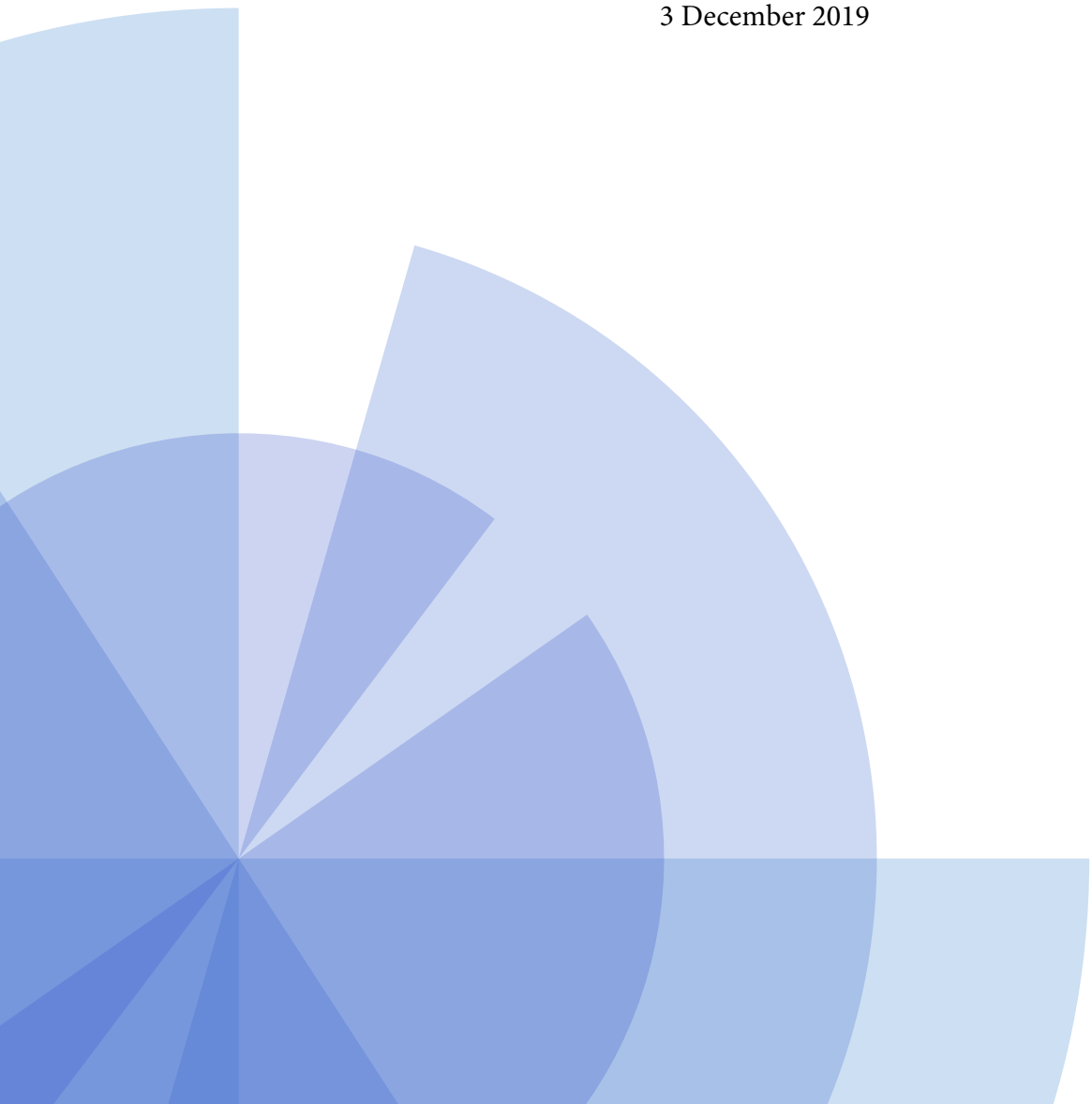
GRADIENT INSTITUTE

Practical Challenges For Ethical AI

Gradient Institute White Paper

3 December 2019



Gradient Institute White Paper

Practical Challenges for Ethical AI

3 December 2019

Gradient Institute Ltd.
Sydney and Canberra, Australia
https://gradientinstitute.org
info@gradientinstitute.org

Executive Summary

This White Paper examines four key challenges that must be addressed to make progress towards developing ethical artificial intelligence (AI) systems. These challenges arise from the way existing AI systems reason and make decisions. Unlike humans, AI systems only consider the objectives, data and constraints explicitly provided by their designers and operators. They possess no intrinsic moral awareness or social context with which to understand the consequences of their actions. To build ethical AI systems, any moral considerations must be explicitly represented in the objectives, data and constraints that govern how AI systems make decisions.

The first challenge in creating ethical AI is to define ethical objectives and constraints as precise, measurable quantities. This is necessary because AI systems reason mathematically, rather than through written or spoken language that is open to interpretation. Any mathematical representation will only be able to approximate the 'true' intention motivating the deployment of the system, and will inevitably fail to capture the complexity of human experience. It is therefore crucial that designers maximise the quality of such mathematical approximations by incorporating diverse viewpoints, and building robust mechanisms to detect and mitigate risks that arise when these approximations are unsatisfactory.

