Narrow Rules are not Enough

Ondrej Bajgar 2022-08-11T16:27:40

As artificial intelligence (AI) will continue getting integrated into important places in our society and getting more capable, the associated risks may start becoming too large and start coming too quickly for reactive regulation to be effective. At the same time, humans‘ ability to foresee future harms is inevitably limited, especially at the level of specificity usually required for detailed legal regulation. If these two considerations limit our ability to get specific in the rules we are laying down, a path forward is to lean onto more general principles, but give them teeth by making them legally binding. We think negative human rights could go a long way in protecting us against unforeseen harms from AI, but we need to go beyond how they are understood today in human rights law.

To ensure AI — any of a broad family of human-made systems autonomously performing tasks that people would generally associate with intelligence — stays safe over the decades to come, we need both technical and regulatory solutions. These cannot be created in isolation from each other — regulation cannot demand what is technically infeasible, while technical solutions that cannot be enforced legally may fall short of ensuring safety. We think a framework based on negative human rights has a lot to offer on both sides.

AI and its unforeseen harms

The long-term AI safety community has been arguing that rather than narrower rules and tasks, we will need to equip advanced AI systems with *general* principles to decide which actions or outcomes are desirable to humans and which should be avoided. AI systems — especially the more capable future ones — are likely to come up with innovative solutions that fulfil any narrower set of instructions exceptionally well in letter but oppose wider human values (and possibly even the spirit of the original instructions).

For instance, we have already seen claims that recommender systems may be pushing users toward more extreme views, which helps the algorithms fulfil their narrow goal of maximising time spent on a website by making the users more predictable but violates wider human preferences over what the algorithm should be doing.[1) As AI systems continue becoming more capable, the costs of such specification failures are likely to grow and may go all the way to existential risks to humanity.[2)

Thus, a part of the long-term AI safety community has been advocating for AI systems to be taught full human values or preferences. However, if we accept the need for global regulation of AI, is there a full set of human values that regulation could mandate for all AI systems globally? Whose values should those be? At least in the foreseeable future, we don’t consider an agreement on a set of human values for regulatory purposes realistic.

Human rights and the specification problem

Negative human rights are in a better position to inspire broad international consensus. Though often under different names (such as constitutional rights or fundamental rights), they are to some extent recognized by most countries, including China as well as many African and South American countries. We are not making a particular moral argument for human rights. Different people and nations may found their protections of rights on different bases — religious, ethical, or political. However, we empirically observe that the results of their arguments largely intersect in certain basic protections and thus form what Rawls termed an *overlapping consensus*.

The protection of most negative human rights has much deeper roots than the concept of human rights as such. What we mean by protecting human rights largely serves the same objectives that criminal law and some other areas of law have served for centuries across the world, and what is thus less emphasised under human rights law for historical reasons — we would like to include those rights, aiming to include also what we could term generalised criminal law. New rights would need to be included — for instance, Stuart Russell suggests a right to mental security, which includes e.g. a right not to be lied to by AI systems — while other rights would inevitably need to get added over time as consensus on them emerges.

Building human-rights-respecting AI

Human rights may seem like an abstract concept, so we need to take care to give them concrete meaning. Then we need to convey this meaning to AI systems and ensure they behave accordingly. We think the legal world already has a tentative answer to the first question in the form of judicial systems, and we think the European Court for Human Rights can serve as an illustrative example.

While the European Convention on Human Rights is written in similarly abstract terms as the Universal Declaration of Human Rights, its meaning becomes much more concrete thanks to the European Court for Human Rights — for any concrete behaviour brought forward, the court makes in essence a binary decision: either the behaviour did violate human rights, or it didn’t.

We think an (eventually AI-augmented) human adjudication system can similarly give concrete meaning to human rights in the case of AI. AI could be eventually trained to internally predict whether a behaviour it is planning risks being classified by the adjudication system as violating human rights, and if yes, it should refrain from behaving so. This could be split, for instance, into training AI systems to predict the effects of their behaviour on the world first (as is done in model-based reinforcement learning or planning), and then classifying the predicted trajectories into permissible and impermissible ones.

The most obvious learning source for the latter would be case law from existing courts; however, other useful sources include reinforcement-learning-like reward or knowledge captured in large language models, such as GPT-3.

Framing compliance with human rights as such a classification problem turns it from an abstract aspiration into a kind of task that the machine learning research community is used to solving. The fact that, relative to many alternative proposals and principles for AI, human rights are amenable to both technical and regulatory work is one of their distinguishing strong points.

Human rights as a path to sustainable regulation

What could this mean for the legal regulation of AI? In most cases, if an AI system violates human rights, it would be illegal already now. However, in most cases, it would be illegal through more specific law that does not explicitly mention human rights. Future AI systems may start causing problems very distant from current legal practice and those would risk falling through the cracks of any law more specific than human rights. To prevent this we would suggest directly making it a legal obligation for the producers of autonomous AI systems to prevent violations of negative human rights as defined in a particular convention, with liability in case of violation, instead of relying on a patchwork of more specific regulation.

More specific regulation is needed, also to make the legal environment more predictable and to provide more specific guidance on what preventive measures should be taken. However, human rights could stand in the background to step in when needed, as well as to provide an aspirational framing, which a part of the more specific regulation could be implementing.

If there is a risk of the AI system violating human rights as a result of a decision of its users or operators, then for sufficiently powerful future systems, we could require there to be a safety mechanism that would prevent such behaviour on the part of AI anyway. Or in cases where this is not practical, such a risk should be clearly communicated to the users in operation instructions so that they could assume legal responsibility similarly to other cases where a user is responsible for the consequences of using a product with harmful consequences.

Conclusion

With continuing proliferation of increasingly capable AI systems, we will need regulation to address the associated risks. Since our ability to foresee such future risks is very limited, our best bet is to base such regulation on relatively general principles, rather than narrow rules. As we also argue in our upcoming paper, we think that negative human rights with their existing broad international support could form a suitable foundation both for flexible regulation and for the associated technical solutions.

References

- Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Allen Lane.
- As has been argued e.g. by Bostrom, Petit, Critch & Krueger, or Ngo.