We begin by reviewing how principles have evolved in AI ethics over the last two years. Drawing on comparisons with bioethics - a field with a robust and well-developed tradition in using principles to govern medical practice - we discuss some of the limitations of principles. We make the case that all areas of AI ethics would benefit from a more rigorous exploration of the tensions that arise when we try to apply principles to concrete cases. We outline some key tensions that already arise from the use of AI in society, and discuss what work might be needed to resolve them. To our knowledge, this is the first paper to explicitly examine the role and limits of principles in AI ethics, and the importance of focusing more on tensions as a next step.

2 THE EMERGENCE OF PRINCIPLES IN AI ETHICS

Though the field is in its infancy, there is widespread agreement on some of the core issues (such as bias) and values (such as fairness) that AI ethics should focus on. Over the last two years, these have begun to be codified in sets of 'principles' or 'tenets'. The Asilomar AI principles, developed in 2017 in conjunction with the Asilomar conference for Beneficial AI, outline guidelines on how research should be conducted, ethics and values that use of AI must respect, and important considerations for thinking about long-term issues [12]. The principles were signed by several thousand AI researchers and others, including many academic ethicists and social scientists. Around the same time, the US Association for Computing Machinery (ACM) issued a statement and set of seven principles for Algorithmic Transparency and Accountability, addressing a narrower but closely related set of issues [1].

Over the course of 2017, several other initiatives and organisations published additional sets of principles: including the Japanese Society for Artificial Intelligence's Ethical Guidelines in February 2017 [16]; a set of draft principles from the Montreal Declaration on Responsible AI in November [25]; and the IEEE's General Principles of Ethical Autonomous and Intelligent Systems in December [14]. This proliferation of principles has continued into 2018: with the Partnership on AI publishing a set of 'tenets' which its members agree to uphold [18]; the UK House of Lords suggesting five principles for a cross-sector AI code which could be adopted internationally [23], and Google publishing their 'AI ethics principles' in June [19].

These different sets of principles have considerable overlap. There is widespread agreement that AI-based technologies should be used for the common good, should not be used to harm people or undermine their rights, and should respect widely-held values such as fairness, privacy, and autonomy. [9] suggest that many of the different existing sets can be synthesised into five key principles: the four that are already used in bioethics - autonomy, beneficence, non-maleficence, and justice [4] - plus the additional principle of explicability, which captures the challenges of intelligibility and accountability unique to AI systems. While this convergence is encouraging, it is unclear at this point whether this reflects a deep consensus about what is important, arrived at independently by numerous different actors, or merely a shallow consensus due to the fact that different groups have read similar papers and built on the work of one another.

Principles can be a valuable part of applied ethics; agreeing on high-level principles is therefore an important step for ensuring that AI is developed and used for the benefit of society. Principles help condense complex ethical issues into a few central elements which can be clearly understood and agreed upon by people from diverse fields and sectors. They encourage widespread commitment to a shared set of values, and can give them a more prominent role in institutional decision-making processes. Principles can form a basis for more formal commitments in professional ethics, internationally agreed standards, and regulation. They can also help address public concerns, by clarifying the ethical commitments of researchers and industry.

However, while principles are important, they are not in themselves enough to ensure society can reap the benefits and mitigate the risks of new technologies. In order to be useful in practice, principles need to be able to guide action - to help people navigate the competing demands and considerations of concrete situations. As we articulate in the next section, there are several obstacles to this.

3 THE LIMITS OF PRINCIPLES

The four principles of beneficence, non-maleficence, autonomy, and justice have played a prominent role in bioethics [4], a field with decades of experience in managing the challenges posed by new technologies. These principles aim to articulate general values on which everyone can agree, and to function as practical guidelines. But they have spurred substantial debate: some argue that we should put no weight on principles and focus entirely on the elements of specific cases [10], while others have advocated a more moderate view, whereby principles should be considered in close conjunction with analysis of 'paradigm' cases [17].

