Interpretability of Black-Box Models

Context, Formalization, and a focus on Cooperative Game Theory

[1]**EDF Lab Chatou** - Département PRISME
[2]**Institut de Mathématiques de Toulouse**
[3]**SINCLAIR AI Lab**

Vendredis de la Data Science
MAIF - Niort.
2nd of December, 2022

Marouane Il Idrissi[123], Nicolas Bousquet[13], Fabrice Gamboa[2], Bertrand Iooss[123], Jean-Michel Loubes[2], Vincent Chabridon[1,3]

Context

*“**ML interpretability methods** to **certify A.I models** on **critical systems**.“*

Context

*“**ML interpretability methods** to **certify A.I models** on **critical systems**.“*

- **ML interpretability**: Tools that allow to *better understand* a (black-box) model.
- **Critical systems**: Hydroelectric dams, wind turbines, stocks, a human body...
- **A.I. model certification**: Guarantee a certain level of confidence on the A.I via (usually) a set of *meaningful* (*rigid*) *rules*.

Context

*"**ML interpretability methods** to **certify A.I models** on **critical systems**."*

- **ML interpretability**: Tools that allow to *better understand* a (black-box) model.
- **Critical systems**: Hydroelectric dams, wind turbines, stocks, a human body...
- **A.I. model certification**: Guarantee a certain level of confidence on the A.I via (usually) a set of *meaningful* (*rigid*) *rules*.

Main goal: Develop *meaningful* tools in order to *better understand* a black-box model, in order to *ensure confidence* in their use for *critical systems*.

Context

*“**ML interpretability methods** to **certify A.I models** on **critical systems**.“*

- **ML interpretability**: Tools that allow to *better understand* a (black-box) model.
- **Critical systems**: Hydroelectric dams, wind turbines, stocks, a human body...
- **A.I. model certification**: Guarantee a certain level of confidence on the A.I via (usually) a set of *meaningful* (*rigid*) *rules*.

Main goal: Develop *meaningful* tools in order to *better understand* a black-box model, in order to *ensure confidence* in their use for *critical systems*.

Who needs convincing: *Safety authorities*, which are not receptive to *empirical results* (i.e., SOTA statements).

Context

*“**ML interpretability methods** to **certify A.I models** on **critical systems**.“*

- **ML interpretability**: Tools that allow to *better understand* a (black-box) model.
- **Critical systems**: Hydroelectric dams, wind turbines, stocks, a human body...
- **A.I. model certification**: Guarantee a certain level of confidence on the A.I via (usually) a set of *meaningful* (*rigid*) *rules*.

Main goal: Develop *meaningful* tools in order to *better understand* a black-box model, in order to *ensure confidence* in their use for *critical systems*.

Who needs convincing: *Safety authorities*, which are not receptive to *empirical results* (i.e., SOTA statements).

Meaningful tools to safety authorities = Theoretical guarantees on the interpretability method.

Sommaire

The need of a formal definition

From the litterature (Barredo Arrieta et al. 2020):

- **Interpretability:** *"[...] The ability to explain or to provide the meaning in understandable terms to a human."* - **Provide meaningful (mathematical) tools.**
- **Explainability:** *"[...] Associated with the notion of explanation as an interface between humans and a decision maker* [...]*."* - **Use of these tools in a human interaction context.**

The need of a formal definition

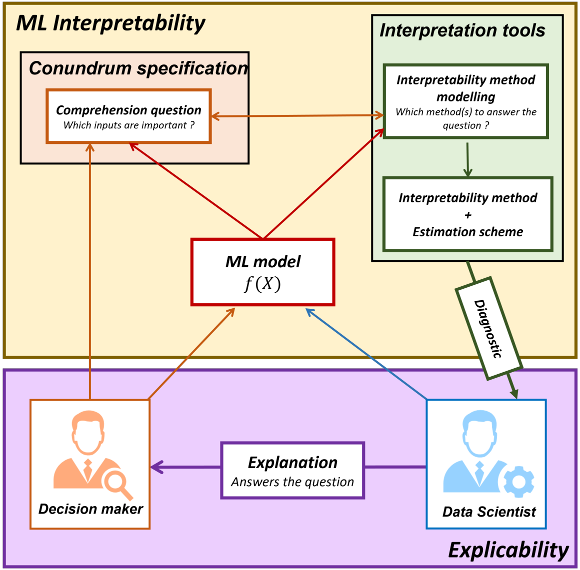
From the litterature (Barredo Arrieta et al. 2020):

