Horizon Scanning Series

The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing

Liability

This input paper was prepared by Olivia Erdélyi and Gábor Erdélyi

Suggested Citation

Erdélyi, O and Erdélyi, G (2018). Liability. Input paper for the Horizon Scanning Project "The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing" on behalf of the Australian Council of Learned Academies, www.acola.org.

The views and opinions expressed in this report are those of the author and do not necessarily reflect the opinions of ACOLA.

Considerations on the Liability of Artificial Intelligence Systems

Olivia J. *Erdélyi* School of Law University of Canterbury

Gábor Erdélyi School of Economic Disciplines University of Siegen and School of Mathematics and Statistics University of Canterbury

August 5, 2018

1 Background

In more or less perceptible forms, artificial intelligence (AI) increasingly impacts our lives — autonomous vehicles, growing reliance on predictive analytics in health care, the criminal justice system, and in financial services, and the deployment of ever more autonomous weapons systems are but a few examples to mention here. Incrementally, these highly transformative technologies will (and have already begun to) fundamentally change our world, putting humanity in front of some tough choices regarding the very essence of core societal values and institutions. With the proliferation of human-AI interactions both in the civil and military context, issues around the civil and criminal liability of AI systems are among the more pressing problems, which require urgent solution and are hence moving to the forefront of policy debates. Accordingly, a growing number of commentators are taking a first crack at examining the topic from various perspectives, pointing out dangerous gaps in the existing legal framework [6] and providing conflicting accounts on the most appropriate type of liability provisions to capture AI liability and the legal system's overall ability to adapt to this latest wave of technological innovation [4, 2, 3].

While we share the view that policy action addressing AI liability across various domains is sorely needed, we maintain that the time is not yet ripe for advancing specific regulatory proposals in this regard. This is because we currently lack conceptual clarity both on the notion of AI in general and on most of its relevant properties, meaning essentially that virtually all key parameters that could serve as benchmarks for regulation are at best ill-defined. Instead of formulating any proposal on how AI systems could be held liable in different situations, the next sections therefore seek to show the extent and implications of this conceptual ambiguity, which characterizes both legal and AI research. In light of these findings, we would also like to caution against premature and at this point by definition speculative action, and stress that further research both in law and AI to develop precise and universally accepted definitions must precede concrete regulatory proposals.

2 Conceptual Ambiguity in Legal and AI Research

Legal contributions discussing various aspects of the regulatory treatment of AI systems typically either handle the concept of AI as given and thus refrain from defining what they mean by AI, or choose a working definition that is best suited to their particular inquiry. This is perhaps unsurprising, given that even AI researchers have so far not managed to work out a universal definition. Instead, the tacit assumption is that AI is a system that mimics certain aspects of human cognition, and approaches to defining AI have broadly focused on comparing AI systems' cognitive and behavioral abilities to human and rational behavior [9, 1].

Although the absence of a universally agreed-upon definition may not have hampered AI re- search, a consistent understanding and definition of the concept of AI or at least of its particular aspects subject to regulation are indispensable for the purpose of adequately regulating it. For ex- ample, does a mental picture of AI as a being with the ability to think or act humanly warrant a commensurate regulatory treatment when it comes to holding the system itself or — given the absence of AI's legal personality — the people contributing to its design and/or distribution liable for personal injuries or property damage? Can we expect an AI to *know* enough of the world to *foresee* a certain negative outcome and apply *reasonable care* to prevent such harm from occurring? And can we hold it liable for negligent behavior in case it fails to *behave* this way?

Just posing these hitherto unanswered questions reveals a much more fundamental, yet barely recognized problem: the importance of approaching the AI regulatory debate with the right mindset and the fact that, as things stand, society — whether policymakers or the general public — has not yet fully understood the nature and potential of AI technologies. We are still searching for the right *metaphor*, as Richards and Smart put it in regard to robotics [8]. This, they go on to point out, is a much graver problem than it may seem at first sight, because it is human nature — reinforced by meticulous legal training — to think in analogies, which entails that the metaphors we choose to understand AI will then critically influence the design of different AI instantiations, as well as their regulation and social acceptance.