However, even the strongest advocates of principlism in bioethics acknowledge that principles alone are not enough. [4] suggest that principles should be taken as guidelines, which need to be made specific for use in policy and clinical decision-making. They elaborate that in order to be action-guiding, principles need to be accompanied by an account of how they apply in specific situations, and how to balance them when they conflict. In this section, we review some of the main limitations of principles that have been highlighted in the bioethics literature and illustrate why these also apply to the principles proposed for AI ethics.

3.1 Different Groups May Interpret Principles Differently

The central terms used in principles are often ambiguous, masking conceptual complexity and differences in interpretation across populations. As [8] point out, the principle of 'justice' in bioethics does not say anything about what is just or unjust, leaving this to the agent to decide for themselves. Clouser and Gert argue that principles often mask important moral disagreements rather than presenting a well-developed unified theory as they propose to, and that it would be better if these disagreements were articulated and understood more explicitly.

In particular, lists of broadly agreed-upon principles cannot recognise that important and legitimate differences in values exist across people and populations. While everyone might agree

in principle that 'fairness' is important, there exist deep political disagreements about what exactly constitutes fairness [5]. Groups may also vary in how much weight they put on one value relative to others in situations of conflict: more individualist cultures may put more weight on personal privacy than more collectivist cultures, for example. An important step in making principles more practical is to formalize them into standards and regulation [27]. But this process is not a straightforward unpacking of the relevant principles, as different principles will come into conflict when applied to concrete cases. In the next section, we make the case that in order for principles to inform more practical aspects of AI ethics, including professional ethics, standards, and regulation, we need to begin by exploring in detail the different kinds of tensions that arise when principles are applied.

3.2 Principles Are Highly General

Relatedly, principles are by their nature highly general: their value is that they indicate important moral themes that apply across a wide range of scenarios. This means that they can be useful as a kind of checklist: as a set of important considerations that need to be taken into account in specific scenarios. However, the generality of most principles also limits their ability to guide practical action [3]). Many of the principles proposed in AI ethics are too broad to be action-guiding. For example, ensuring that AI is used for 'social good' or 'the benefit of humanity' is a common thread among all sets of principles. These are phrases on which a great majority can agree exactly because they carry with them few if any real commitments. A very wide range of differing ideological, political and philosophical standpoints could claim to be for the good, or for the benefit of humanity. Only principles that are narrower and more specific are likely to be useful in practice. Recent industry commitments to not develop technology for autonomous weapons are an example of a principle that is specific, action-guiding and can be used to hold people to account. But at the same time, exactly that specificity means its relevance is limited to one sector, and it has many dissenters.

3.3 Principles Come into Conflict in Practice

The gap between principles and practical judgement grows larger still when we consider that principles will inevitably conflict with each other. For example, the UK House of Lords AI Committee report states that, "it is not acceptable to deploy any artificial intelligence system which could have a substantial impact on an individual's life, unless it can generate a full and satisfactory explanation for the decisions it will take." The intentions behind this principle are important, but it masks a crucial tension between using algorithms for social benefit ('beneficence') and ensuring those algorithms are fully intelligible to humans ('explicability'). For example, algorithms exist today that can diagnose medical conditions more accurately than doctors, potentially saving lives [24], but for which a full and satisfactory explanation cannot necessarily be provided (depending on how this is defined). In some situations, the benefit of using an algorithm may be high enough, and its accuracy reliable enough, that all users agree it is worth using even if a fully comprehensive explanation of its decisions cannot be given. There are complex and important trade-offs involved here [15, 22], and a principle that simply states that it is not acceptable to deploy AI systems without full explainability fails to recognise this.

In conclusion, there is a risk that high-level principles give the impression of being the outcome of meaningful debate about how AI should be developed, but in reality they are simply postponing it.

4 WHY THE FIELD SHOULD FOCUS ON TENSIONS

We use the term 'tension' to refer to any conflict, whether apparent, contingent or fundamental, between important values or goals, where it appears necessary to give up one in order to realise the other. For example, the use of socially beneficial data-driven technologies might make it impossible for us to fully guarantee otherwise desirable levels of data privacy. If the potential gains of these technologies are significant enough - new and highly effective cancer treatments, say - we might decide that a higher risk of privacy breaches is a price worth paying.

