IMPOSSIBILITY OF COLLECTIVE INTELLIGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

This work provides a minimum requirement in terms of intuitive and reasonable axioms under which an empirical risk minimization (ERM) is the only rational learning algorithm when learning in heterogeneous environments. We provide an axiomatization of any learning rule in terms of choice correspondences over a hypothesis space and seemingly primitive properties. Then, we show that the only feasible algorithm compatible with these properties is the standard ERM that learns arbitrarily from a single environment. This impossibility result implies that Collective Intelligence (CI), the ability of algorithms to successfully learn across heterogeneous environments, cannot be achieved without sacrificing at least one of these basic properties. More importantly, this work reveals an incomparability of performance metrics across environments as one of the fundamental limits in critical areas of machine learning such as out-of-distribution generalization, federated learning, algorithmic fairness, and multi-modal learning.

1 INTRODUCTION

In this work, we study the aggregation rule

$$F : (r_1, r_2, \dots, r_n) \mapsto (\mathscr{H}, \mathscr{B}, \mathbb{A}_{\mathbf{r}}), \quad n \in \mathbb{N}, \tag{1}$$

where $\mathbf{r} := (r_1, \dots, r_n)$ represents a *risk profile* and $(\mathscr{H}, \mathscr{B}, \mathbb{A}_{\mathbf{r}})$ is called a *learning structure*. The risk profile is an n-tuple of risk functionals evaluated on $h \in \mathscr{H}$ which provides the performance measures for each hypothesis across n heterogeneous environments. The learning structure is composed of a hypothesis space $\mathscr{H}$, a collection $\mathscr{B}$ of non-empty subsets of $\mathscr{H}$, and a learning rule $\mathbb{A}_{\mathbf{r}}$ which is a choice correspondence such that $\mathbb{A}_{\mathbf{r}}(\mathcal{H}) \subseteq \mathcal{H}$ for all hypothesis class $\mathcal{H} \in \mathscr{B}$.[1] In words, $\mathbb{A}_{\mathbf{r}}$ specifies for any feasible non-empty hypothesis class $\mathcal{H} \in \mathscr{B}$ a non-empty subset $\mathbb{A}(\mathcal{H}) \subseteq \mathcal{H}$ that consists of "optimal" solutions to the problem. All in all, the learning structure is an abstraction of any conceivable learning algorithm, while equation 1 models the task of designing a learning algorithm given data from n heterogeneous environments.

Equation 1 characterizes modern machine learning problems including out-of-distribution (OOD) generalization (aka domain generalization) [6, 16, 2], federated learning (FL) [13, 15, 14, 11], algorithmic fairness [8, 5, 26, 12, 24], and multi-modal learning [17, 19, 4]. In domain generalization, the risk profile $\mathbf{r}$ corresponds to empirical risk functionals across n domains and the goal is to design an algorithm $\mathbb{A}_{\mathbf{r}}$ that chooses the hypothesis from $\mathscr{H}$ that generalizes well to the previously unseen domains. In federated learning, the risk profile encodes the *local* empirical risk functionals based on private data sets residing at n local nodes, e.g., mobile phones and hospitals. We then aim at designing an efficient algorithm that can learn across these nodes without exchanging private information. In algorithmic fairness, the risk profile encodes the model error rates across demographic groups and we want a learning algorithm that cannot only achieve high overall accuracy but also satisfies some fairness constraints. Lastly, the risk profile in multi-modal learning is composed of score functions, e.g., negative log-likelihood functions, of different data modalities such as images, texts, and audio. An effective algorithm should be able to leverage information across these modalities.

