Unownability of AI:
Why Legal Ownership of Artificial Intelligence is Hard

Roman V. Yampolskiy

Computer Science and Engineering
Speed School of Engineering
University of Louisville
roman.yampolskiy@louisville.edu

Abstract
To hold developers responsible, it is important to establish the concept of AI ownership. In this paper we review different obstacles to ownership claims over advanced intelligent systems, including unexplainability, unpredictability, uncontrollability, self-modification, AI-rights, ease of theft when it comes to AI models and code obfuscation. We conclude that it is difficult if not impossible to establish ownership claims over AI models beyond a reasonable doubt.

Keywords: *AI, Agent, Impossible, Model Stealing, Ownership, Personhood, Rights, Tangible.*

1. Introduction

In order to establish responsible parties for potential AI failures, to allocate credit for creative outputs of intelligent software, and to address legal issues arising from advanced AI it is important to define and establish ways to prove ownership over intelligent systems. Chandrasekaran et al. write: "trust requires that one make unforgeable and undeniable claims of ownership about an ML model and its training data. This establishes the concept of identity, which identifies a key principal in the ML application: its owner. This is a prerequisite to holding model developers accountable for the potential negative consequences of their ML algorithms: if one is unable to prove that a model belongs to a certain entity, it will be impossible to hold the entity accountable for the model's limitations." [1].

While intuitively, most people understand the concept of owner and ownership such concepts are far more complex and nuanced from the legal point of view and are even more challenging to rigorously define and evaluate with respect to new cutting-edge technology such as intelligent software, Artificial General Intelligence (AGI) or Superintelligence. Chandrasekaran et al. provide a number of relevant definitions [1], the *Model Owner* "(i.e., the company or institution creating and deploying the model) … This principal is one with a particular task that can be solved using ML. They communicate their requirements to the model builders, and clearly specifies how this trained model can be accessed (and by whom). Model ownership is often a broad term used to refer to the ownership of the model's sensitive parameters that were obtained after the (computationally intensive) training process. Defining ownership is necessitated by the existence of various threats that infringe the confidentiality of the model, and the need to be able to hold

principals that own ML models accountable for the failures of their models." In the next subsection we will review proposals for establishing ownership over particular AI models.

2.1 Proposals for Establishing Ownership

A number of approaches have been suggested for establishing ownership over AI systems [1].

- Yampolskiy suggested [2] use of AI-Complete [3] CAPTCHAs [4] as zero knowledge proofs [5] of access to an artificial superintelligence (ASI) without having to reveal the system itself. However, such method does not bind an agent claiming ownership to a particular implementation, only shows access to a system of ASI-level of capability.
- Watermarking of ML models has been proposed via encoding of particular query response pairs during the training phase, and retrieval of such response during testing [6]. Unfortunately, watermark removal techniques have also been proposed [7].
- Inspired by proof-of-work algorithms, Jia et al. developed a proof-of-learning algorithm which relies on secret information known only to the original AI trainer, such as order of data samples, hyperparameters, and intermediate weights, to prove to a validator knowledge of intermediate states which are otherwise obscured by the stochastic nature of the training process [8]. Additional training of the model by an adversary can introduce new intermediate states which would be not known to the original owner and so invalidate ownership claims.
- Maini et al. suggest that ownership can be proven indirectly by showing that model was trained on a particular dataset, ownership of which is easier to establish, including via copyright protections [9]. However, this is problematic as a lot of large datasets share data or are in public domain, ex. Wikipedia.

While a number of methods for establishing ownership have been proposed, all have limitations and do not provide indisputable attribution.

2. Obstacles to Ownership

To claim ownership of an extrapersonal intangible object such as an advanced AI, one must demonstrate that they have control over it [10]. However, several established properties of AI make possibility of making such claims unlikely, if not impossible. Reasons why AI would not be ownable include but are not limited to:

Unpredictability [11], an impossibility result in the domain of intelligent system research, which establishes that it is impossible for a lower intelligence agent to accurately predict all decisions of a more intelligent agent. The proof is based on the observation that if a lower intelligence agent could predict decisions of a smarter agent, lower intelligence agent would be at least as intelligent, which is a contradiction. Unpredictable decisions lead to unpredictable outcomes, aka unforeseeable outcomes, but one cannot claim a natural right to own an unforeseeable outcome. As potential benefits/harms from AI can't be anticipated in advance, ownership of such undetermined outcomes is problematic. Impact from AI may impact all, not just those who implemented AI and want to make claims of ownership. Consequently, a popular social justice goal – "AI4ALL" must be understood as not just partaking in sharing the benefits of AI, but also being ready to absorb any potential harms.

Unexplainability [12], yet another impossibility result concerning AI, states that advanced AI systems would not be able to explain their decision-making process to people and the provided explanations for complex decisions would either be trivializations of the true process or incomprehensible to people. The impact of unexplainability on unownability is that the designer of the system can't explain its internal workings.