Given a well-defined set of ethical objectives, the next challenge is to create a system that will realise them. Doing so requires careful analysis of data bias, causal relationships and predictive uncertainty. Real-world data sets inevitably contain biases for which designers must account. It is also necessary to model the causal effect that design choices are likely to have on the objectives to better ensure the decisions produced by the AI system lead to the intended consequences. A quantitative treatment of uncertainty is also crucial to understand and manage risks associated with deploying the system.

The next challenge is to leverage human reasoning and judgement to provide effective oversight over AI-driven decisions. Effective oversight relies on nuanced approaches to transparency and interpretability; simplistic approaches such as transparent source code or intuitive explanations for individual decisions are unlikely to be the answer to create robust, reliable and ethical AI systems. More effective approaches will likely be context- and domain-specific, and require a deeper understanding of how to combine human and machine reasoning.

The fourth and final challenge this White Paper discusses is how to ensure regulation keeps up with advances in AI development. The urgent need to establish effective systems of accountability for AI-driven decision-making demands a proactive approach to regulation, however, the fact that the scientific understanding of AI ethics is still in its infancy means that policymakers should proceed with caution. A multidisciplinary approach is required to solve the challenges outlined in this White Paper, and it is unlikely that general-purpose, cross-domain regulation will prove suitable. Instead, we should consider bolstering sectoral regulatory bodies by identifying how current regulation falls short, making changes to cover existing gaps, and ensuring that all regulation is flexible enough to respond to rapid advances in AI technology.

The challenges of developing ethical AI systems are substantial, but so too are the opportunities to do good. With the right knowledge, we can engineer automated decision-making systems that deliberately minimise harm and can be configured to achieve a variety of ethical objectives. In doing so, we create the opportunity to discuss, re-examine, and perhaps even advance the values that we as a society wish to live by.

Contents

1 Introduction

Every day we give AI decision-making systems more power to affect the world. Not only are these systems directly steering the lives of billions of people, their decisions are influencing the long-term direction of our society in ways that are often unknown and difficult to predict. It is vital that we design these systems to improve the wellbeing of the people they affect. Designing ethical AI systems is an immense technological, philosophical and societal challenge but one we cannot afford to ignore.

AI SYSTEMS IN THE WORLD

AI systems are already ubiquitous in society, and make consequential decisions that impact people's lives. Companies use AI to make decisions about who gets insurance, who gets a loan, and who gets a job. Parole and sentencing risk scores, social media feeds, web search results, traffic routes, advertising, job recruitment, and online dating recommendations are all curated by AI systems.

As a society, we have delegated a lot of important decisions to AI because it is a powerful technology. AI can take advantage of the ever-increasing amounts of data we generate to make decisions faster and more consistently than a human, and in many cases, the predictions informing those decisions are more accurate than a human could achieve.

The impact of AI is not limited to simple improvements in the speed and accuracy of predictions. Just as with any technology, when and how we use AI systems are choices that can have major ethical implications. These choices can encourage a particular way of thinking, make some actions easier and others harder, and produce outcomes that benefit some at a cost to others. In other words, how we create and

apply AI steers our path towards a particular future. Some paths lead to a better world, but others do not.

Unfortunately, there is already mounting evidence of AI systems doing unintentional harm. Researchers have identified examples of unfair discrimination resulting from decisions made by AI systems that triage patients for special care programs [21], screen job applicants for interviews [13], recognise faces in images [6], and return web search results [20]. Harms such as these are unlikely to result from malicious intentions on the part of designers. They are unintended consequences arising from designs that do not adequately account for the differences between AI and human decision-making.

AI DECISION-MAKING

An AI system takes data, objectives and constraints as input, and produces as output predictions, decisions or actions determined exclusively to meet the objectives and respect the constraints. The consideration of supplied objectives and constraints exclusive of anything else is one of the crucial differences from human decision-making.

Consider an AI-driven loan approval application system. A bank has credit card transaction data and credit history data on its customers. The bank wants to know which new loan applications should be approved to maximise profit in the long term. An AI-driven solution to this problem might proceed along the following lines. First, designers build an AI model that uses historical transaction data to predict the probability a customer will repay a loan in the future. The model then analyses the transactional data of customers that did and did not repay their loans, to learn which patterns of transactions make a person more or less likely to repay. The bank applies this model to new applicants to estimate their repayment probability. From that probability, the system determines when to give out loans in order to maximise the bank's profit objective.

As this example illustrates, the AI designer encodes an objective — for example, to maximise profit — and the system uses data to learn which actions it can take to achieve that objective. In this way, AI design differs from most traditional soft-

ware development in which the programmer directly encodes the actions the computer should perform. By contrast, the AI designer needs to only program high level objectives that the system should fulfil, and these objectives may become detached from the detailed domain knowledge of the setting in which it operates. This in turn may limit the designer's insight into how the system will behave, particularly when faced with exceptional circumstances. This is problematic because the AI approach to decision-making ignores any considerations that are not specified as objectives or constraints - including ethical considerations. We cannot rely on AI systems to be reasonable, decent or to expect them to apply common sense without representing these concepts explicitly in computer language. If the bank loan system described above could maximise profit by denying loans to all people belonging to a certain racial minority, it would do so without any understanding that this approach was discriminatory.

To create AI systems that operate ethically, we must build ethics into system design. This White Paper aims to explore four of the biggest challenges of doing so. The first challenge is the difficulty of precisely defining the objectives and constraints of AI systems when the omission or misspecification of these can easily create unintentional harm.

The next challenge is building systems that act in alignment with these objectives and constraints, a task that requires careful consideration of the systems' input data as well as their models of causation and uncertainty.