We seek to understand fundamentally the limit of the above problems by establishing a minimum requirement in terms of intuitive and reasonable axioms, namely, Pareto Optimality (PO), Independence of Irrelevant Hypotheses (IIH), and Invariance Restriction (IR), under which the empirical risk

[1] We assume throughout that $\emptyset \notin \mathscr{B}$ and that $\mathbb{A}_{\mathbf{r}}(\mathcal{H}) \neq \emptyset$ for all hypothesis class $\mathcal{H} \in \mathscr{B}$.

minimization (ERM)[2] is the only algorithm possible (cf. Theorem 1 and Corollary 1). Our impossibility result is based on the novel characterization of learning algorithms as choice correspondences on a hypothesis space (cf. Definition 1). This result suggests that Collective Intelligence (CI), the ability of algorithms to learn across heterogeneous environments, cannot be achieved without sacrificing at least one of these fundamental properties.

2 Algorithmic Choice

Let $\mathscr{H}$ be a set of hypotheses from which the solutions can be chosen. A learning problem involves choosing the best solutions from $\mathscr{H}$ via a learning rule. A learning rule $\mathbb{A}$ specifies for any feasible non-empty subset $\mathcal{H} \subset \mathscr{H}$ a non-empty subset $\mathbb{A}(\mathcal{H}) \subseteq \mathcal{H}$.[3] In what follows, the set of natural numbers is denoted by $\mathbb{N}$, and for $n \in \mathbb{N}$, $[n] := \{1, 2, \dots, n\}$.

Definition 1 (Learning structure)**.** *A learning structure is a triple $(\mathscr{H}, \mathscr{B}, \mathbb{A})$ where $\mathscr{H}$ is a hypothesis space, $\mathscr{B}$ is a collection of non-empty subsets of $\mathscr{H}$, and $\mathbb{A}$ is a learning algorithm represented as a choice correspondence from $\mathscr{B}$ to $2^{\mathscr{H}}$ such that $\mathbb{A}(\mathcal{H}) \subseteq \mathcal{H}$ for all $\mathcal{H} \in \mathscr{B}$.*

The learning structure is a blueprint of any *conceivable* learning rule. For each feasible hypothesis class $\mathcal{H} \in \mathscr{B}$, $\mathbb{A}(\mathcal{H})$ outputs the solutions that would be obtained by executing the learning rule. It is instructive to note that $(\mathscr{H}, \mathscr{B}, \mathbb{A})$ can itself be an outcome of the extensive choice, e.g., model selection procedures. For brevity, we omit the detail of model selection as it does not alter our main result. What matters here is that for each $(\mathscr{H}, \mathscr{B}, \mathbb{A})$, $\mathbb{A}$ remains the same for *all* $\mathcal{H}$ in $\mathscr{B}$.[4]

2.1 Internal Consistency

Since not all learning structures implied by Definition 1 exhibits desirable properties, we argue that any *reasonable* learning algorithm must satisfy an *internal consistency* property.

Definition 2 (Internal consistency)**.** *Given a learning structure ($\mathscr{H}$,$\mathscr{B}$,$\mathbb{A}$), the algorithm $\mathbb{A}$ is internally consistent if for $\mathcal{F}, \mathcal{G} \in \mathscr{B}$, the following two conditions hold. (i) If $h \in \mathbb{A}(\mathcal{F})$ and $h \in \mathcal{G} \subseteq \mathcal{F}$, then $h \in \mathbb{A}(\mathcal{G})$. (ii) If $h \in \mathcal{G}$ and, for $\mathcal{G} \subseteq \mathcal{F}$, $h \in \mathbb{A}(\mathcal{F})$, then $\mathbb{A}(\mathcal{G}) \subseteq \mathbb{A}(\mathcal{F})$.*

