Input-Attribution Methods Cannot Justify AI Decisions

I. The Duty of Justification

> *Someone must have been telling tales about Josef K., for one morning, without having done anything wrong, he was arrested.*[1]

In *The Trial*, Josef K. is subjected to indiscernible laws, implicated by nameless accusers, charged for an unknown crime. We feel the injustices committed against K. throughout the story—we feel the wrongfulness of subjecting someone to outcomes for which no explanation is offered. When a student receives a poor grade on an essay, we think they deserve feedback. When an employee faces termination at work, we think their boss owes them a reason. When a defendant faces conviction, they are legally entitled to an articulation of the case against them. There exists a duty of justification basic to interpersonal life, and we feel its absence immediately.

This feeling has served as the normative foundation behind many of our institutional and legislative processes. The notion of procedural due process traces back to the Magna Carta and is a cornerstone of American constitutional law.[2,3] The Equal Credit Opportunity Act requires lenders to provide the specific reasons for denial.[4] GDPR Article 22 grants individuals a right to information about the logic involved in automated decisions to which they are subject.[5] Across society, we have embedded this duty of justification: that the morality of an action is grounded not just in its consequences, but in the reasoning behind it.

In *What We Owe to Each Other*, T. M. Scanlon formalizes this idea as contractualism, which holds that an action is morally permissible only if it is grounded in principles that no one affected by it could reasonably reject.[6] On this view, justification is itself constitutive of moral permissibility, not merely instrumental to it. A seemingly just action whose reasons cannot be made available to those affected by it is, for that reason, unjust—even if the outcomes of the action are themselves good. The duty of justification demands that rationale be available, operative, and answerable to those who must live with its consequences.

---

[1] Kafka, Franz. *The Trial*. Translated by Mike Mitchell, Oxford UP, 2009. Oxford World's Classics.
[2] "Magna Carta, Clause 39 (1215)." *The Magna Carta Project*, U of East Anglia, magnacarta.cmp.uea.ac.uk/read/magna_carta_1215/Clause_39.
[3] U.S. Constitution. Amends. V, XIV.
[4] Equal Credit Opportunity Act, 15 U.S.C. § 1691, 1974.
[5] "General Data Protection Regulation." Regulation (EU) 2016/679, Art. 22, *Official Journal of the European Union*, 2016.
[6] Scanlon, T. M. *What We Owe to Each Other*. Belknap Press of Harvard UP, 1998.

AI systems are increasingly making consequential decisions in domains across hiring, lending, medical triage, criminal justice. The moral permissibility of these decisions, however, does not rest just on the valence of outcomes, but on whether the reasoning behind them can be justified. But our current methods of explainability in AI do not, and in principle cannot, deliver justification of the kind contractualism requires. I therefore argue that the employment of AI in domains where individuals are owed reasons for the decisions made about them is presently not morally permissible. More specifically, my argument is as follows:

> **Premise 1.** An action or decision regarding an individual is morally permissible *only if* it satisfies a duty of justification.
> **Premise 2.** Current explainability methods in AI, namely input-attribution methods such as LIME and SHAP, do not and in principle cannot satisfy a duty of justification.
> **Conclusion.** Therefore, input-attribution methods are insufficient grounds for the moral permissibility of AI-mediated decisions.

Beyond this, I offer a recommendation that computational-decomposition methods such as mechanistic interpretability, while nascent, represent our most promising path to explanations that do satisfy the duty of justification.

II. A Taxonomy of Explanations

Satisfying the duty of justification requires explanations of a particular kind. Specifically, it requires that three conditions be met: (1) the *causal condition*, that the explanations be operative in producing the action, (2) the *right-kind condition*, that the operative causes function as reasons, integrated with other internal states and responsive to relevant considerations, and (3) the *reasonable-rejection condition*, that the reasons offered be ones the affected party could not reasonably reject. An explanation that satisfies all three conditions is **justifiable**. It identifies operative causes that function as reasons which the affected party has no reasonable grounds to refuse.

This framework then yields a four-part categorization of explanations:

- **Behavioral** explanations may accurately describe input-output relationships but fail to engage the causal infrastructure underlying the decision. However accurate, they remain characterizations of statistical patterns, not accounts of what actually produced a given output. Behavioral explanations fail the *causal condition*.
- **Mechanistic** explanations correctly identify operative causes but not ones which function as reasons. A person who reflexively hits another in response to a sudden movement is rarely thought liable for the act. There is a direct causal explanation, the firing of their

sympathetic nervous system, but said firing is not a reason in any proper sense. It is a mechanism operating outside the domain of deliberative reasoning, and as such passes the *causal condition* but fails the *right-kind condition*.

- **Unjustifiable** explanations are both causal and of the right kind. Compared to the above, a person who, after a long deliberation, punches someone because they thought the person deserved it has an explanation that is, however unjust, at least of the right kind and may be morally evaluated. Here, the cause functions as a reason. But the reason identified is one the affected party would likely reasonably reject, and so the action fails the *reasonable-rejection condition*.
- **Justifiable** explanations, then, are those which pass all three conditions. They identify the operative causes that function as reasons, reasons which an affected party could not reasonably reject. As the name suggests, justifiable explanations are the only category that meets the requirements of moral permissibility under the duty of justification.