So, as a preliminary matter, both policymakers and society at large need to be conscious of the fact that AI does not *know*, *think*, *foresee*, *care*, or *behave* in the anthropomorphic sense, rather it applies what could be best described as *machine logic*. To illustrate the potential implications of that distinction, consider the following example: Machine learning (ML)-based systems — which raise the biggest technical and legal challenges due to their unpredictability stemming from their independent learning property — do not *know* why a given input should be associated with a specific label (e.g., that a small, red, circular object is a ball), only that certain inputs are *correlated* with that label [5]. That is, the system identifies outputs based on a set of predefined parameters and probability thresholds through a process that is fundamentally different from human thinking. What is more, this type of machine reasoning always implies a certain probability of failure, where the failure tends to occur in — from a human perspective — unexpected ways. These failures can again have different reasons. Let us give two examples.

In the first example, the failure can be a bad classifier as illustrated by Ribeiro et al. in their Husky vs. Wolf experiment [7]. The task is to distinguish between pictures of wolves and huskies. In order to do so, they trained the system with 10 wolf and 10 husky pictures. On purpose, all wolf pictures had snow in the background but none of the husky pictures. Since snow is a common element in the wolf pictures and is not present in the husky pictures, the system regards snow as a classifier for wolves. Thus, in the experiment the system predicts huskies in pictures with snow as wolves and vice-versa. The second example shows a way of cheating a facial recognition system (FRS) introduced by Sharif et al. [10]. FRS's are usually using neural networks in order to recognize patterns in big datasets, in this case the differences between millions of faces (for example the relative position of nose and eyebrows, size of the nose, etc.). Sharif et al. used a pair of glasses with a colorful frame which basically interfered with the system's pattern recognition. It not just blocked the *view* to crucial parts of the faces but, due to the colorful frame, gave the system the impression that it sees some patterns. This way, the FRS often made mistakes despite indicating a high probability of confidence.

Another frequently discussed but poorly defined concept used in the context of AI liability is the *black box* attribute of certain ML-based AI systems. Here again, commentators are either inclined to omit defining this term altogether or give incorrect definitions exposing their lack of understanding of the distinction between the notions black box and *transparency*. Russel and Norvig [9] use the term black box in the context of atomic representation — the mode of representation of a learning agent's environment with the lowest level of complexity and expressive power, in which each state of the world is indivisible, i.e., has no internal structure. By that logic, with respect to ML-based AI systems, the black box metaphor refers to a state of zero knowledge of the system's internal workings — we do not know how the system works because we have no information of it, not because it is too complex for us to understand, as is sometimes suggested. On the contrary, problems of comprehending the system's workings logically presuppose knowledge of its operation. Such a system is said to be transparent, although, admittedly, AI systems significantly vary in their degree of transparency. Thus, black box and transparency are mutually exclusive concepts.

Contemplating ways in which AI systems could potentially be held liable without conceptual clarity on this attribute is problematic, as it crucially impacts on the foreseeability requirement, which is central to any form of legal liability, whether or not it involves some sort of mental element. Karnow [3] raises this point in regard to autonomous robots and classic tort doctrines — negligence and the various forms of strict liability — but at the most basic level his reasoning can also be applied to not embodied ML-based AI systems and criminal liability: holding someone responsible for a certain harm requires that said person can to some extent anticipate that harm; we cannot intend for or be negligent about something we cannot foresee. The picture is more complex in cases where no mental intent is required [4], but even here it can be argued that the rationale behind strict liability offenses is that they foreseeably lead to some undesired outcome. Intuitively, one would assume that while foreseeability cannot be given in the case of black box systems, it should not be a major problem as long as we are dealing with a transparent system, where we can comprehend the system's every move.