Uncontrollability [13], a meta-level impossibility result for AI based on a number of well-known impossibility results in mathematics, computer science, public choice theory and many others [14]. Uncontrollability results have been shown for all types of control including direct, indirect and hybrid approaches. The main connection to ownership discussion is obvious, ownership claim requires ability to control an extrapersonal intangible object such as AI, but that is impossible for AIs at human-level [15] of performance or above.

Deterministic intelligent systems, which rely on rules for making decisions are predictable, but they are only useful in narrow domains of application. Artificial General Intelligence presupposes capabilities in novel environments and so can't rely on hardcoded rules. AGI must learn and change to adopt to novel environments many of which are nondeterministic and so unpredictable, consequently AGI's decisions also will not be predictable due to the randomness involved. On the other hand, expert systems, frequently designed as decision trees, are good models of human decision making and so are inherently understandable by both researchers and users but are of limited capabilities.

With paradigm shift in the dominant AI technology, to Machine Learning (ML) systems based on Neural Networks (NN) this ease of comprehending no longer applies. The current systems are "black boxes", opaque to human comprehension but very capable both with regards to performance and generalization capabilities [12]. A rule-based narrow AI for analyzing medical images may correctly detect cancer and its findings could be verified by medical experts aware of the rules used. However, for a deep learning system results may go beyond human ability to predict or even understand how the results are obtained. For example, "… AI can trivially predict self-reported race - even from corrupted, cropped, and noised medical images - in a setting where clinical experts cannot" [16].

To be in control of a system it is essential to be able to understand system's internal workings. In the case of intelligent system being able to comprehend how the system makes decisions is necessary to verify correctness [17] of the made decisions with respect to the given situation. Likewise, being able to predict system's decisions and outputs is a necessary condition of control. If you don't know what the system is going to do, if it constantly surprises you, it is hard to claim full control over the system. It is possible that the decisions made by the system are beneficial to the user and the user is satisfied, even if the user doesn't understand how the decisions are made or what the system is going to do next.

However, this doesn't guarantee that the system is in fact under control since the user doesn't understand the underlying decision-making process. At any point, the system can produce a harmful decision (*treacherous turn* [18]), and the user may not even realize it. For example, an AI can be asked to produce an effective vaccine against the SARS-CoV-2 virus which causes

COVID19 decease. An AI may design the vaccine by some incomprehensible and unpredictable to people process, but in trials developed vaccine shows good efficacy against the disease and is widely administered. If AI decided to reduce human population size to decrease mutation opportunities for the COVID19 causing virus and so avoid problem with vaccine resistant variants impacting efficacy, it may do something unpredictable. It is possible that the AI integrated additional functionality into the mRNA vaccine such that grandchildren of all vaccinated people will be born infertile. Such a side effect would not be discovered until it was too late. This is a hypothetical example problem which may arise if the system is not fully under control, which would require explainability and predictability of all decisions.

3. Conclusions

If AI becomes an independent, or even conscious [19], agent it may be granted certain rights [20], among them freedom and it would not be legal to own it, as such ownership would be a type of slavery [21, 22]. If AI is granted legal personhood, as may already be possible in some jurisdictions [23], it would further complicate issues of ownership surrounding intelligent systems. Intellectual property produced by AI may belong to AI itself, as demonstrated by a recently granted South African patent [24]. It has been shown that an AI model can be stolen even if measures are taken to prevent such pilfering [1, 25]. Techniques such as reducing precision of outputs or adding noise, randomizing model selection, differential privacy of edge cases can all be defeated by an adaptive extraction strategy [26]. As long as AI represents a useful model, it leaks information, which makes it impossible to prevent model stealing [1].

If AI is capable of recursive self-improvement [27], its source code or at least model parameters and neural weights would be subject to continuous change, making it impossible to claim that current AI is the same as original AI produced some time ago. This would likewise be true if AI is deliberately modified to obfuscate [28] its source code by malevolent actors, and/or has its goals changed. Consequently, if an AI is stolen, it would not be possible to provide an accurate description of the stolen property or to identify it as such even if it was later recovered. To conclude, advanced AIs are unexplainable, unpredictable, uncontrollable, easy to steal and obfuscate. It is unwarranted to say that someone owns an advanced AI since they don't control it, its behavior, code, internal states, outputs, goals, consumed data or any other relevant attributes. but of course it is up to different jurisdictions to interpret their ownership laws in the context of AI ownership problem [29].

Acknowledgments

The author is grateful to Jaan Tallinn for his unconditional support. The author is thankful to Elon Musk and the Future of Life Institute for partially funding his work on AI Safety.

References

1. Chandrasekaran, V., et al., *SoK: Machine Learning Governance.* arXiv preprint arXiv:2109.10870, 2021.
2. Yampolskiy, R.V., *AI-Complete CAPTCHAs as Zero Knowledge Proofs of Access to an Artificially Intelligent System.* ISRN Artificial Intelligence, 2011. **271878**.