These two conditions, known as Property α [7][20, Ch. 1*] and Property β [21, pp. 320], are together one of the interpretations of the rationality of choice in mainstream economic theory. In this context, the former says that any hypothesis h that is chosen from $\mathcal{F}$ must also be chosen from $\mathcal{G}$, if it is a contraction of $\mathcal{F}$ that also contains h. For example, suppose that $\mathcal{F}$ is composed of all polynomials of degree smaller than p and $\mathcal{G}$ consists of all polynomials of degree smaller than q where $q \leq p$, i.e., $\mathcal{G} \subseteq \mathcal{F}$. Then, if $\mathbb{A}$ chooses the polynomial of degree $t < q \leq p$ from $\mathcal{F}$ as a solution, it must also be chosen again from $\mathcal{G}$. The latter, albeit less intuitive, can be interpreted as follows. If there is the hypothesis h that is chosen from $\mathcal{G}$ and subsequently from $\mathcal{F}$, which is an expansion of $\mathcal{G}$, then all other hypotheses that are considered *equally good*[5] to h in $\mathcal{G}$ must also be chosen from $\mathcal{F}$. For instance, let $f, g \in \mathcal{G}$ be polynomials of degree t with different values of coefficients. If $\mathbb{A}$ chooses both f and g from $\mathcal{G}$, and f from $\mathcal{F}$, then it must also choose g from $\mathcal{F}$. That is, it characterizes the behavior of the algorithm under the expansion of the hypothesis class.

The following two results follow immediately from the internal consistency property.

Proposition 1. *Let $\mathbb{A}_r$ be a risk minimization (RM) algorithm, i.e., $\mathbb{A}_r(\mathcal{H}) = \{h \in \mathcal{H} : r(h) \leq r(g), \forall g \in \mathcal{H}\}$ for some real-valued function $r : \mathscr{H} \to \mathbb{R}$. Then, $\mathbb{A}_r$ satisfies internal consistency.*

A majority of learning algorithms including empirical risk minimization (ERM) [25], structural risk minimization (SRM) [23], and invariant risk minimization (IRM) [2, 1] fall into this category.

[2] In this paper, ERM loosely refers to algorithms that minimize a risk function associated with a single homogeneous environment.

[3] Further restriction can be made so that any $\mathbb{A}(\mathcal{H})$ must be a unit set, with only one hypothesis chosen from $\mathcal{H}$, but we stick to the more general setting throughout.

[4] In practice, both $\mathscr{B}$ and $\mathbb{A}$ can depend on some hyperparameters, e.g., kernel choice, neural architecture, and regularization parameters, that render $\mathcal{H} \in \mathscr{B}$ and $\mathbb{A}$ interdependent.

[5] In this work, h and g are considered by $\mathbb{A}$ as *equally good* if and only if $h, g \in \mathbb{A}(\mathcal{H})$.

Proposition 2. *Suppose that $\mathbb{A}$ satisfies the internal consistency property. For every pair $f, g \in \mathscr{H}$, let $f \succeq_{\mathbb{A}} g$ if $f \in \mathbb{A}(\{f, g\})$. Then, the binary relation $\succeq_{\mathbb{A}}$ is complete and transitive.*[6] *Moreover, for every $\mathcal{H} \in \mathscr{B}$, $\mathbb{A}(\mathcal{H}) = \mathbb{A}_{\succeq_{\mathbb{A}}}(\mathcal{H}) := \{h \in \mathcal{H} : h \succeq_{\mathbb{A}} g, \forall g \in \mathcal{H}\}$.*

The binary relation $\succeq_{\mathbb{A}}$ is known as a *revealed preference* of $\mathbb{A}$ and this proposition implies that as long as $\mathbb{A}$ satisfies internal consistency, $\mathbb{A}(\mathcal{H})$ for any $\mathcal{H} \in \mathscr{B}$ coincides with those obtained from the learning algorith $\mathbb{A}_{\succeq_{\mathbb{A}}}$ defined in terms of preferences that are revealed by $\mathbb{A}$ operated on one- and two-element subsets of hypotheses; see, e.g., Sen [22] for details on revealed preference.

More importantly, Proposition 2 highlights the essence of the internal consistency property as inconsistent algorithms can exhibit incomplete or intransitive preference over the hypothesis space.

3 The Impossibility Result

What we seek to understand is whether one can take the risk profile $\mathbf{r} = (r_1, \ldots, r_n)$ computed from data across n heterogeneous environments and create learning algorithms $\mathbb{A}_{\mathbf{r}}$ that are internally consistent; see equation 1. To answer this question, we argue that any of such mechanisms should satisfy the following four primitive properties.