However, even this is not so simple, not least because the notion of transparency is itself subject to considerable conceptual ambiguity in the ML literature. In fact, a whole line of research in ML is concerned with the issue of how to ensure transparency (or interpretability) of ML models. Yet here again, as Lipton points out, many papers assume transparency axiomatically, and existing definitions reveal that transparency is far from being a monolithic concept. He identifies three distinct model properties used to facilitate *ex ante* transparency, namely simulatability, decomposability, and algorithmic transparency [5]. In *simulatability* we assume that a person can reflect the whole ML model at once. In *decomposability*, each part of the ML model (i.e., input, parameter, and calculation) admits an intuitive explanation. And finally, in *algorithmic transparency* we require a full under- standing of the learning algorithm itself, i.e., we expect to fully understand and reconstruct each and every step it makes. Without having to delve into further technicalities, this analysis already suggests that each of these notions of transparency may well require different levels of expertise in order to establish foreseeability. Additionally, these cases have to be distinguished from *ex post* transparency/interpretability, that is when we are able to understand how the system has achieved a given output for instance to seek explanation for an unforeseen — and from an *ex ante* perspective perhaps even unforeseeable — outcome. Note that this does not mean, that we can fully back-trace every step our ML model did. A final aspect of the transparency issue worth mentioning here is that there is always a trade-off between AI-performance and transparency. Transparent models usu- ally have much simpler structures than black-box models in order to be understandable for humans. This, of course, deduces a severe limitation in regards of accuracy and flexibility [7].

3 Summary

By discussing these conceptual ambiguities, we have only meant to provide a narrow snapshot of problems that currently stand in the way of devising concrete policy initiatives in relation to AI liability. Lawyers may have to befriend with the idea that foreseeability as the primary benchmark for imposing liability needs to be replaced with something else in the context of AI — or face a different set of unexpected challenges. In any case, society will need to adapt the law to the changing realities of our AI-driven world and our guiding principles in the course of those reflections should probably be the core societal values we intend to preserve. We also believe that the design of AI related policies — whether in the context of liability or in any other area — will require looking at things from a broader perspective, taking account of multidisciplinary imperatives in collaboration with multiple stakeholders.

References

[1] R. Calo. Artificial Intelligence Policy: A Primer and Roadmap, 2017. August 8, 2017, Available at SSRN: https://ssrn.com/abstract=3015350.

[2] F. P. Hubbard. Allocating the risk of physical injury from ”sophisticated robots”: Efficiency, fairness, and innovation. In R. Calo, A. M. Froomkin, and I. Kerr, editors, *Robot Law*, pages 25–50. Edward Elgar Publishing, 2016.

[3] C. E. A. Karnow. The application of traditional tort theory to embodied machine intelligence. In R. Calo, A. M. Froomkin, and I. Kerr, editors, *Robot Law*, pages 51–77. Edward Elgar Publishing, 2016.

[4] J. K. C. Kingston. Artificial intelligence and legal liability. In M. Bramer and M. Petridis, editors, *Research and Development in Intelligent Systems XXXIII*, pages 269–279. Springer International Pub- lishing, 2016.

[5] Z. C. Lipton. The Mythos of Model Interpretability. *Queue*, 16(3):30:31–30:57, June 2018.

[6] T. McFarland and T. McCormack. Mind the Gap: Can Developers of Autonomous Weapons Systems be Liable for War Crimes. *International Law Studies*, 90(1):361–385, 2014.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin. ”Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, pages 1135–1144, New York, NY, USA, 2016. ACM.

[8] N. M. Richards and W. D. Smart. How should the law think about robots? In R. Calo, A. M. Froomkin, and I. Kerr, editors, *Robot Law*, pages 3–22. Edward Elgar Publishing, 2016.

[9] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 2 edition, 2003.

[10] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS ’16, pages 1528–1540, New York, NY, USA, 2016. ACM.