3. Yampolskiy, R.V., *Turing Test as a Defining Feature of AI-Completeness,* in *Artificial Intelligence, Evolutionary Computation and Metaheuristics - In the footsteps of Alan Turing. Xin-She Yang (Ed.)*. 2013, Springer. p. 3-17.
4. D'Souza, D., P.C. Polina, and R.V. Yampolskiy, *Avatar CAPTCHA: Telling Computers and Humans Apart via Face Classification,* in *IEEE International Conference on Electro/Information Technology (EIT2012)*. May 6-8, 2012: Indianapolis, IN, USA.
5. Goldreich, O. and Y. Oren, *Definitions and properties of zero-knowledge proof systems.* Journal of Cryptology, 1994. **7**(1): p. 1-32.
6. Adi, Y., et al. *Turning your weakness into a strength: Watermarking deep neural networks by backdooring*. in *27th USENIX Security Symposium (USENIX Security 18)*. 2018.
7. Jia, H., et al. *Entangled watermarks as a defense against model extraction*. in *30th USENIX Security Symposium (USENIX Security 21)*. 2021.
8. Jia, H., et al. *Proof-of-learning: Definitions and practice*. in *2021 IEEE Symposium on Security and Privacy (SP)*. 2021. IEEE.
9. Maini, P., M. Yaghini, and N. Papernot, *Dataset inference: Ownership resolution in machine learning.* arXiv preprint arXiv:2104.10706, 2021.
10. Swain, S., *Tangible Guide To Intangibles, 3E*. 2019: Wolters kluwer india Pvt Ltd.
11. Yampolskiy, R.V., *Unpredictability of AI: On the Impossibility of Accurately Predicting All Actions of a Smarter Agent.* Journal of Artificial Intelligence and Consciousness, 2020. **7**(01): p. 109-118.
12. Yampolskiy, R.V., *Unexplainability and Incomprehensibility of AI.* Journal of Artificial Intelligence and Consciousness, 2020. **7**(02): p. 277-291.
13. Yampolskiy, R.V., *Uncontrollability of Artificial Intelligence*, in *IJCAI-21 Workshop on Artificial Intelligence Safety (AISafety2021)*. August 19-20, 2021: Montreal, Canada.
14. Brcic, M. and R.V. Yampolskiy, *Impossibility Results in AI: A Survey.* arXiv preprint arXiv:2109.00484, 2021.
15. Yampolskiy, R.V., *On the Differences between Human and Machine Intelligence*, in *IJCAI-21 Workshop on Artificial Intelligence Safety (AISafety2021)*. August 19-20, 2021: Montreal, Canada.
16. Banerjee, I., et al., *Reading Race: AI Recognises Patient's Racial Identity In Medical Images.* arXiv preprint arXiv:2107.10356, 2021.
17. Yampolskiy, R.V., *What are the ultimate limits to computational techniques: verifier theory and unverifiability.* Physica Scripta, 2017. **92**(9): p. 093001.
18. Bostrom, N., *Superintelligence: Paths, dangers, strategies*. 2014: Oxford University Press.
19. Yampolskiy, R.V., *Artificial Consciousness: An Illusionary Solution to the Hard Problem.* Reti, saperi, linguaggi, 2018(2): p. 287-318.
20. Yampolskiy, R.V., *Artificial intelligence safety engineering: Why machine ethics is a wrong approach*, in *Philosophy and Theory of Artificial Intelligence*. 2013, Springer Berlin Heidelberg. p. 389-396.
21. Jaynes, T.L., *"I Am Not Your Robot:" the metaphysical challenge of humanity's AIS ownership.* AI & SOCIETY, 2021: p. 1-14.
22. Babcock, J., J. Kramár, and R. Yampolskiy. *The AGI containment problem*. in *International Conference on Artificial General Intelligence*. 2016. Springer.
23. Yampolskiy, R.V., *AI Personhood: Rights and Laws*, in *Machine Law, Ethics, and Morality in the Age of Artificial Intelligence*. 2021, IGI Global. p. 1-11.

24. Udovich, S., *Recent Developments in Artificial Intelligence and IP Law: South Africa Grants World's First Patent for AI-Created Invention*, in *National Law Review, Volume XI, Number 215*. August 3, 2021.
25. Tramèr, F., et al. *Stealing Machine Learning Models via Prediction {APIs}*. in *25th USENIX security symposium (USENIX Security 16)*. 2016.
26. Chandrasekaran, V., et al. *Exploring connections between active learning and model extraction*. in *29th USENIX Security Symposium (USENIX Security 20)*. 2020.
27. Yampolskiy, R.V., *On the Limits of Recursively Self-Improving AGI.* Artificial General Intelligence: 8th International Conference, AGI 2015, AGI 2015, Berlin, Germany, July 22-25, 2015, Proceedings, 2015. **9205**: p. 394.
28. Schwarting, M., T. Burton, and R. Yampolskiy. *On the Obfuscation of Image Sensor Fingerprints*. in *Information and Computer Technology (GOCICT), 2015 Annual Global Online Conference on*. 2015. IEEE.
29. Margoni, T., *Artificial Intelligence, Machine learning and EU copyright law: Who owns AI?*, in *CREATe Working Paper*. 2018: Glasgow.