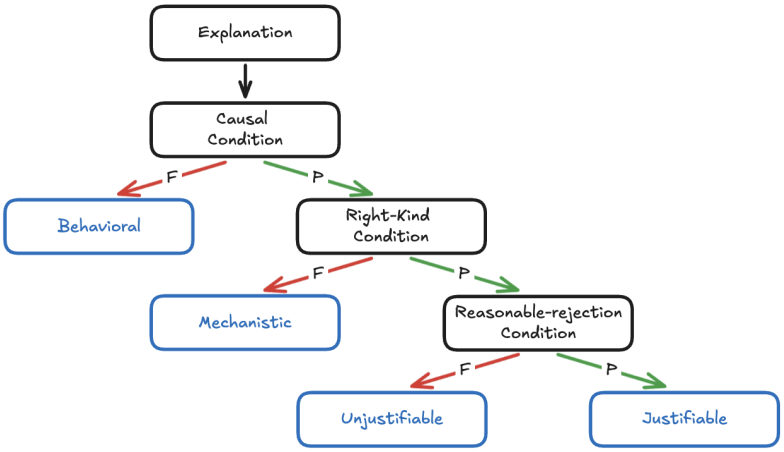

Figure 1: A taxonomy of explanations. Each condition is sequentially gating. Explanations failing a condition are routed to a terminal category.

This framework can be understood as a binary decision tree with three sequential gating conditions, offering a logical schema through which we may categorize explanations. The question for current XAI methods is therefore which conditions they pass and whether they are oriented toward offering explanations which satisfy the duty of justification.

III. Input-Attribution Methods Cannot Satisfy The Duty of Justification

The majority of methods employed today are *input-attribution methods*, which seek to quantify the relevant contributions of input features in black-box neural networks. Two highly influential techniques, Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP), together form the backbone of contemporary XAI. The explanations they

produce, however, remain fundamentally behavioral and therefore cannot in principle satisfy the duty of justification.

LIME trains a simple, interpretable surrogate model that approximates a black-box model's behavior around a given datapoint. It samples perturbations around the target, queries the black-box model on those points, and fits an interpretable model locally. The objective minimizes a weighted squared loss between the outputs of the interpretable and original models, where the weights are functions of proximity to the original datapoint.[7]

SHAP addresses an adjacent question of how to fairly distribute credit among features for a particular prediction, offering a "unified framework for interpreting predictions." Any prediction can be decomposed into a baseline and a sum of feature contributions. Each feature's contribution is its SHAP value, a weighted sum of its marginal contribution averaged across feature coalitions. SHAP is further grounded in Lloyd Shapley's four original axioms of efficiency, symmetry, dummy, and additivity. In their paper, Lundberg and Lee "introduce the perspective of viewing *any* explanation of a model's prediction as a model itself." Under this perspective, they show that their class of additive feature attribution methods unifies six existing XAI methods, including LIME, and that each of these input-attribution methods shares a common foundation. LIME and SHAP are not just two adjacent explainability methods, but instances of the same kind of method.[8]

Both papers explicitly frame their methods as enabling *trust*, and each method serves to bolster users' confidence in model predictions. Trust, however, is not justification. A user may trust a model's predictions on the basis of consistency, and trust does not require the understanding of predictions in terms of causal or operative rationale. Trust in a model is no trivial matter, but the duty of justification requires more than trust.

By both sets of authors' own qualification, LIME and SHAP are fundamentally about constructing and interpreting a separate explanation model. Each produces a model that approximates the behavioral relationship between features and predictions, failing to engage with the inner workings of the original model. Explanations derived from these methods are therefore behavioral, fail the causal condition, and cannot ground the moral permissibility of AI as required under a duty of justification.

---

[7] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2016, pp. 1135–44.

[8] Lundberg, Scott M., and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems 30*, edited by I. Guyon et al., Curran Associates, 2017, pp. 4765–74.

IV. Objections

Two primary objections warrant direct response. First, one might argue that input-attribution methods are sufficient for whatever moral duty we owe affected parties. A SHAP explanation that correctly identifies which features contributed to a given output provides the affected party the information required to determine moral rightness. Demanding causal-mechanistic explanations is a standard that the duty of justification does not require. The reply is that this is precisely what justification requires. A feature can have a high SHAP value merely because of a statistical correlation with an output, even when the actual computation routes through structures that have nothing to do with the feature. This confabulation is a significant weakness of feature attribution methods. They cannot delineate between a plausible account of an output based on feature attributions and one that is causally determinative. Such explanations can be useful in certain instances, but they cannot ground reasonable rejection.