In some cases, a tension may reflect a strict moral tradeoff: a situation where two values or goals conflict and it is not possible to get more of one without sacrificing another. However, many tensions in AI are more contingent, and arise as a result of current technological or societal constraints. Using machine learning for social benefit may not be fundamentally in tension with privacy, transparency, or fairness, but many current methods employed for the former goal do conflict with these ideals. We do not yet know how far new technological or governance solutions could go to dissolve these tensions.

Others have acknowledged the importance of recognising conflicts between values in AI ethics, but to our knowledge none have explored in detail why this would be beneficial or what it would look like in practice. For example, [9] say that, "Ensuring socially preferable outcomes of AI relies on resolving the tension between incorporating the benefits and mitigating the potential harms of AI, in short, simultaneously avoiding the misuse and underuse of these technologies", but do not discuss specific tensions in detail or how to resolve them.

In this section, we discuss some of the benefits of focusing on tensions. We outline four reasons this is an important next step for AI ethics: (1) bridging the gap between principles and practice, (2) acknowledging differences in values, (3) highlighting areas where new solutions are needed and (4) identifying ambiguities and knowledge gaps.

4.1 Bridging the Gap between Principles and Practice

In general, we see focussing on tensions as an important way of bridging the gap between abstract ethical principles and specific cases, and therefore an important first step towards an ethics of AI that is practical and action-guiding.

To identify tensions, we need to consider how different values and goals might come into practice in concrete cases. For example, when Google DeepMind collaborated with the Royal Free Hospital, they encountered a conflict between protecting the privacy of patient data, and their goal of using AI to improve early diagnosis

of acute kidney injury [20]. Since similar tensions will likely arise across a range of different cases, focusing on tensions means we are neither driven entirely by the specifics of an individual case, nor are we relying on abstract high-level values. If we can articulate important tensions by looking at a range of cases, and find ways to resolve them in specific scenarios, what we learn from this can then be used to develop standards and regulation that are more sensitive to how principles apply differently across scenarios.

4.2 Acknowledging Differences in Values

Focusing on tensions forces us to consider how different values might be interpreted and endorsed differently across groups. While some important tensions are due to conflicts between principles in practice, others arise because there are conflicting meanings or values within a single principle: broad terms like 'fairness' or 'justice' for example are subject to substantial moral and political disagreement [5, 8].

It may never be possible to totally resolve all of these disagreements. But clearly articulating them is a crucial starting point for ensuring that all aspects of AI ethics are as inclusive as possible: for example, to ensure that international standards take full account of and accommodate cultural differences, and that agreement on such standards is meaningful.

4.3 Highlighting Areas Where New Solutions Are Needed

Noting a tension between two values does not necessarily mean we are forced to choose between them: often, we may be able to find some way to get more of both things we value. Recognising these tensions can therefore highlight high priority areas for both researchers and policymakers. For example, acknowledging that there is currently a tradeoff between performance and interpretability in state-of-the-art machine learning systems has motivated technical research that attempts to reduce or eliminate this trade-off [2].

Acknowledging tensions will also help direct the development of AI in beneficial directions more generally. It is currently far from clear whether advances in AI will augment or degrade human capabilities and agency, but making this tension explicit focuses attention on the important question of which trajectories of development are most likely to lead to the former.

4.4 Identifying Ambiguities and Knowledge Gaps

Finally, a tension-focused approach helps to clearly highlight ambiguities and gaps in our understanding of how uses of AI are impacting society. Again, this can help identify new and important research directions.

To think clearly about all tensions, we need to recognise and clarify ambiguities in terms: what do we really mean by things like 'fairness', 'justice', and 'autonomy', and how might these be interpreted differently across groups and contexts? To understand the nature of many tensions we need to understand what is currently technically possible: what are the best current methods for ensuring data privacy in machine learning, for example, and what are the costs of these methods? To understand how tensions arise in practice, we need better evidence on how AI is actually being applied in society today: what effect is automation already having on individual lives across different sectors? And to articulate and resolve conflicts between the interests of different groups, we need to really understand the needs and values of affected communities: how do the trade-offs people are willing to make differ based on demographic factors, for example? Focussing on tensions should help to drive this important work.

