor, M. L. Shuffler, and E. Salas. 2021. "An Integrative Review & Practical Guide to Team Development Interventions for Translational Science Teams: One Size Does Not Fit All." *Journal of Clinical and Translational Science* 5: e198.

Belfield, H. 2020. "Activism by the AI Community: Analysing Recent Achievements and Future Prospects." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 15–21 New York

Bellamy, R. K., K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, et al. 2019. "AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias." *IBM Journal of Research and Development* 63(4/5): 4:1–15.

Berg, A. 2020. "Assessing the Impact of the NIH CTSA Program on Clinical Trials Registered with ClinicalTrials. gov." *Clinical and Translational Science* 13(4): 818–25.

Bird, S., M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker. 2020. "Fairlearn: A Toolkit for Assessing and Improving Fairness in AI." Tech. Rep. MSR-TR-2020-32, Microsoft.

Black, E., J. Williams, M. A. Madaio, and P. L. Donti. 2020. "A Call for Universities to Develop Requirements for Community Engagement in AI Research." In *Proceedings of the Fair and Responsible AI Workshop at the 2020 CHI Conference on Human Factors in Computing Systems*.

Börner, K., O. Scrivner, M. Gallant, S. Ma, and X. Liu. 2018. "Skill Discrepancies between Research, Education, and Jobs Reveal the Critical Need to Supply Soft Skills for the Data Economy." *Proceedings of the National Academy of Sciences* 115(50): 12630–7.

Bughin, J., J. Seong, J. Manyika, M. Chui, and R. Joshi. 2018. *Notes from the AI Frontier: Modeling the Impact of AI on the World Economy*. McKinsey Global Institute. Available online at: https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy

Butler, D. 2008. "Translational Research: Crossing the Valley of Death." *Nature News* 453(7197): 840–2.

Dignum, V. 2018. ""I'm just an engineer" is unacceptable! Urgent need to redo education/training curricula to include reflection on impact of #AI. #AIethics [@vdignum]." Twitter.

Division of Agriculture and Natural Resources. "About UC Cooperative Extension." https://ucanr.edu/sites/ucanr/County_Offices/. Accessed 2022. © 2022 Regents of the University of California.

Fam, D., E. Clarke, R. Freeth, P. Derwort, K. Klaniecki, L. Kater-Wettstädt, S. Juarez-Bourke, et al. 2020. "Interdisciplinary and Transdisciplinary Research and Practice: Balancing Expectations of the 'Old'Academy with the Future Model of Universities as 'Problem Solvers.". *Higher Education Quarterly* 74(1): 19–34.

Fisher, M., S. E. Brewer, J. M. Westfall, M. Simpson, L. Zittleman, S. T. O'Leary, D. H. Fernald, A. Nederveld, and D. E. Nease Jr. 2019. "Strategies for Developing and Sustaining Patient and Community Advisory Groups: Lessons from the State Networks of Colorado Ambulatory Practices and Partners (SNOCAP) Consortium of Practice-Based Research Networks." *The Journal of the American Board of Family Medicine* 32(5): 663.

Foster, I., R. Ghani, R. S. Jarmin, F. Kreuter, and J. Lane. 2020. *Big Data and Social Science: Data Science Methods and Tools for Research and Practice*. edited by I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter, and J. Lane. BocaRaton, FL: Chapman and Hall/CRC, Taylor & Francis Group.

Garfinkel, S. L., J. M. Abowd, and S. Powazek. 2018. "Issues Encountered Deploying Differential Privacy." In *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, 133–7.

Gilliland, C. T., J. White, B. Gee, R. Kreeftmeijer-Vegter, F. Bietrix, A. E. Ussi, M. Hajduch, et al. 2019. "The Fundamental Characteristics of a Translational Scientist." *ACS Pharmacology & Translational Science* 2(3): 213–6.