Even with a well-specified objective and a rigorous design process, an AI system is likely to need constant adjustment, improvement and oversight. The third challenge is designing transparency and interpretability mechanisms to address this need.

The final challenge is to create regulations that help ensure AI operates ethically without over-constraining technological development. This task will likely require a careful, technically-informed approach that builds on existing regulations across a range of sectors, yet is flexible enough to keep up with the pace of ongoing technological change.

2 Specifying the Right Intent

Unlike people, AI systems have no default moral awareness or contextual understanding to help recognise the unintended consequences of their actions. They see only those objectives and constraints specified in their design. It is therefore imperative that designers provide AI systems with an explicit, carefully constructed, representation of the intent motivating the deployment of the system.

Unfortunately, ethical considerations rarely feature in existing AI systems. A typical implementation of an AI algorithm has the objective of maximising a single quantity such as the mean accuracy of its predictions. It neglects matters such as who will benefit or suffer harm from its operation, and the social impacts that this creates.

One reason many AI systems omit ethical considerations from their design, is that including them is very difficult. Doing so requires defining objectives that matter for a particular system, translating these into precise and measurable quantities, balancing them when they inevitably conflict, and monitoring and adapting them through design and deployment of the system. The following sections expand on these difficulties.

IDENTIFYING ETHICAL CONSIDERATIONS

In deciding what ethical considerations should shape AI systems, we can draw from the range of ethical frameworks that have been developed throughout human history. Though different ethical traditions contain areas of bitter dispute, some general aspirations are widely accepted, including [4]:

- generating good outcomes
- achieving purpose

- fulfilling obligations
- acting virtuously
- taking responsibility
- retaining trust and legitimacy

In practice, organisations use AI systems within established decision-making structures. Therefore, when designing AI systems, organisations should also look for guidance in existing processes and regulations, principles and governance structures.

Leveraging existing structures may not be enough to ensure AI systems operate ethically. Objectives set for AI systems may lead to a range of potential harms that can manifest in the short or long term. Because most AI objectives are based on data that typically reflects overall, aggregate performance, there is a particular risk that minorities and disadvantaged groups may be adversely affected. When designing objectives, it is crucial to consider the broad range of diverse viewpoints and experiences of people who may be affected by AI systems, in order to help mitigate these problems [12, 16]. Increasing the diversity of those in both senior decision-making and practitioner roles is a vital step towards developing the right ethical intent for an AI system.

Finally, even if an existing decision-making system appears to be operating ethically, it can be risky to 'copy' its objective directly into an AI system. Sometimes ethical considerations are not explicit in existing decision-making processes, but instead, rely on human judgement about when to bend or break the rules.

Consider the case of a call centre staffed by human operators. A caller rings to ask for an extension on a bill that they cannot currently afford to pay. The business rules state the operator should not offer an extension in this case, but the caller explains they are a pensioner who has had to pay a considerable sum of money for a funeral, following the sudden death of their partner. Despite extensions being 'against the rules', a human operator may be able to understand and respond to this exceptional circumstance and provide the customer with an extension. If this call centre was automated with a simple AI system, the system might have concluded that the business

rules should be followed without question, thereby ending up with an intent absent of the empathetic judgement found in the human system.

QUANTIFYING AND MEASURING ETHICAL CONSIDERATIONS

When designing an AI system, it is not enough to enumerate ethical considerations in written language. Designers must encode these considerations mathematically as objectives or constraints. The translation of ethical concepts to mathematics is a challenging problem that requires making value judgements. Mathematics is more specific and precise than written language, and a single ethical concept such as 'fairness' can have many mathematical definitions with different ethical implications. Researchers have already identified dozens of them [3, 26]. Each of these definitions differ in exactly how they will benefit some people and disadvantage others. When developing an AI system, a designer has to make a conscious choice about which mathematical definition of fairness to use, knowing that a different choice could lead to a substantially different outcome.

Ethical concepts in written language are likely ambiguous enough that mistranslation between intent and outcome is a significant risk. Consider an AI system designed to screen résumés to select job applicants for interview. The system's designers determine that one of its objectives is to not unfairly discriminate against female applicants. The designers take this concept and translate it into code: they remove the gender column from the input data, based on the belief that the system cannot discriminate if it has no gender information with which to do so. Unfortunately, this apparently reasonable approach likely represents a mistranslation between the concept of discrimination and the coding of this concept in the system's design. Denying the system access to the gender status of applicants is often not only ineffective in eliminating gender-based discrimination (since the algorithm can implicitly recognise gender from other features that correlate with gender), but may even exacerbate it. For instance, if total work experience is a success factor in the application, women who stayed out of work for extended periods to care for a young family can be mistaken for males otherwise similar but who stayed out of work due to unemployment [22, 23]. If the system has no access to the gender column in the data, it

can't even detect there is a potential issue, let alone use the gender information in any proactive effort to ameliorate the situation.

Proxies

Not all quantities are practically measurable, even if they can be precisely defined. The best designers can hope for if they encounter an essential, but unmeasurable quantity, is to develop a proxy, or a measurable stand-in. Designers might, for example, use an arrest rate (which can be measured) as a proxy for a crime rate (which also includes crimes that went unnoticed).

Failing to develop accurate proxies creates risks of AI systems causing unintended harm. To AI systems, the proxy is the consideration: the systems' actions will be optimised to achieve proxy objectives, and may be indifferent to, or even detrimental to, the actual objectives.