5 WHICH TENSIONS?

There are several different ways that applications of AI can introduce tensions between important goals and values.

Some tensions arise due to the very nature of AI and machine learning: these techniques allow us to use and draw inferences from very large amounts of (often personal) data, and so challenge important notions of privacy. The most useful models also often quickly become very complex, introducing new issues around human interpretability [11, 26]. This means we face tensions between using these technologies for socially beneficial goals: improving healthcare, justice, or security, for example, and other goals such as respecting privacy and maintaining trust and understanding in automated systems.

Another possibility is that AI systems exacerbate already existing ethical or societal tensions: between different conflicting notions of fairness, for example. Often this is due to the fact that machine learning models are trained on historical data, and so inherit the biases or mistakes they contain [13]. Here, applications of AI in society do not necessarily introduce new tensions, but increase the importance of already-existing ones such as how to make decision-making more accurate and efficient without inadvertently discriminating against minority groups.

Other tensions arise because the harms and benefits of AI systems are unequally distributed in various ways. For example, the impacts of automation may be unequally distributed across populations and cultures: enhancing the agency of some groups by automating mundane tasks while wiping out the livelihood of others, thus threatening their basic needs. The risks and benefits of AI systems could also be unequally distributed over time, and uses of AI that present opportunities in the near term may compromise important long-term values. Increasing personalisation of messages and services may make our lives more convenient and enjoyable in the short run, but begin to undermine important aspects of autonomy, equality and solidarity over time [21].

Finally, AI may have the potential to both enhance and threaten a given value. For example, depending on the precise direction in which technology develops, it could be used to either greatly enhance human capabilities - if we can develop sophisticated methods of intelligence augmentation [7] - or to degrade them: if our own capabilities atrophy as we outsource more and more tasks [6]. As mentioned above, automation might enhance the agency of some groups while threatening the autonomy of others: whether we see AI as enhancing or degrading human agency could depend on how narrow or global a view we take of its impacts.

6 FOUR KEY TENSIONS

Given the wide range of tensions that may arise from applications of AI, now or in the future, there is unlikely to be an exhaustive list

of all possible tensions. However, we believe that the following four tensions will be particularly central to thinking about the ethical issues arising from the applications of AI systems in society today. These capture a range of issues which are already salient or likely to grow in importance moving forward.

Tension 1: Using data to improve the quality and efficiency of services vs. respecting privacy and autonomy of individuals. Machine learning and big data are already being used to improve various public services (including healthcare, education, and social care). These improvements could be hugely beneficial to citizens, but require large amounts of personal data, raising concerns about how to best protect privacy and ensure meaningful consent.

Tension 2: Using algorithms to make decisions and predictions more accurate vs ensuring fair and equal treatment. This tension arises when public or private bodies base decisions on predictions about future behaviour of individuals (e.g. when probation officers estimate risk of reoffending) and when they employ machine learning algorithms to improve their predictions. These algorithms may improve accuracy overall, but discriminate against specific subgroups for whom representative data is not available.

Tension 3: Reaping the benefits of increased personalisation in the digital sphere vs enhancing solidarity and citizenship. Companies and governments can use personal data to tailor the messages, offers, and services people see. This personalisation can make it easier for people to find the right products and services for them, but differentiating between people in such fine-grained ways may threaten societal ideals of citizenship and solidarity.

Tension 4: Using automation to make people's lives more convenient and empowered vs promoting selfactualisation and dignity. Automated solutions may genuinely improve people's lives by saving them time on mundane tasks that could be better spent on more rewarding activities. But they also risk disrupting some of the practices that are an important part of what makes us human. With automation we may see the gifts of arts, languages and science become more accessible to those who were excluded in the past - but we may also see widespread deskilling, atrophy, ossification of practices, homogenisation and cultural diversity.