Grosz, B. J., D. G. Grant, K. Vredenburgh, J. Behrends, L. Hu, A. Simmons, J. Waldo, et al. 2019. "Embedded EthiCS: Integrating Ethics across CS Education." *Communications of the ACM* 62(8): 54–61.

Heller, C., and I. de Melo-Martín. 2009. "Clinical and Translational Science Awards: Can they Increase the Efficiency and Speed of Clinical and Translational Research?" *Academic Medicine* 84(4): 424–32.

Holstein, K., J. Wortman Vaughan, H. Daumé III, M. Dudík, and H. Wallach. 2019. "Improving fairness in machine learning systems: What do industry practitioners need?" In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–16.

Holzer, J., and N. Kass. 2014. "Community Engagement Strategies in the Original and Renewal Applications for CTSA Grant Funding." *Clinical and Translational Science* 7(1): 38–43.

Hupfer, S. 2020. "Talent and workforce effects in the age of AI: Insights from Deloitte's State of AI in the Enterprise." 2nd Edition survey. 03 March 2020. DeloitteInsights. Available online at: https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/ai-adoption-in-the-workforce.html

Hutson, M. 2018. "Artificial Intelligence Could Identify Gang Crimes—and Ignite an Ethical Firestorm." *Science*. https://www.science.org/content/article/artificial-intelligence-could-identify-gang-crimes-and-ignite-ethical-firestorm

Hutson, M. 2021. "Who Should Stop Unethical A.I.?" The New Yorker, February 15, 2021.

Jasny, L., J. Sayles, M. Hamilton, L. Roldan Gomez, D. Jacobs, C. Prell, P. Matous, E. Schiffer, A. M. Guererro, and M. L. Barnes. 2021. "Participant Engagement in Environmentally Focused Social Network Research." *Social Networks* 66: 125–38.

Kalluri, P. 2020. "Don't Ask If Artificial Intelligence is Good or Fair, Ask How it Shifts Power." *Nature* 583(7815): 169.

Kearns, M., S. Neel, A. Roth, and Z. S. Wu. 2018. "Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness." In *Proceedings of the International Conference on Machine Learning*, 2564–72 PMLR.

Lawson, D. M., K. R. Hall, L. Yung, and C. A. F. Enquist. 2017. "Building Translational Ecology Communities of Practice: Insights from the Field." *Frontiers in Ecology and the Environment* 15(10): 569–77.

LeClair, A. M., V. Kotzias, J. Garlick, A. M. Cole, S. C. Kwon, A. Lightfoot, and T. W. Concannon. 2020. "Facilitating Stakeholder Engagement in Early Stage Translational Research." *Plos One* 15(7): e0235400.

Lee, M. S. A., and J. Singh. 2021. "The Landscape and Gaps in Open Source Fairness Toolkits." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13.

Littell, J. S., A. J. Terando, and T. L. Morelli. 2017. "Balancing Research and Service to Decision Makers." *Frontiers in Ecology and the Environment* 15(10): 598.

Liverman, C. T., A. M. Schultz, and S. F. Terry. 2013. *The CTSA Program at NIH: Opportunities for Advancing Clinical and Translational Research*. Washington, DC: The National Academies Press.

Llewellyn, N., D. R. Carter, D. DiazGranados, C. Pelfrey, L. Rollins, and E. J. Nehl. 2020. "Scope, Influence, and Interdisciplinary Collaboration: The Publication Portfolio of the NIH Clinical and Translational Science Awards (CTSA) Program from 2006 through 2017." *Evaluation & the Health Professions* 43(3): 169–79.

Llewellyn, N., D. R. Carter, L. Rollins, and E. J. Nehl. 2018. "Charting the Publication and Citation Impact of the NIH Clinical and Translational Science Awards (CTSA) Program From 2006 Through 2016." *Academic Medicine: Journal of the Association of American Medical Colleges* 93(8): 1162–70.