1. **Pareto Optimality (PO):** For all $f, g \in \mathscr{H}$, $r_i(f) < r_i(g)$ for all $i \in [n]$ implies that $\{f\} = \mathbb{A}_{\mathbf{r}}(\{f, g\})$.
2. **Independence of Irrelevant Hypotheses (IIH):** For any pair of risk profiles $\mathbf{r} = (r_1, \ldots, r_n)$, $\mathbf{r}' = (r'_1, \ldots, r'_n)$ and any pair of hypotheses $f, g \in \mathscr{H}$ such that $r_i(f) = r'_i(f)$ and $r_i(g) = r'_i(g)$, $f \in \mathbb{A}_{\mathbf{r}}(\{f, g\})$ if and only if $f \in \mathbb{A}_{\mathbf{r}'}(\{f, g\})$.
3. **Invariance Restriction (IR):** For any pair of risk profiles $\mathbf{r} = (r_1, \ldots, r_n)$, $\mathbf{r}' = (r'_1, \ldots, r'_n)$ for which there exists $(a_1, \ldots, a_n) \in \mathbb{R}^n$ and $(b_1, \ldots, b_n) \in \mathbb{R}^n_+$ such that $r_i(h) = a_i + b_i r'_i(h)$ for all $i \in [n]$, $\mathbb{A}_{\mathbf{r}}(\mathcal{H}) = \mathbb{A}_{\mathbf{r}'}(\mathcal{H})$ for all $\mathcal{H} \in \mathscr{B}$.
4. **Collective Intelligence (CI):** There exists no $i \in [n]$ such that for all pair $f, g \in \mathscr{H}$ and for all $\mathbf{r}$ in the domain of F, $r_i(f) < r_i(g)$ implies that $\{f\} = \mathbb{A}_{\mathbf{r}}(\{f, g\})$.

Firstly, PO simply requires that the algorithm chooses f over g if f is strictly better than g in *all* environments. It is hard to argue against PO as a desirable property. If f is unanimously superior to g, there is no reason for $\mathbb{A}_{\mathbf{r}}$ to choose otherwise. Secondly, IIH requires that when choosing between two hypotheses, the algorithm should not rely on any other information apart from the risk values of these two hypotheses across environments. Thirdly, IR demands the algorithm to be invariant to any transformation of the risk profile that is *informationally identical* to the original one. Most learning algorithms in a single environment are invariant to the strictly increasing transformation of the risk functional and IR simply generalizes this property to a set of n-tuples of risk functionals. Last but not least, CI demands that the algorithm leverages information from multiple environments.

Nevertheless, we show that for a finite number of two or more environments, no aggregation rule F can satisfy these four properties simultaneously.

Theorem 1. *For a finite number of two or more environments and at least three distinct hypotheses, there exists no aggregation rule F that produces an internally consistent structure $(\mathscr{H}, \mathscr{B}, \mathbb{A})$ and satisfies PO, IIH, IR, and CI simultaneously.*

The following corollary follows immediately from Theorem 1.

Corollary 1. *For a finite number of two or more environments and at least three distinct hypotheses, a unique algorithm that is internally consistent and is compatible with PO, IIH, and IR simultaneously is of the form: $\mathbb{A}_{\mathbf{r}}(\mathcal{H}) = \{h \in \mathcal{H} : r_i(h) \leq r_i(g), \forall g \in \mathcal{H}\}$, $\mathcal{H} \in \mathscr{B}$ for some $i \in [n]$.*

To summarize, if we treat PO, IIH, and IR as primitive properties that any learning algorithms in heterogeneous environments should satisfy, the only possibility that can come out of equation 1 is the algorithm that does not leverage information across environments. To move forward, we must therefore give up at least one of these properties.

[6] The binary relation $\succeq$ on $\mathscr{H}$ is *complete* if for every pair f and g from $\mathscr{H}$, either $f \succeq g$ or $g \succeq f$ (or both). It is *transitive* if $f \succeq g$ and $g \succeq h$ implies $f \succeq h$.