- **Interpretability:** *"[...] The ability to explain or to provide the meaning in understandable terms to a human."* - **Provide meaningful (mathematical) tools.**
- **Explainability:** *"[...] Associated with the notion of explanation as an interface between humans and a decision maker* [...]*."* - **Use of these tools in a human interaction context.**

Asher et al. (2022) provides a first step towards a formal model of explanability:

- 2-player explanation game between an **explainee** (decision maker) and an **explainer** (data scientist), to model the interaction:
  - Each player aim at solving a **conundrum** (a question on the black-box model).
  - The explainer provides **diagnostics** (interpretability method) at each turn.
  - The explainee either **accepts or rejects** the explanation (based, e.g., on the theoretical foundations of the interpretability method).

Explanation game



In the rest of this presentation, ML interpretability can then be understood as:

- Development of interpretability methods relative to different conundrums.
- Theoretical understanding of their meaning and their validity.
- Development of efficient estimation schemes in order to compute relevant indices.

Example:

- **Conundrum:** Which feature is responsible for which part of a prediction ?
- **Interpretability method:** SHAP.
- **Estimation scheme:** KernelSHAP.
- **Diagnostic:** Compute the results given by KernelSHAP.
- **Explanation:** Interpret the results and provide an answer to the question.

ML interpretability

In the rest of this presentation, ML interpretability can then be understood as:

- Development of interpretability methods relative to different conundrums.
- Theoretical understanding of their meaning and their validity.
- Development of efficient estimation schemes in order to compute relevant indices.

Example:

- **Conundrum:** Which feature is responsible for which part of a prediction ?
- **Interpretability method:** SHAP.
- **Estimation scheme:** KernelSHAP.
- **Diagnostic:** Compute the results given by KernelSHAP.
- **Explanation:** Interpret the results and provide an answer to the question.

But is SHAP the only interpretability method that answers this question ? Is it the best ? What are its limits ?

Sommaire

Meaningfulness of interpretability methods

In order to decide if an interpretability method is **meaningful**, i.e., if it should be accepted by the decision maker, its **theoretical properties must be understood**.

Example: What does SHAP quantify ?

Hence, one has to focus on the **theoretical quantities**, in order to provide **mathematical proofs** of the behavior of the interpretability method.

Example: When the features are independent, SHAP quantifies...

These properties allow to **accurately interpret the diagnostic**, and further enhances the impact of the **explanation**.

Parenthesis: Statistical inference

Goal: Make a distinction between a **theoretical quantity** and an **estimator**.

One is interested in the average salary in France, denoted μ. If the salary in France is assumed to be a random variable X, one would have that:

$$\mathbb{E}[X] = \mu.$$

This theoretical quantity could be measured exactly if **everyone in France would indicate its salary**. However, this is very costly.

In order to have an **educated guess** on μ, i.e., a "good" approximation, one could randomly ask their salary to n persons living in France, resulting in $x_1, \ldots, x_n$ observations.

Then, an **estimator** of μ would be the empirical mean:

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Parenthesis: Statistical inference

For instance, the theoretical values of **SHAP** can be estimated using several methods: **TreeSHAP, KernelSHAP...**

Here, we are interested in knowing if SHAP is a **meaningful theoretical quantity**, i.e., does it have theoretical properties that would make it the basis of a **good explanation** ?

In order to (partly) answer this question, one can start by taking an interest in the field of **Cooperative Game Theory.**

Sommaire

Cooperative game theory

In a nutshell, cooperative game theory can be summarized as **“the art of cutting a cake”**.

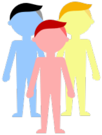
Given a **set of players** $D = \{1, \ldots, d\}$, who produces a **quantity** $v(D)$, how can one allocate shares of $v(D)$ among the d players ?