A recent US study that assessed algorithmic health risk estimation highlights this issue well [21]. In the study, the algorithm's goal was to direct patients to a special care program by estimating their health risk. The algorithm quantified this health risk by using a proxy: the estimated future cost of medical treatment for the patient. The study showed that the quality of this proxy varied with race: black patients had a higher real health risk than white patients for the same expected cost of future treatment.

In other words, the proxy used by the algorithm underestimated real health risk for black patients compared to white patients. The difference was likely driven by sick black patients incurring lower costs due to poorer access to care. The use of this proxy then denied black patients special care at a higher rate than white patients for the same level of real health risk. Designers expected the cost of service to be an accurate proxy for health, but did not anticipate that the approximation it applied systematically disadvantaged black patients.

Using a better quality proxy that incorporated real health indicators would have likely lessened or removed this bias. However, acquiring additional data may have impacted patients' privacy and reduced their willingness to engage with the system. The people responsible for the deployment of the system would have to pay careful attention to balancing these conflicting considerations.

BALANCING OBJECTIVES

It is rarely possible to satisfy every objective that designers encode into an AI system. Instead, such objectives are often in conflict: any decision made by an AI system results in some benefits and some costs. The people responsible for the system's deployment are limited to determining how the system should distribute these benefits and costs within the limitations of their capability and knowledge [15]. Trade-offs are inevitable in any decision-making process, and balancing competing ethical objectives is part of what makes designing ethical systems so difficult.

Consider the fraught question of whether to remove children considered to be at risk of harm from their parents. Should child protection authorities resist removing children from an Indigenous community with evidence of higher rates of serious abuse in order to equalise removal rates with a similar, non-Indigenous community nearby? This question requires addressing unavoidable trade-offs between the harm done to the children, family and community by removing children, and the risk of harm to the children if they remain in their current situation. It also requires deep engagement with many issues beyond the child protection system [10], such as the history of the Stolen Generations [1], intergenerational trauma, and discrimination [5].

Even simpler problems with considerably fewer nuanced sociocultural layers will contain unavoidable trade-offs. Many quantitative measures of fairness demand a trade-off between fairness and accuracy, meaning that AI system designers must increase overall error rates to decrease error rate disparity across different groups. Similarly, many different fairness measures are impossible to satisfy simultaneously [9].

The need to balance competing ethical objectives is not new to AI systems. What is new, is the need to specify the balance of these objectives with mathematical precision. This represents a new way of thinking for human decision-makers, which will likely require developments in methods of eliciting preferences, and of visualising and exploring potential outcomes.

CONCLUDING REMARKS

Accurately and precisely encoding the ethical objective of an AI system is difficult, and at times impossible. Any AI system is therefore likely to be built with an approximation of the real intention motivating the deployment of the system. In some cases, it may be possible to build oversight mechanisms to detect and correct areas of misspecification (explored in section four), but building such oversight systems is itself a critical challenge. In other cases, it may not be possible to encode a reasonable ethical intent into an AI system before it is used to make consequential decisions: this is a good indication that AI may not be a suitable tool for solving the problem at hand.

3 Designing a System to Deliver on the Right Intent

For an AI system to operate ethically, the consequences of its actions must align with the ethical intent motivating the deployment of the system. Understanding how a design choice influences a system's ethical impact requires careful analysis. The data provided to the system, the causal effects of actions and design choices on outcomes, and the treatment of uncertainty all contribute to the system's ability to realise the intent with which it is designed.

ACCOUNTING FOR BIASED DATA

An AI system's model of the world is informed by the data its operators provide. Problems with this data can result in the system making unethical decisions, even if the designers have specified an appropriate ethical intent. Data may encode human bias or discrimination directly, or may do so indirectly by under- or over-representing disadvantaged parts of society. If the AI system has not been designed to account for these biases, it can perpetuate disadvantage or discrimination.

Consider an algorithm trained to filter the résumés of prospective childcare workers based on previous human-made hiring decisions. If the previous hiring managers preferentially selected women over men with the same qualifications, then training an algorithm on historical hiring decisions would perpetuate the hiring managers' bias. This kind of discrimination in data can be impossible to detect without access to additional data sources. Only explicit estimation and correction of the discriminatory effect will prevent an AI system from perpetuating it.

Another problem is that the availability of data that is used to train AI systems

often correlates with disadvantage: data sets for crime and welfare over-represent disadvantaged populations, while the opposite is true of data sets like purchase history, credit history and education. Unless explicitly instructed otherwise, AI systems will make decisions assuming that good outcomes are typically attained by groups for whom there is data on good outcomes, and bad outcomes for groups for whom there is data on bad outcomes. The result is, again, a mirroring and compounding of existing inequalities.

Take the development of an educational app by designers living in a big city. The schools from which they could readily obtain training data are likely to be proximate, and have the expertise and resources to collaborate. Such schools would look quite different from those in remote, disadvantaged communities. An app trained using data from city schools is therefore likely to perform better for children similar to those whose attributes were captured in the training data. If designers fail to notice and address that the design of this app is optimised to perform well among children in privileged city schools, they will continue to perpetuate disadvantage for children in remote or disadvantaged communities.