7 IDENTIFYING FURTHER TENSIONS

The above tensions are important and represent areas where exploring tensions is likely to be fruitful for AI ethics. Going forward, further such areas can and should be identified. In order to do so, it is helpful to ask a range of questions, including:

- Where AI is being used to serve a particular goal or value, or for 'social benefit' in general, what risks to other values are introduced?
- Where might uses of AI that benefit one group, or the population as a whole, have negative consequences for a specific subgroup? How do we balance the interests of different groups?
- Where might applications of AI that are beneficial in the near-term introduce risks in the long-term? How do we trade-off short and long-term impacts of society?
- Where might future developments in AI either enhance or threaten important values, depending on the direction they take?

8 RESOLVING TENSIONS

The best approach to resolving a tension will depend on the nature of the tension in question.

Where a strict trade-off between two values exists, a choice must be made to prioritise one set of values over another. For example, this may mean judging what risks to privacy it is acceptable to incur for the sake of better public health, or where to reject innovative automation technologies because the threats they pose to human skills and autonomy are too great.

Making these trade-off judgements will be a complex political process. Weighing the costs and benefits of different solutions can be an important part of the process but alone is not enough, since it fails to recognise that values are vague and unquantifiable, and that numbers often hide complex value judgements. In addition, resolving trade-offs will require extensive public engagement, to give voice to a wide range of stakeholders and articulate their interests with rigour and respect.

On the other hand, where tensions are more practical in nature, strict trade-offs may not be inevitable. It may be that we simply lack the knowledge or tools to advance conflicting values, and investing in further research could identify solutions that better serve all relevant values or goals. For example, it might be possible to use automation to improve people's lives without sacrificing self-actualisation and devaluing human skills, if a clear line can be drawn between the contexts where we do and do not want to pursue automation.

In these situations we face a choice. Even if a tension between two goals is not a fundamentally irresolvable one, if we want to apply current technology in society, we will still need to make the kinds of trade-offs described above. On the other hand, if we can hold-off from implementing technologies that introduce tensions - certain kinds of automation, say - then we could instead invest in more research on how technological or governance solutions might reduce the need to navigate potentially difficult value trade-offs. Of course, this is not a binary choice: we can choose to strike a balance by making the trade-offs necessary to implement technology where doing so is relatively unproblematic, while still investing in research to explore how these tensions might be resolvable in future. This choice can be thought of as involving its own tension, between short- and long-term interests: to what extent should we postpone the benefits of new technologies in order to invest the time and resources necessary to resolve the tensions they introduce?

9 CONCLUSION

Over the last two years, many different sets of principles for the ethical use of AI have been developed. We argue that these high-level principles, rather than representing the outcome of meaningful debate on how AI should be developed, risk simply postponing it. In order to make AI ethics practical, the field needs to now focus more on the tensions that arise when principles are applied to concrete cases.

Though most of these tensions cannot be resolved straightforwardly, we believe articulating them more clearly and explicitly has several benefits. To be useful in practice, principles need to be formalized in standards, codes and ultimately regulation. To be

effective, these in turn must acknowledge that there are tensions between the different high-level goals of AI ethics, and provide some guidance on how they should be resolved in different scenarios. They also need to acknowledge and accommodate different perspectives and values as far as possible, if they are to reflect genuine agreement.

A focus on tensions can also help to direct research priorities in AI ethics. Articulating tensions can help to highlight important ambiguities and gaps in our understanding of how AI is currently being applied in society which need further research. More generally, much current research in AI ethics appears to be driven by questions of how we ensure that uses of AI respect important values, such as privacy, transparency, or fairness. Reframing research questions to be more focused on understanding and resolving tensions is an important step towards solving practical problems arising from the use of AI in society, since it directs attention to where new technological or governance solutions might help push the development of AI in robustly beneficial directions.

ACKNOWLEDGMENTS

This work was funded by the Nuffield Foundation and by a Leverhulme Trust Research Centre Grant.