The **“cake cutting process”** is often described through **axioms** (i.e., desired properties), and results in an **allocation**.

Formally, a cooperative game is denoted (D, v) where D is a **set of players**, and $v : \mathcal{P}(D) \to \mathbb{R}$ is a **value function**, mapping every possible subset of players to a real value.

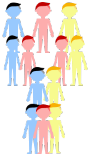
Players

D



Coalitions

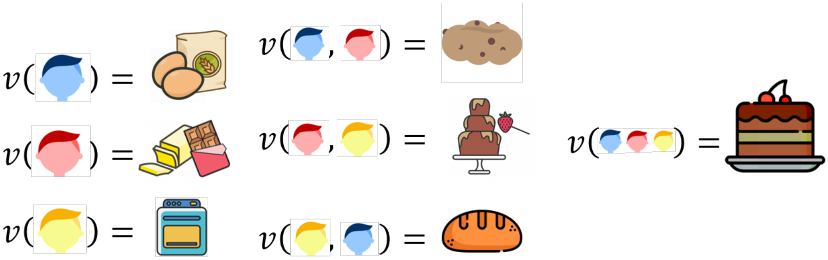
$\mathcal{P}_d$



Cost Function

v

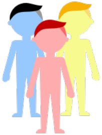
Quantifies the value produced by a coalition



Cooperative game: illustration

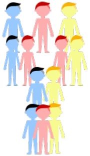
Players

D



Coalitions

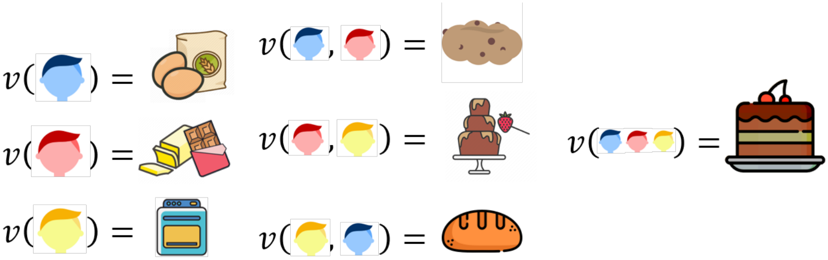
$\mathcal{P}_d$



Cost Function

v

Quantifies the value produced by a coalition



How can we share the cake between the players ?

Allocations

The cake-cutting process must respect one fundamental property:

- The **whole cake** and **nothing but the cake** must be shared (**Efficiency**).

For prediction decomposition (SHAP), we want **every feature contribution** (shares of cake) to **sum-up to the prediction itself** (the cake).

Example of efficient allocations:

- Give the whole cake to the yellow player.
- Divide the cake in three equal parts.
- Give half of the cake to blue, and one fourth to red and yellow.

Allocations

The cake-cutting process must respect one fundamental property:

- The **whole cake** and **nothing but the cake** must be shared (**Efficiency**).

For prediction decomposition (SHAP), we want **every feature contribution** (shares of cake) to **sum-up to the prediction itself** (the cake).

Example of efficient allocations:

- Give the whole cake to the yellow player.
- Divide the cake in three equal parts.
- Give half of the cake to blue, and one fourth to red and yellow.

Can we define efficient allocations in a smarter way ?

The Weber set

Random order model allocations allow to easily define **efficient allocation**.

First, to every cooperative game (D, v) can be associated its **dual** (D, w) where $\forall A \subseteq D$:

$$w(A) = v(D) - v(D \setminus A).$$

The Weber set

Random order model allocations allow to easily define **efficient allocation**.

First, to every cooperative game (D, v) can be associated its **dual** (D, w) where $\forall A \subseteq D$:

$$w(A) = v(D) - v(D \setminus A).$$

Instead to quantifying the **value** $v(A)$ produced by a coalition A, we rather focus on its **cost** $w(A)$ (i.e., how much value did we lose by removing A). w can be interpreted as the **bargaining power** of a coalition.

The Weber set

Random order model allocations allow to easily define **efficient allocation**.