Machanavajjhala, A. 2020. "Enabling privacy in AI and ML using Differential Privacy." AI and Privacy, Artificial Intelligence & Ethics. Duke Kunshan University. November 4–6. https://sites.duke.edu/dkuhumanities/projects/petal/petal-conference/

Marrero, D. G., E. J. Hardwick, L. K. Staten, D. A. Savaiano, J. D. Odell, K. Frederickson Comer, and C. Saha. 2013. "Promotion and Tenure for Community-engaged Research: An Examination of Promotion and Tenure Support for Community-engaged Research at Three Universities Collaborating through a Clinical and Translational Science Award." *Clinical and Translational Science* 6(3): 204–8.

Matthews, A. K., A. Castillo, E. Anderson, M. Willis, W. Choure, K. Rak, and R. Ruiz. 2018. "Ready or not? Observations from a long-standing community engagement advisory board about investigator competencies for community-engaged research." *Journal of Clinical and Translational Science* 2(3): 129–34.

McCormick, J. 2021. "AI Ethics Teams Bulk Up in Size, Influence at Tech Firms; A Flaw in a Photo-cropping Algorithm at Twitter Prompted the Attention of its Ethics Team." Wall Street Journal (Online), May 27, 2021.

National Center for Advancing Translational Sciences. 2021a. "National Center for Advancing Translational Sciences (NCATS) FY 2021 Budget." https://ncats.nih.gov/files/FY21-justification.pdf

National Center for Advancing Translational Sciences. 2021b. "Translational Science Spectrum." https://ncats.nih.gov/translation/spectrum

National Institutes of Health. 2007. "Institutional clinical and translational science award (U54).".

National Institutes of Health. 2021. "Clinical and Translational Science Award (U54) PAR-18-464." https://grants.nih.gov/grants/guide/pa-files/PAR-18-940.html

Nease, D. E., D. Burton, S. L. Cutrona, L. Edmundson, A. H. Krist, M. Barton Laws, and M. Tamez. 2018. "Our Lab is the Community": Defining Essential Supporting Infrastructure in Engagement Research." *Journal of Clinical and Translational Science* 2(4): 228–33.

Pegasystems. 2017. "What Consumers Really Think About AI: A Global Study." Reportno. Report Number|, Date. Place Published|: Institution|.

Pelfrey, C. M., K. D. Cain, M. E. Lawless, E. Pike, and A. R. Sehgal. 2017. "A Consult Service to Support and Promote Community-based Research: Tracking and Evaluating a Community-based Research Consult Service." *Journal of Clinical and Translational Science* 1(1): 33–9.

Pinsoneault, L. T., E. R. Connors, E. A. Jacobs, et al. 2019. "Go Slow to Go Fast: Successful Engagement Strategies for Patient-centered, Multi-site Research, Involving Academic and Community-based Organizations." *Journal of General Internal Medicine* 34(1): 125–31.

Rabin, M. 1995. "Moral Preferences, Moral Constraints, and Self-Serving Biases." Unpublished (but well cited) Manuscript.

Rakova, B., J. Yang, H. Cramer, and R. Chowdhury. 2021. "Where Responsible AI Meets Reality: Practitioner Perspectives on Enablers for shifting Organizational Practices." *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW1): 1–23.

Raynor, K. 2019. "Participatory Action Research and Early Career Researchers: The Structural Barriers to Engagement and Why We Should Do It Anyway." *Planning Theory & Practice* 20(1): 130–6.

Richardson, B., J. Garcia-Gathright, S. F. Way, et al. 2021. "Towards Fairness in Practice: A Practitioner-Oriented Rubric for Evaluating Fair ML Toolkits." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13.

Rose, D. C., M. C. Evans, and R. M. Jarvis 2020. "Effective Engagement of Conservation Scientists with Decision-makers." In *Conservation Research, Policy and Practice*, edited by W. J. Sutherland, P. N. M. Brotherton, Z. G. Davies, N. Ockendon, N. Pettorelli, and J. A. Vickery, 162–82. Ecological reviews. Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/9781108638210.010

Safford, H. D., S. C. Sawyer, S. D. Kocher, J. K. Hires, and M. Cross. 2017. "Linking Knowledge to Action: The Role of Boundary Spanners in Translating Ecology." *Frontiers in Ecology and the Environment* 15(10): 560–68.