CAUSAL RELATIONSHIPS

Ethical AI systems should prevent avoidable harms. An ethical AI system must therefore be able to model the consequences of its actions. Such causal reasoning is a hard, context-sensitive problem requiring multidisciplinary domain expertise, and is unlikely to be automated in the foreseeable future. AI systems typically avoid this kind of reasoning by making strong assumptions about causal relationships and relying on statistical correlations to make predictions. When these causal assumptions are wrong, there is a real risk that the systems' actions may cause harm even if it has been designed with an ethical intent.

Examples of harm caused by mistaken causal reasoning abound: rabbits were introduced into Australia with the purpose of providing food and rabbit hunting, and cane toads were brought into the country for the purposes of controlling the cane beetle. Despite good intentions, the consequences of these decisions led to the catastrophic destruction of native fauna [11, 17]. Similarly, interventions to make processes

fairer for disadvantaged minorities may harm the groups they are intended to benefit. For example, a US ban on employers asking about criminal history in job applications may have increased discrimination against young black and Hispanic males — in the absence of evidence to the contrary, employers assumed they might have a criminal record [2].

AI systems are not immune to such mistakes. Take, for example, a system built to help with hospital admissions of pneumonia patients [8]. The system was designed to predict a pneumonia patient's risk of death when they arrived at the hospital. This information was intended to be used in a new admission protocol based on that risk of death, which would replace an existing system that relied on a doctor's judgement.

On inspection of the AI model, the designers noticed that the algorithm viewed patients with asthma as having a low risk of death. This relationship was a real correlation in the original admission protocol: doctors knew that patients with asthma were at high risk, so they were often admitted straight to intensive care and thus rarely died. However, the model did not account for the causal effect of changing the treatment protocol by removing these doctors' judgements. Placing this algorithm in charge of informing a new treatment protocol, in which asthma patients are considered 'low risk', would have been a terrible mistake.

The issue in the above example can be identified relatively easily by thoughtful and informed introspection of the correlations produced by the system. The general problem of relating the thousands of small design decisions that go into an AI system's design and implementation to the many possible ethical consequences that result from these decisions is significantly harder. This difficulty points to the importance of modelling uncertainty when estimating the impacts of AI systems.

UNCERTAINTY

It is impossible to predict with perfect accuracy the outcomes of future events. Some consequences will be likely, some will be unlikely, and quantifying this uncertainty is vital to understanding the risks associated with different choices people make in the process of conceiving and deploying an AI system. Like with causality, modelling uncertainty is a hard, context-sensitive problem that still requires extensive human

thinking and investigation. Failing to accurately quantify uncertainty in the outcomes of a system leads to that system making bad decisions: if the system wrongly believes a possible negative outcome is unlikely, it may fail to avoid it.

Amongst the failures apparent from the Australian Government's recent online compliance intervention (OCI) debt recovery system episode — colloquially known as 'Robodebt' — is a failure to address uncertainty in an automated system [7]. The system designers arguably failed to take proper care in acknowledging and quantifying the margins of error associated with an averaging procedure used to estimate overpayment recovery amounts for welfare recipients. While welfare payments were based on actual fortnightly earnings, the system used the fortnightly average of annual earnings to estimate any overpayments in cases where fortnightly earnings data was not provided by recipients.

As an estimate, this fortnightly average of annual data had an associated uncertainty that appears not to have been estimated. Had it been, this uncertainty would have been substantial in many cases, and this may have prompted the system operators to be more careful when issuing overpayment notices. Instead, the averaging procedure resulted in large-scale issuing of overpayment notices that greatly exceeded the correct amounts owed, causing significant and avoidable distress for many welfare beneficiaries.

CONCLUDING REMARKS

Understanding how the design and operation of an AI system delivers on its stated intent is a complicated scientific and technical challenge. It requires building an understanding of the cause and effect relationships between the design choices, the resulting AI systems' decisions, and the ethical consequences of these decisions. Understanding the consequences of AI systems' decisions requires understanding the bias and limitations of the available data, and quantifying the uncertainty of systems' predictions. Even AI systems built with good intentions will continue to cause unnecessary harm as long as these problems are not addressed.

4 Leveraging Human Judgement and Oversight

As section two described, AI systems cannot independently apply contextual information to problems as they arise, or see the broader consequences of their operations beyond the data and objectives their designers supply. Human oversight over AI systems is therefore critical to ensuring they operate ethically. By making the operation of AI systems more transparent and interpretable, we provide human overseers the ability to anticipate, detect, and correct problems with systems.

A transparent AI system is one whose design or operation is available for inspection by a broader set of people than just those operating it. Increasing the transparency of AI for citizens, customers, regulators or experts may increase the chance that a mistake in the design of the system will be noticed and corrected, and may also help to ensure that these systems earn and maintain the trust of the people they affect.

However, AI systems are typically complex mathematical and computational artefacts that have reasoning processes that differ significantly from those used by humans. As such, merely exposing the inner workings of an AI system may do nothing to help a human understand their operation. Such understanding requires explainability: that is, representing the machine's operations in a way that allows humans to reason about it.

The tools of transparency and explainability both help combine the predictive power and consistency of AI technology with the contextual judgement and causal understanding of people to make better decisions. The following sections explore challenges associated with each.

TRANSPARENCY

The justification of transparency in AI systems is similar to that for other powerful systems like governments and corporations: it reduces the risk that they will be motivated, designed or operated in a socially unacceptable way.

Quantities like the outcomes of a system, how these outcomes are distributed amongst different groups, the data and code the system relies on, and the specified set of objectives and constraints, may all be beneficial to make transparent to the right people under the right circumstances. These people may include users, auditors, regulators, oversight bodies, academic experts, or society at large.