It is instructive to understand intuitively how these assumptions respectively yield Theorem 1 and Corollary 1. First, we consider all conceivable algorithms defined on $\mathscr{H}$. The undesirable ones are then eliminated by demanding that they satisfy some essential properties. By successively imposing internal consistency, PO, IIH, and IR, the algorithm in Corollary 1 remains the only possibility. Adding CI eliminates this only possibility, giving rise to the impossibility result in Theorem 1. The detailed proofs of these two results can be found in Appendix A.

4 Discussion

To sum up, our results imply that it is impossible to achieve collective intelligence without sacrificing at least one of the above primitive properties. The first way out is to remove internal consistency, which implies that the revealed preferences of learning algorithms may not be complete and transitive (see Proposition 2). The second way out is to restrict the domain of F. In fact, existing assumptions such as task relatedness in multi-task learning and (causal) invariance in domain generalization are domain restrictions in disguise. The drawback however is that it may be non-trivial to check in practice. Pareto optimality [18] is a very simple and highly appealing criterion of comparison of hypotheses in the multi-objective setting which generalizes the notion of "minimum risk". Therefore, the consequence of dropping PO as a necessary criterion must be immense. It also implies that the information contained in the risk profile is not sufficient for learning and that some "irrelevant" information must be used. Hence, a violation of PO requires some caution. Dropping IIH also poses a similar concern regarding the use of irrelevant information.

What about the IR property? To answer this question, consider two hypotheses h and h' from $\mathscr{H}$ and risk functionals r_i and r_j from the same risk profile. Suppose that $r_i(h') - r_i(h) = r_j(h') - r_j(h) < 0$, i.e., h' is better than h in both environments i and j and by the same margin. Then, whether one can drop the IR condition depends on whether or not one possesses enough information to claim that "*h' leads to the same improvement over h in environment i as it does in environment j*". For instance, will the COVID-19 AI diagnosis system h' lead to the same improvement over the old system h for Johns Hopkins Hospital in Baltimore as it does for Siriraj Hospital in Thailand? Will the new autocorrection system h' lead to the same improvement over the existing one h in terms of satisfaction for users in Japan as it does for users in South Africa? and so on. When the answer to these questions is yes, we can drop the IR property. In reality, however, such a detailed comparison between environments may not be possible, a shortcoming we call *informational incomparability*.[7]

Informational incomparability What makes heterogeneous learning particularly challenging is the fact that it requires a meaningful inter-environment comparison. In retrospect, while risk functions and the like have been commonly used as a proxy for the performance of algorithmic models in the real world and as a basis to compare them, the inter-environment comparison of risk functions is harder than one might expect. The first challenge is a physical one. In federated learning, for example, it is physically impossible to share all the data across a massive network of mobile devices. Matters pertaining to privacy, security, and access rights will also limit data sharing across environments. The second challenge is a cultural one. An algorithmic model might have varied effects across different demographic groups simply because of cultural differences. It is difficult to believe that there will be means for us to tell all the differences between any two cultures. In algorithmic fairness, for example, there can be a mismatch between measurement modeling and operationalization of social constructs, i.e., abstractions that describe phenomena of theoretical interest such as socioeconomic status and risk of recidivism, which makes it difficult to meaningfully compare different operationalizations [9]. The third challenge is a subjective matter. In multi-modal learning, the relationship between modalities is often open-ended or subjective [19, 4]. Language is often seen as symbolic, but audio and visual data are represented as signals. Moreover, likelihood functions defined on different data types are generally incomparable [10]. Last but not least, the obstacle can simply be a legal one. To protect its people, a government might regulate what kind of and to what extent information can be shared. Well-known examples of this attempt are the EU's General Data Protection Regulation (GDPR) and its upcoming AI Act.[8]

[7] In case of cardinal utility functions, this problem is known in economics as an interpersonal incomparability of utility; see, e.g., Sen [21, Ch. 7]. There has been a long debate on whether one can make a meaningful comparison of the welfare of different individuals.