REFERENCES

[1] ACM US Public Policy Council 2017. Statement on Algorithmic Transparency and Accountability. https://www.acm.org/binaries/content/assets/publicpolicy/2017_usacm_statement_algorithms.pdf
[2] Tameem Adel, Zoubin Ghahramani, and Adrian Weller. 2018. Discovering interpretable representations for both deep generative and discriminative models. In *International Conference on Machine Learning*. 50–59.
[3] Tom L. Beauchamp. 1995. Principlism and its alleged competitors. *Kennedy Institute of Ethics Journal* 5, 3 (1995), 181–198.
[4] Tom L. Beauchamp and James F. Childress. 2001. *Principles of biomedical ethics*. Oxford University Press, USA.
[5] Reuben Binns. 2017. Fairness in machine learning: Lessons from political philosophy. *arXiv preprint arXiv:1712.03586* (2017).
[6] Adam J. Carter. 2018. Autonomy, cognitive offloading and education. *Educational Theory* (2018).
[7] Shan Carter and Michael Nielsen. 2017. Using artificial intelligence to augment human intelligence. *Distill* 2, 12 (2017), e9.
[8] Danner K. Clouser and Bernard Gert. 1990. A critique of principlism. *The Journal of medicine and philosophy* 15, 2 (1990), 219–236.
[9] Josh Cowls and Luciano Floridi. 2018. Prolegomena to a White Paper on an Ethical Framework for a Good AI Society. (2018).
[10] Jonathan Dancy. 2004. *Ethics without principles*. Oxford University Press on Demand.
[11] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).
[12] Future of Life Institute 2017. Asilomar AI Principles. https://futureoflife.org/ai-principles/
[13] Sara Hajian, Francesco Bonchi, and Carlos Catillo. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2125–2126.
[14] IEEE 2017. Ethically Aligned Design, Version 2. https://ethicsinaction.ieee.org/
[15] W. Nicholson Price II. 2017. Regulating black-box medicine. *Mich. L. Rev.* 116 (2017), 421.
[16] Japanese Society for Artificial Intelligence 2017. The Japanese Society for Artificial Intelligence Ethical Guidelines. http://ai-elsi.org/archives/514
[17] Albert R. Jonsen and Stephen E. Toulmin. 1988. *The abuse of casuistry: A history of moral reasoning*. Univ of California Press.
[18] Partnership on AI 2018. Tenets. https://www.partnershiponai.org/tenets/
[19] Sundar Pichai. 2018. AI at Google: Our Principles. https://www.blog.google/technology/ai/ai-principles/.
[20] Julia Powles and Hal Hodson. 2017. Google DeepMind and healthcare in an age of algorithms. *Health and technology* 7, 4 (2017), 351–367.
[21] Barbara Prainsack and Alena Buyx. 2017. *Solidarity in biomedicine and beyond*. Vol. 33. Cambridge University Press.
[22] Andrew D. Selbst and Solon Barocas. 2018. The intuitive appeal of explainable machines. *Fordham L. Rev.* 87 (2018), 1085.
[23] Select Committee on Artificial Intelligence 2018. AI in the UK: Ready, Willing, and Able? https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf
[24] Ming Song, Yi Yang, Jianghong He, Zhengyi Yang, Shan Yu, Qiuyou Xie, Xiaoyu Xia, Yuanyuan Dang, Qiang Zhang, Xinhuai Wu, Yue Cui, Bing Hou, Ronghao Yu, Ruxiang Xu, and Tianzi Jiang. 2018. Prognostication of chronic disorders of consciousness using brain functional networks and clinical characteristics. *eLife* 7 (2018), e36173.
[25] University of Montreal 2017. Montreal Declaration on Responsible AI. https://www.montrealdeclaration-responsibleai.com/the-declaration
[26] Adrian Weller. 2017. Challenges for transparency. *arXiv preprint arXiv:1708.01870* (2017).
[27] Alan F. Winfield and Marina Jirotka. 2018. Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2133 (2018), 20180085.