The unanswered research problem is to understand better what information about which system needs to be made transparent, and to whom, in order to ensure the system operates ethically.

Some cases may be intuitive. For instance, it seems reasonable that AI systems interacting with humans in ways that deliberately mimic human behaviour should disclose that fact. The increasing deployment of chatbots and AI voice assistants makes this kind of transparency a live issue.

There are also circumstances in which some kinds of transparency might not be practical or even desirable. Transparency of source code and data would allow external parties to reproduce a system in its entirety, revealing its behaviour in real and hypothetical situations, its assumptions and its practical limitations. In some cases, this might be an appropriate level of transparency, but too much transparency may also damage the privacy of entities in the data, or create disincentives for organisations to develop new algorithms.

In certain sensitive situations transparency can completely undermine the utility of a system's decisions. The implementation details for algorithms targeting investigations like customs inspections or tax audits, if made transparent, could provide those wishing to avoid the attention of the algorithm an understanding of exactly how to do so.

In sum, transparency is not a panacea but rather a powerful tool to improve the design of AI systems if used in the right way and with the right purpose.

INTERPRETABILITY

Transparency may not be sufficient for useful human oversight. The mechanisms by which an AI system makes decisions may not be possible for a person to fully comprehend. Making AI systems more interpretable allows people to understand their reasoning processes, explain how mistakes occurred, or inform users how to adapt their behaviour to obtain different decisions from systems in the future.

In the previous example of the pneumonia risk prediction algorithm (section three), the designers noticed the system correlating asthma with a low risk of death because they could scrutinise its reasoning. This correlation was not an error, but seeing it helped the designers catch a potential error in their thinking: mistaking the predictive model they had built for a causal one. By providing a summary of an AI system's workings or conclusions, interpretability can allow people to leverage their relational understanding of the world to detect problems.

Interpretability also helps us understand or produce justifications for the decisions AI systems make. An individual impacted by an AI-made decision may have a right to understand the basis for it. Justifying a decision is especially important if the decision could be harmful, or denies someone a benefit they could have otherwise received.

For example, an individual denied a loan might want an explanation so that they might contest the decision, or change their behaviour in order to obtain a loan in the future. One explanation may be that the applicant is statistically similar to people that did not repay loans in the past, based on gigabytes of financial transactional information about the applicant and other customers. This explanation (perhaps along with the associated data) is not interpretable or useful to the applicant. A more interpretable and useful justification might look more like the following:

> *The bank denied your loan because you have missed credit card repayments, and because your account balances have been consistently going down. If you do not miss any payments in the next 12 months, and if your balances do not decrease significantly in that period, your chances of having a loan application approved will be significantly higher.*

Unfortunately, ensuring such justifications are simple enough to understand but

also accurately describe the underlying decision processes is a significant technical challenge [24].

Even with further research, there is likely to be a fundamental design trade-off between interpretability and predictive power. A critical part of operating an ethical AI system is deciding on that trade-off: determining what attainable compromise between predictive power and effective human oversight results in the best ethical outcomes.

One proposed solution to this problem is to suggest that when a decision must be completely interpretable, it should be made by a human. This approach, however, has its own challenges. Although humans are excellent at providing a narrative to justify how they reached a decision, there is ample evidence from social psychology that such explanations often do not accurately reflect the actual cognitive processes involved [14, 19]. Humans possess the most complex known neural networks, and their decisions are influenced by unique experiences, remembered imperfectly. These memories cannot be audited, corrected, or reliably presented.

By contrast, machines can keep precise records of all the data they process, as well as records of every instruction they execute. If we judged human decision-makers by the same standards we do algorithms, they would be the least interpretable models we have. Even the most opaque AI systems allow us to answer which output will be produced for a given input. On the other hand, we never know for sure what a human would have done had the information available to them been different.

Finally, it is not always worthwhile trading any predictive power for interpretability. There are problems such as image classification where the gains in accuracy from using complex, difficult-to-interpret models like deep neural networks dramatically outweigh the losses from poor interpretability. In cases such as using AI systems to detect cancer in images of tissue samples, highly accurate but hard-to-interpret models appear to be the most ethical choice. Like with transparency, interpretability serves a purpose and isn't an end in itself.

CONCLUDING REMARKS

AI systems can outperform experts for some prediction tasks, but can also make mistakes obvious to a child. The right approach to building an AI system is to realise that it will always have both human and machine components. A well-designed system will take advantage of the skills of both, combining AI systems' ability to ingest vast amounts of data, make accurate predictions and quantify uncertainty, with humanity's superior contextual understanding, moral reasoning, domain knowledge, and adaptability.

5 Regulation and AI Systems

The rising role AI is playing in almost all aspects of our lives and its potential for transformative impact suggests we must be proactive in developing regulation to mitigate and prevent harm. There are however a number of issues that make regulating in the face of AI very challenging: mechanisms for attributing responsibility and accountability for AI-made decisions need to be established, the field is technically complex and developing very rapidly, and the potential issues and concerns around the use of AI differ substantially across sectors [18].

ACCOUNTABILITY

Society has built complex mechanisms to assign accountability in the government, legal and corporate worlds. However, as these mechanisms were not designed with AI-driven automated decision making in mind, there is work to be done in evaluating how well they apply and if there are any gaps. Because humans and AI systems process information in such different ways, there are particular complexities that become apparent in settings where a flawed decision was made jointly by a human and an AI system. For example, the designer of an AI system may recommend human oversight, but the person tasked with providing oversight may lack the knowledge required to properly evaluate the system's conclusions.