[8] `https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence`

REFERENCES

[1] K. Ahuja, K. Shanmugam, K. Varshney, and A. Dhurandhar. Invariant risk minimization games. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 145–155. PMLR, 2020.

[2] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *ArXiv*, abs/1907.02893, 2019.

[3] K. J. Arrow. A difficulty in the concept of social welfare. *Journal of Political Economy*, 58(4): 328–346, 1950.

[4] T. Baltrusaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443, 2019.

[5] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[6] G. Blanchard, G. Lee, and C. Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2178–2186. 2011.

[7] H. Chernoff. Rational selection of decision functions. *Econometrica*, 22(4):422–443, 1954.

[8] M. Hardt, E. Price, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, pages 3323–3331. Curran Associates, Inc., 2016.

[9] A. Z. Jacobs and H. Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 375–385. Association for Computing Machinery, 2021.

[10] A. Javaloy, M. Meghdadi, and I. Valera. Mitigating modality collapse in multimodal VAEs via impartial optimization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9938–9964. PMLR, 2022.

[11] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. A. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, S. E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konečný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, M. Raykova, H. Qi, D. Ramage, R. Raskar, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, and S. Zhao. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14:1–210, 2021.

[12] N. Kilbertus, M. G. Rodriguez, B. Schölkopf, K. Muandet, and I. Valera. Fair decisions despite imperfect predictions. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 277–287. PMLR, 2020.

[13] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *CoRR*, abs/1610.02527, 2016. URL `http://arxiv.org/abs/1610.02527`.

[14] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[15] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 2017.

[16] K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 10–18, 2013.

[17] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. pages 689–696. Omnipress, 2011.

[18] V. Pareto. The new theories of economics. *Journal of Political Economy*, 5, 1897.

[19] D. Ramachandram and G. W. Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017.

[20] A. Sen. *Collective Choice and Social Welfare*. Mathematical economics texts. Holden-Day, 1970.

[21] A. Sen. *Collective Choice and Social Welfare: An Expanded Edition*. Harvard University Press, 2017.

[22] A. K. Sen. Choice functions and revealed preference. *The Review of Economic Studies*, 38(3): 307–317, 1971.

[23] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. A framework for structural risk minimisation. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, pages 68–76. Association for Computing Machinery, 1996.

[24] B. van Giffen, D. Herhausen, and T. Fahse. Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research*, 144:93–106, 2022.

[25] V. Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1991.

[26] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.

A PROOFS

A.1 PROOF OF PROPOSITION 1

Proof. (α Property): For any $\mathcal{F}, \mathcal{G} \in \mathscr{B}$ such that $\mathcal{G} \subseteq \mathcal{F}$, let $h \in \mathbb{A}_r(\mathcal{F})$. Thus, $r(h) \leq r(g)$ for all $g \in \mathcal{F}$. Assume that $h \in \mathcal{G}$ but $h \notin \mathbb{A}_r(\mathcal{G})$. This implies that there exists another hypothesis $f \in \mathcal{G}$ for which $r(f) < r(h)$. However, since f is also in $\mathcal{F}$, it contradicts with $r(h) \leq r(g)$ for all $g \in \mathcal{F}$. Hence, h must also be in $\mathbb{A}_r(\mathcal{G})$. ($\beta$ Property): For any $h, g \in \mathcal{G}$, let $h, g \in \mathbb{A}_r(\mathcal{G})$. Assume that $h \in \mathbb{A}_r(\mathcal{F})$ but $g \notin \mathbb{A}_r(\mathcal{F})$. This implies that $r(h) \leq r(f)$ for all $f \in \mathcal{F}$ and $r(g) > r(h)$, which contradicts the fact that $g \in \mathbb{A}_r(\mathcal{G})$. Hence, g must also be in $\mathbb{A}_r(\mathcal{F})$. This implies that $\mathbb{A}_r(\mathcal{G}) \subseteq \mathbb{A}_r(\mathcal{F})$. □