Saha, D., C. Schumann, D. Mcelfresh, et al. 2020. "Measuring non-expert Comprehension of Machine Learning Fairness Metrics." In *Proceedings of the International Conference on Machine Learning*. PMLR, 8377–87.

Saleiro, P., B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani. 2018. "Aequitas: A Bias and

Fairness Audit Toolkit." http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/

Schaich Borg, J. 2021. "Four Investment Areas for Ethical AI: Transdisciplinary Opportunities to Close the Publication-to-practice Gap." *Big Data & Society* 8(2): 20539517211040197.

Schiff, D., B. Rakova, A. Ayesh, A. Fanti, and M. Lennon. 2020. "Principles to Practices for Responsible AI: Closing the Gap." arXiv preprint *arXiv:2006.04707*.

Schiffer, Z. 2021. "Google Fires Second AI Ethics Researcher Following Internal Investigation." The Verge.

Schlesinger, W. H. 2010. "Translational Ecology." *Science* 329(5992): 609.

Strenge, B., and T. Schack. 2019. "AWOSE-A Process Model for Incorporating Ethical Analyses in Agile Systems Engineering." *Science and Engineering Ethics* 26: 1–20.

Teufel-Shone, N. I. 2011. "Community-based Participatory Research and the Academic System of Rewards." *AMA Journal of Ethics* 13(2): 118–23.

Towfighi, A., A. Z. Orechwa, T. J. Aragón, M. Atkins, A. F. Brown, J. Brown, O. Carrasquillo, et al. 2020. "Bridging the Gap between Research, Policy, and Practice: Lessons Learned from Academic–Public Partnerships in the CTSA Network." *Journal of Clinical and Translational Science* 4(3): 201–8.

Vakkuri, V., K.-K. Kemell, J. Kultanen, and P. Abrahamsson. 2020. "The Current State of Industrial Practice in Artificial Intelligence Ethics." *IEEE Software* 37(4): 50–7.

Verma, S., and J. Rubin 2018. "Fairness Definitions Explained." In *Proceedings of the 2018 IEEE/ACM International Workshop on Software Fairness (Fairware)*, 1–7 IEEE.

Vincent, W., B. Matthijs, and others. 2019. "scikit-fairness." https://github.com/koaning/scikit-fairness

Vogel, A. L., K. L. Hall, H. J. Falk-Krzesinski, et al. 2019. "Broadening our Understanding of Scientific Work for the Era of Team Science: Implications for Recognition and Rewards." In *Strategies for Team Science Success*, 495–507. Cham: Springer.

Washington, A. L., and R. Kuo. 2020. "Whose Side are Ethics Codes On? Power, Responsibility and the Social Good." In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 230–40 Barcelona, Spain: Association for Computing Machinery.

Wexler, J., M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. 2019. "The What-if tool: Interactive Probing of Machine Learning Models." *IEEE Transactions on Visualization and Computer Graphics* 26(1): 56–65.

AUTHOR BIOGRAPHY

Jana Schaich Borg is an Associate Research Professor of Neuroscience and Data Science at the Social Science Research Institute at Duke University. Her research leverages neuroscience, computational tools, and emerging technologies to understand and predict social interactions, and to optimize decisions that affect other people. She co-directs the Duke Moral AI Lab and the Duke Moral Attitudes and Decision-Making Lab, and was Faculty Director and Director of Duke University's Master in Interdisciplinary Data Science (MIDS) until 2020.

How to cite this article: Schaich Borg, J. 2022. "The AI field needs translational Ethical AI research." *AI Magazine* 43: 294–307. https://doi.org/10.1002/aaai.12062