Having someone (a person, corporation or public body) ultimately responsible and accountable for any product or service (including automated decision making) is crucial. It provides two key benefits. First it affords people who have suffered harm the potential to seek redress. Secondly, it provides an incentive for those offering products or services to avoid causing harm in the first place so as to avoid being sued, thrown out of office, or having their company lose value. It may be possible to

hold organisations who deploy AI systems accountable for the actions of those systems under existing legal and political frameworks. However, the complexity of AI systems, their dependence on the data supplied to them and the difficulty of anticipating how they will operate in every contingency raise unique challenges. Although AI systems are complex and senior decision-makers within organisations may lack the appropriate background to understand their inner workings, the solution cannot be to move accountability down the chain until it finds someone who does have that background. The person who rivets the skin to the aircraft should not be the one deciding what the safety margin should be.

THE GENERAL-PURPOSE NATURE OF AI

AI is a general-purpose technology. Everything from an excel spreadsheet, to a smart toaster, to a data-driven fraud-detection system that automatically triggers audits of taxpayers could be considered to utilise AI. The type and scale of harm that might arise from the use of AI in different applications can vary significantly. This suggests that implementing regulations such as labelling requirements, special taxation or regulatory approval processes for 'AI systems' broadly construed is unlikely to be helpful.

Take the example of software, another general-purpose technology. Software is not itself the subject of regulation, but rather specific industries or classes of products or services that make use of it. Similarly, it's not AI that needs regulation, but rather the application context in which it is used. For instance, it is reasonable to expect that regulatory frameworks should evolve to govern the deployment of autonomous vehicles. Likewise, there is a clear need for anti-discrimination legislation to be revisited in light of automated decision systems. This argument suggests that a pragmatic approach may be to focus on the application side of AI, such as providing appropriate support to existing sector-specific regulators [25].

EVIDENCE-BASED AND TECHNICALLY-INFORMED REGULATION

Given the rate at which the field of AI is developing, the fact that applications for AI systems continue to emerge and that the ethical impacts of AI are so poorly understood, there is a risk of developing regulation which is naive or poorly designed, and results in unintended consequences.

Regulation should be evidence-based. For instance, consider regulation promoting consumer rights. Regulatory strategies such as granting rights to opt-out of automated decision making, or to have one's data deleted, may be beneficial to some people in some circumstances. However, many individuals may not have the capacity, resources or knowledge to exercise such rights. This is a crucial piece of evidence. It suggests that such regulations if applied in isolation have the potential to exacerbate inequality by selectively benefiting those who are already in privileged positions. Whenever considering a new regulatory intervention, an important question to ask is: who will benefit the least with this regulation, and what can be done to compensate for that?

Regulation must not only be evidence-based, but also technically informed. Consider the case of law. Because legislation is written in ambiguous human language and AI systems operate with precise computer code, there is an increased propensity for the emergence of regulatory loopholes that can be exploited by organisations to subvert the authority of governing bodies. It is therefore critical that technical subject matter experts are involved in the development of policy and regulation. Their expertise is required to translate the requirements of legislation into lines of code in an AI system.

CONCLUDING REMARKS

Regulation in the face of AI is challenging due to the complex and rapidly changing nature of the field and the broad scope of potential applications. However, given the potential future impacts of AI technology, both good and bad, it is critical that we take a proactive approach to developing appropriate regulations. Given the broad scope of potential applications, it is unlikely that blanket regulation of AI technolo-

gies will be effective, and it is likely that sector-specific legislation will be required. As a first step, we should support existing regulators and oversight agencies to address the challenges introduced by AI, and identify specific areas of legislation that need to be adapted when the powers of these existing bodies are insufficient.

6 Conclusion

> *Everything is vague to a degree you do not realise till you have tried to make it precise, and everything precise is so remote from everything that we normally think, that you cannot for a moment suppose that is what we really mean when we say what we think."*
>
> (Bertrand Russell in "The Philosophy of Logical Atomism" (1918-19))

We can only build ethical AI if we manage to address the challenges laid out in this White Paper. It will be difficult: each challenge contains unsolved research problems, and questions that may be impossible to answer fully. It will require broad engagement and collaboration: between technical, legal, domain and ethical experts, decision-makers, the public, and critically, those most at risk of harms caused by AI systems.

To fail to address these challenges is to continue in the direction of having powerful technology make inscrutable decisions on our behalf, with no consideration of broader ethical consequences. The harms already done by existing AI systems provide a glimpse of this potential future. However, it is not only mitigation of risk that should drive the development of ethical AI. The challenges set out in this White Paper are also opportunities for us to improve the standard of ethical decision-making.

One of the reasons we deploy AI systems in the first place is to improve upon human decision-making. People can get tired, take shortcuts, rely on first impressions, mischaracterise risk, and apply a myriad of other cognitive biases to their decisions. In carefully chosen situations, AI-driven decision-making can avoid these biases and make better decisions.

AI systems force us to be explicit and precise about our intent, our model of the world, and how we want to balance objectives. In doing so, they stimulate discus-

sion and debate about what the right intents and trade-offs are. Re-examining these questions in consequential situations may help us improve upon the status quo.