A.2 PROOF OF PROPOSITION 2

Proof. Since $\mathbb{A}(\{f,g\}) \neq \emptyset$, $\succeq_{\mathbb{A}}$ is complete. If $f \succeq_{\mathbb{A}} g$ and $g \succeq_{\mathbb{A}} h$, then $f \in \mathbb{A}(\{f,g\})$ and $g \in \mathbb{A}(\{g,h\})$. By β Property, if $g \in \mathbb{A}(\{f,g,h\})$, then $f \in \mathbb{A}(\{f,g,h\})$. Also, if $h \in \mathbb{A}(\{f,g,h\})$, then $g \in \mathbb{A}(\{f,g,h\})$. Hence, we have $f \in \mathbb{A}(\{f,g,h\})$ in any case. By α Property, $f \in \mathbb{A}(\{f,h\})$ and $f \succeq_{\mathbb{A}} h$, which shows that $\succeq_{\mathbb{A}}$ is transitive. Next, we show that $\mathbb{A}(\mathcal{H}) = \mathbb{A}_{\succeq_{\mathbb{A}}}(\mathcal{H})$ for every $\mathcal{H} \in \mathscr{B}$. Assume that $f \in \mathbb{A}(\mathcal{H})$. By α Property, we have for every $g \in \mathcal{H}$ that $f \in \mathbb{A}(\{f,g\})$. This implies that $f \succeq_{\mathbb{A}} g$ and thus $f \in \mathbb{A}_{\succeq_{\mathbb{A}}}(\mathcal{H})$. Now, let us assume that $f \neq g$, $f \in \mathbb{A}_{\succeq_{\mathbb{A}}}(\mathcal{H})$, and $g \in \mathbb{A}(\mathcal{H})$. Then, $f \in \mathbb{A}(\{f,g\})$ and by β Property, $f \in \mathbb{A}(\mathcal{H})$, which completes the proof. □

A.3 PROOF OF THEOREM 1 AND COROLLARY 1

This section provides a proof of Theorem 1, which relies heavily on the insights from the original proof of Arrow's General Possibility Theorem [3] and its simplification in Sen [21, pp. 286].

Let $\mathcal{E}$ be a set of environments. Crucial to the proof is the idea of a set $\mathcal{E}$ being "decisive".

Definition 3 (Decisiveness)**.** *A set of environments $\mathcal{E}$ is said to be* locally decisive *over a pair of hypotheses f, g if $r_e(f) < r_e(g)$ for all $e \in \mathcal{E}$ implies that $\{f\} = \mathbb{A}(\{f,g\})$. It is said to be* globally decisive *if it is locally decisive over every pairs of hypotheses.*

The following two intermediate results provide basic properties of decisive set of environments $\mathcal{E}$.

Lemma 1. *If a set of environments $\mathcal{E}$ is decisive over any pair $\{f,g\}$, then $\mathcal{E}$ is globally decisive.*

Proof. Let $\{p,q\}$ be any other pair of hypotheses that is different from $\{f,g\}$. Assume that in every environment e in $\mathcal{E}$, $r_e(p) < r_e(f)$, $r_e(f) < r_e(g)$, and $r_e(g) < r_e(q)$. For all other environments e' not in $\mathcal{E}$, we assume that $r_{e'}(p) < r_{e'}(f)$ and $r_{e'}(g) < r_{e'}(q)$ and leave the remaining relations unspecified. By PO condition, $\{p\} = \mathbb{A}(\{p,f\})$ and $\{g\} = \mathbb{A}(\{q,g\})$. By the decisiveness of $\mathcal{E}$ over $\{f,g\}$, we have $\{f\} = \mathbb{A}(\{f,g\})$. Then, it follows from the transitivity implied by Proposition 2 that $\{p\} = \mathbb{A}(\{p,q\})$. By IIH condition, this must be related only to the relation between p and q. Since we have only specified information in $\mathcal{E}$, $\mathcal{E}$ must be decisive over $\{p,q\}$ and for all other pairs. Hence, $\mathcal{E}$ is globally decisive. □