A more practical objection is that requiring causal explanations of this kind imposes substantial costs on AI deployment and blocks systems that could demonstrably help individuals. The refusal of such deployments has its own moral consequences. To this I would first respond that the duty of justification should be applied asymmetrically: most stringently in domains where individuals are owed reasons, domains traditionally subject to procedural justice norms, or instances where such notions are instantiated into the law. In high-stakes domains like lending, hiring, and criminal sentencing, the cost of justification is the price of moral admissibility. Existing institutions and practices already trade efficiency for legitimacy by design, and the same trade-off applies to the AI deployed in these same domains.

V. Mechanistic Interpretability and the Path Forward

The question, then, is what explainability method could satisfy the duty of justification. Mechanistic interpretability offers one potential solution and, despite its nascency, has already begun to show promise. With sparse autoencoders (SAEs), researchers have recovered tens of millions of human-understandable features from activations in frontier models.[9] Circuit tracing has revealed the computational pathways by which models compose features into behavior.[10] Activation patching has isolated specific attention heads responsible for capabilities like indirect object identification and factual recall.[11] Targeted ablation has verified the functional and causal

---

[9] Amodei, Dario. "The Urgency of Interpretability." *Dario Amodei*, Apr. 2025, darioamodei.com/post/the-urgency-of-interpretability.

[10] Ameisen, Emmanuel, et al. "Circuit Tracing: Revealing Computational Graphs in Language Models." *Transformer Circuits Thread*, Anthropic, 2025, transformer-circuits.pub/2025/attribution-graphs/methods.html.

[11] Wang, Kevin, et al. "Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 Small." *arXiv*, 2022, arxiv.org/abs/2211.00593.

role of individual neurons and attention heads.[12] Activation steering has been shown to produce predictable behavioral changes under amplification or suppression.[13]

Mechanistic interpretability's methodology is oriented toward a causal-mechanistic understanding of the internal computations of a model, in a way input-attribution methods are not. Although certain techniques in the field, such as SAEs and circuit tracing, share with input-attribution methods the use of a substitute model designed for interpretability, they differ fundamentally in what that intermediary is meant to capture. Methods like LIME and SHAP remain trapped in questions of input-output correlation, while SAEs and circuits attempt to address the underlying causal structure of the model itself. Mechanistic interpretability has shown it can produce explanations at the *mechanistic* level of the framework, engaging with the internal causal structure. Whether its methods give rise to the *right kind* of explanation, though, which could provide the groundwork for sufficient justification under contractualism, is yet to be determined. The features and circuits identified to date are also a small fraction of what is plausibly there, and whether the identified structures function as reasons rather than mere mechanisms is a question current methods do not settle. That said, the field is growing rapidly and offers our best path forward to XAI that meets the criteria of a duty of justification.

Interpretability research is frequently defended on instrumental grounds: that it aids safety, debugging, and alignment more broadly. This instrumental view, however, misses the real point. Mechanistic interpretability has an intrinsic moral significance and is the most promising existing path toward explanations that could satisfy a duty of justification. Until some method does, AI deployment in domains where reasons are owed remains morally impermissible.

---

[12] Olsson, Catherine, et al. "In-context Learning and Induction Heads." *Transformer Circuits Thread*, Anthropic, 2022, transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

[13] Arditi, Andy, et al. "Refusal in Language Models Is Mediated by a Single Direction." *arXiv*, 2024, arxiv.org/abs/2406.11717.

Works Cited

Kafka, Franz. *The Trial*. Translated by Mike Mitchell, Oxford UP, 2009. Oxford World's Classics.

"Magna Carta, Clause 39 (1215)." *The Magna Carta Project*, U of East Anglia, magnacarta.cmp.uea.ac.uk/read/magna_carta_1215/Clause_39.

U.S. Constitution. Amends. V, XIV.

Equal Credit Opportunity Act, 15 U.S.C. § 1691, 1974.

"General Data Protection Regulation." Regulation (EU) 2016/679, Art. 22, *Official Journal of the European Union*, 2016.

Scanlon, T. M. *What We Owe to Each Other*. Belknap Press of Harvard UP, 1998.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2016, pp. 1135–44.

Lundberg, Scott M., and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems 30*, edited by I. Guyon et al., Curran Associates, 2017, pp. 4765–74.

Amodei, Dario. "The Urgency of Interpretability." *Dario Amodei*, Apr. 2025, darioamodei.com/post/the-urgency-of-interpretability.

Ameisen, Emmanuel, et al. "Circuit Tracing: Revealing Computational Graphs in Language Models." *Transformer Circuits Thread*, Anthropic, 2025, transformer-circuits.pub/2025/attribution-graphs/methods.html.

Wang, Kevin, et al. "Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 Small." *arXiv*, 2022, arxiv.org/abs/2211.00593.

Olsson, Catherine, et al. "In-context Learning and Induction Heads." *Transformer Circuits Thread*, Anthropic, 2022, transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Arditi, Andy, et al. "Refusal in Language Models Is Mediated by a Single Direction." *arXiv*, 2024, arxiv.org/abs/2406.11717.