More broadly, in imposing the challenge of making ethical concepts mathematically precise, AI technologies invite us to refine our understanding of ethics. The very act of making a concept precise can expose ambiguities and questions we had not previously considered, and can force us to address them. Developing ethical AI systems can help improve the standard of decision making beyond forcing us to formalise ethical concerns. The fact that AI is implemented in software has the potential to improve accountability and transparency. It is harder for system owners to provide false explanations or justifications, if the system is available for scrutiny. This also hinders the human propensity to claim a different motivation from the one actually used to make the decision.

These examples point to the possibility that AI systems can be a force for good in the world, but this will only happen through a deliberate commitment to realising ethical AI systems. Our challenge is to build AI that articulates the best humanity has to offer, and in doing so, create a better world for everyone.

Authors and Acknowledgements

This White Paper was written by:

Tiberio Caetano	Finn Lattimore
Lachlan McCalman	Simon O'Callaghan
Linda Przhedetsky	Alistair Reid
Bill Simpson-Young	Dan Steinberg

The authors acknowledge feedback from the following people on earlier drafts of this White Paper:

Will Bateman	Matthew Beard
Lyria Bennett Moses	Terry Caelli
Gregorio Caetano	Jack Dan
Chris Dolman	Kate Fullagar
Thomas Ilbery	Duncan Ivison
Elizabeth Kelly	Seth Lazar
Ben Lever	Simon Longstaff
Iain McCalman	Julianne Schultz
Greg Taylor	Elizabeth Tydd
Kimberlee Weatherall	Bob Williamson

Bibliography

[1] Australian Institute of Aboriginal and Torres Strait Islander Studies. *Stolen Generations*. 2018. URL: https://aiatsis.gov.au/research/finding-your-family/before-you-start/stolen-generations.

[2] A. Agan and S. Starr. "Ban the box, criminal records, and statistical discrimination: A field experiment". In: *University of Michigan Law & Economics Research Paper* 16-012 (2016). URL: http://dx.doi.org/10.2139/ssrn.2795795.

[3] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. 2019. URL: https://fairmlbook.org/.

[4] M. Beard and S. Longstaff. *Ethical by Design: Principles for Good Technology*. The Ethics Centre. 2018. URL: https://ethics.org.au/ethical-by-design/.

[5] *Bringing Them Home. Report of the National Inquiry into the Separation of Aboriginal and Torres Strait Islander Children from Their Families*. Human Rights and Equal Opportunity Commission, 1997.

[6] J. Buolamwini and T. Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Machine Learning Research*. Vol. 81. 2018, pp. 1–15.

[7] T. Carney. "The New Digital Future for Welfare: Debts without Legal Proof or Moral Authority?" In: *2018* (), pp. 1–16.

[8] R. Caruana et al. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission". In: *Proc. 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015, pp. 1721–1730.

[9] A. Chouldechova. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments". In: *"Big Data"* 5.2 (2017).

[10] R. Courtland. "Bias detectives: the researchers striving to make algorithms fair". In: *Nature* 558 (2018), pp. 357–360.

[11] J. Covacevich and M. Archer. "The distribution of the cane toad, Bufo marinus. Australia and its effects on indigenous vertebrates". In: *Memoirs of the Queensland Museum* 17.2 (1975), pp. 305–310.

[12] K. Crawford. "Artificial Intelligence's White Guy Problem". In: *New York Times* (2016). URL: https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html.

[13] J. Dastin. *Amazon scraps secret AI recruiting tool that showed bias against women*. 2018. URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[14] H. Kelley. "The process of causal attribution". In: *American Psychologist* 28.2 (1973), pp. 107–128.

[15] J. Kleinberg, S. Mullainathan, and M. Raghavan. *Inherent Trade-Offs in the Fair Determination of Risk Scores*. 2016. URL: https://arxiv.org/abs/1609.05807.

[16] L. McCalman. *Whose Ethics?* Gradient Institute. 2019. URL: https://gradientinstitute.org/blog/4/.

[17] R. McLeod and A. Norris. *Counting the cost: impact of invasive animals in Australia'*. Canberra: Cooperative Research Centre for Pest Animal Control, 2004.

[18] L. Bennett Moses. "Regulating in the Face of Sociotechnical Change". In: *Oxford Handbook of Law, Regulation and Technology*. Ed. by R. Brownsword, E. Scotford, and K. Yeung. Oxford University Press, 2017.

[19] R. Nisbett and T. Wilson. "Telling more than we can know: Verbal reports on mental processes". In: *Psychological Review,* 84 (1977), pp. 231–259.

[20] S. Noble. *Algorithms of oppression: How search engines reinforce racism*. New York University Press, 2018.

[21] Z. Obermeyer et al. "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464 (2019), pp. 447–453.

[22] D. Pope and J. Sydnor. "Implementing Anti-Discrimination Policies in Statistical Profiling Models". In: *American Economic Journal: Economic Policy* 3 (2011), pp. 206–31.

[23] A. Reid and S. O'Callaghan. *Ignorance isn't bliss*. Gradient Institute. 2018. URL: `https://gradientinstitute.org/blog/2/`.

[24] M. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?: Explaining the predictions of any classifier". In: *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1135–1144.

[25] M. Whittaker et al. *AI Now Report 2018*. AI Now Institute, 2018. URL: `https://ainowinstitute.org/AI_Now_2018_Report.pdf`.

[26] R. Williamson and A. Menon. "Fairness risk measures". In: *Proc. 36th International Conference on Machine Learning*. 2019, pp. 6786–6797.