Lemma 2. *If a set of environments $\mathcal{E}$ consists of more than one element and is decisive, then some proper subset of $\mathcal{E}$ is also decisive.*

Proof. Since there are at least two environments, we can partition $\mathcal{E}$ into two subsets $\mathcal{E}_1$ and $\mathcal{E}_2$. Assume that $r_e(f) < r_e(g)$ and $r_e(f) < r_e(h)$ in every environment $e \in \mathcal{E}_1$ with the relation between g and h unspecified. Let $r_{e'}(f) < r_{e'}(g)$ and $r_{e'}(h) < r_{e'}(g)$ in every environment $e' \in \mathcal{E}_2$. By the decisiveness of $\mathcal{E}$, we have $\{f\} = \mathbb{A}(\{f,g\})$. Now, if h is at least as good as f for some environments over $\{h,f\}$, then we must have $\{h\} = \mathbb{A}(\{h,g\})$ for that configuration. Since we do not specify relation over $\{g,h\}$ other than those in $\mathcal{E}_2$, and $r_{e'}(h) < r_{e'}(g)$ in $\mathcal{E}_2$, $\mathcal{E}_2$ is decisive over $\{g,h\}$. By Lemma 1, $\mathcal{E}_2$ must be globally decisive. That is, some proper subset of $\mathcal{E}$ is indeed decisive for that particular case. To avoid this possibility, we must remove the assumption that h is at least as good as f. But then f must be better than h. However, no environment has this relation over $\{f,h\}$ other than those in $\mathcal{E}_1$ where f is better than h. Clearly, $\mathcal{E}_1$ is decisive over $\{f,h\}$. Thus, by Lemma 1, $\mathcal{E}_1$ is globally decisive. So either E_1 or E_2 must be decisive. □

An important observation from the proofs of Lemma 1 and Lemma 2 is that we rely only on the relative rankings of hypotheses. We are now in a position to prove Theorem 1.

Proof of Theorem 1. Consider any two risk profiles $\mathbf{r} = (r_1, \dots, r_n)$ and $\mathbf{r}^* = (r_1^*, \dots, r_n^*)$ such that for any f, g and for all $i \in [n]$, $r_i(f) < r_i(g) \Leftrightarrow r_i^*(f) < r_i^*(g)$. For every pair $\{f, g\}$, there exists a positive affine transformation $\{\varphi_i\}$ applied to $\mathbf{r}^*$ such that

$$r_i'(f) = \varphi_i(r_i^*(f)) = r_i(f) \quad \text{and} \quad r_i'(g) = \varphi_i(r_i^*(g)) = r_i(g) \text{ for all } i \in [n].$$

By IIH condition, $\{f\} = \mathbb{A}_{\mathbf{r}}(\{f, g\})$ if and only if $\{f\} = \mathbb{A}_{\mathbf{r}'}(\{f, g\})$ and by IR condition, $\{f\} = \mathbb{A}_{\mathbf{r}'}(\{f, g\})$ if and only if $\{f\} = \mathbb{A}_{\mathbf{r}^*}(\{f, g\})$. Since this holds pair by pair, clearly $\mathbb{A}_{\mathbf{r}}(\mathcal{H}) = \mathbb{A}_{\mathbf{r}'}(\mathcal{H})$ for all $\mathcal{H} \in \mathscr{B}$. As a result, what matters when comparing any two hypotheses is the relative ranking between them. Next, by PO condition, the set of all environments $\mathcal{E}$ is decisive. By Lemma 2, some proper subset of $\mathcal{E}$ must also be decisive. Given that smaller subset of environments, some proper subset of it must also be decisive, and so on. Since the number of environments is finite, the set will eventually contain just a single environment that is decisive. However, this violates CI condition, resulting in the impossibility. □

Corollary 1 follows by omitting the last step in the proof of Theorem 1.