# Article information:

[1910.06358] Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability
<https://arxiv.org/abs/1910.06358>

# Article summary:

1. Shapley values are a widely used framework for explaining AI systems, but they ignore causal structure in the data.

2. Asymmetric Shapley values (ASVs) are a less restrictive framework that can incorporate any known causal structure in the data.

3. ASVs can improve model explanations, test for unfair discrimination, enable sequential explanations in time-series models, and support feature-selection studies without retraining the model.

# Article rating:

Appears moderately imbalanced: The article provides some useful information, but is missing several important points or pieces of evidence that would be required to present the discussed topics in a balanced and reliable way. You are encouraged to seek a more balanced perspective on the presented issues by exploring the provided research topics and looking at different information sources.

# Article analysis:

作为一篇关于AI可解释性的论文，本文提出了一种新的框架——不对称Shapley值（ASVs），旨在将因果结构纳入模型无关的解释中。然而，在对该论文进行批判性分析时，我们发现以下几个问题：

1. 潜在偏见及其来源：本文没有明确说明作者的背景和利益相关方，这可能导致潜在偏见。此外，该论文是由一家私营公司发布的，这也可能影响作者的研究结果。

2. 片面报道：本文只介绍了ASVs的优点，并没有探讨其缺点或局限性。例如，ASVs是否适用于所有类型的数据集？是否存在某些情况下ASVs会产生错误或误导性结果？

3. 无根据的主张：本文声称ASVs可以“提供一个明确的测试来检测模型预测中不公平歧视”，但并没有提供任何证据来支持这个主张。如何定义“不公平歧视”？如何确定测试结果是否准确？

4. 缺失考虑点：本文没有考虑到实际应用中可能出现的风险和挑战。例如，在使用ASVs时如何保护用户隐私？如果ASVs产生与人类直觉相反的结果，如何解释和处理？

5. 所提出主张缺失证据：尽管本文声称ASVs可以改进模型解释，并且可以支持特征选择研究等应用场景，但并没有提供足够的证据来支持这些主张。需要更多实验数据和案例研究来验证这些主张。

6. 未探索反驳：本文没有探讨其他学者对Shapley值框架和其他可解释性方法所提出的批评和反驳意见。这可能导致读者对该框架和方法存在误解或过度依赖。

7. 宣传内容：尽管该论文是由一家私营公司发布的，但它似乎试图推销自己开发的AI可解释性工具。文章中多次强调了该工具与其他竞争产品相比具有优势，并且未能充分揭示其商业目标。

总之，虽然本文提出了一个新颖而有前途的框架——ASVs，但它仍然存在许多问题需要进一步探讨和验证。同时，在阅读此类文章时需要保持警惕，并注意作者可能存在潜在偏见或商业动机。

# Topics for further research:

* Author background and potential bias
* Limitations and drawbacks of ASVs
* Lack of evidence for claims
* Risks and challenges in practical applications
* Insufficient evidence for proposed benefits
* Failure to explore criticisms and counterarguments

# Report location:

<https://www.fullpicture.app/item/6f523bd08af8d9d8189e600c759d1360>