First, to every cooperative game (D, v) can be associated its **dual** (D, w) where $\forall A \subseteq D$:

$$w(A) = v(D) - v(D \setminus A).$$

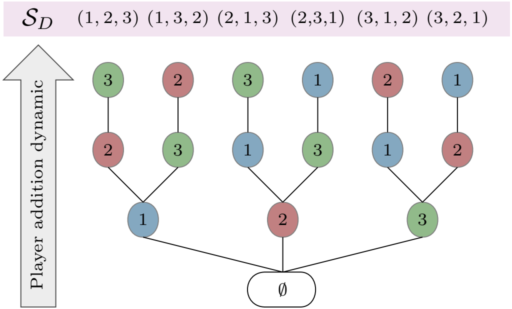
Instead to quantifying the **value** $v(A)$ produced by a coalition A, we rather focus on its **cost** $w(A)$ (i.e., how much value did we lose by removing A). w can be interpreted as the **bargaining power** of a coalition.

A random order allocation is computed as a **weighted average over the permutations of** D: it considers every possible order each player can interact with the others.

Backward-Forward and random order allocations

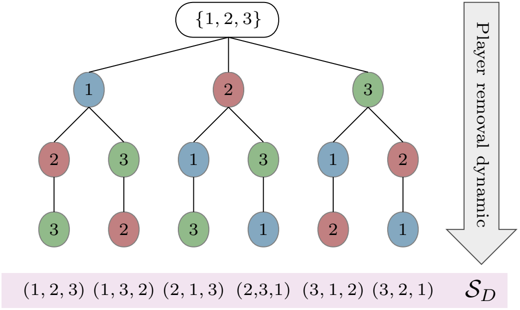
(a.)

Marginal value gain of adding players to $\emptyset$:
$$v\left(C_{\pi(i)}(\pi)\right) - v\left(C_{\pi(i)-1}(\pi)\right)$$



(b.)

Marginal value cost of removing players from $\{1, 2, 3\}$:
$$v\left(D \setminus C_{\pi(i)-1}(\pi)\right) - v\left(D \setminus C_{\pi(i)}(\pi)\right)$$



Random order model allocations

For **any weighting scheme** (probability mass function) put on the permutations of D corresponds an **efficient allocation**.

Random order model allocations

For **any weighting scheme** (probability mass function) put on the permutations of D corresponds an **efficient allocation**.

In this context, **to which weighting scheme corresponds the Shapley values ?**

Random order model allocations

For **any weighting scheme** (probability mass function) put on the permutations of D corresponds an **efficient allocation**.

In this context, **to which weighting scheme corresponds the Shapley values ?**

They correspond to the **specific choice** of assigning the **same weight** to each permutation (i.e., $1/d!$).

The Shapley values as a random order model allocation

The Shapley values of (D, v) and the ones of (D, w) are equal.

How can we understand the Shapley values as a random order model allocation:

- The order to which the players interact **does not matter**.
- The orders to which the players interact are **equally as likely** (this is a huge assumption).
- Allocating **costs** is the same as **allocating profits**.

They can be interpreted as

"[...] *an a priori assessment of the situation, based on either ignorance or disregard of the social organization of the players."* - L. S. Shapley (2016)

Unrelated but relevant parallel question: Uniform prior in the Bayesian context.

Proportional values

Defining an efficient allocation is **as easy as putting a weighting scheme on the permutations of** D.

Are there other known allocations that might be relevant ?

Proportional values

Defining an efficient allocation is **as easy as putting a weighting scheme on the permutations of** D.

Are there other known allocations that might be relevant ?

Yes ! For instance, the **proportional values**. They correspond to a weighting scheme proportional to the **order of importance** of the permutations:

For a profit game (D, v):

- 1 brings the most value ($v(1)$ is bigger than the values of coalitions of 1 player).
- 3 added to 1 brings the most value ($v(\{1, 3\}$ is bigger than the values of the coalitions of 2 players).
- 2 added to $\{1, 3\}$ brings the most value (same thing for coalitions of 3 players).
- ...

Then the order $(1, 3, 2, ...)$ should have a **bigger weight** than the other orders. The same reasoning can be applied to the dual (cost) game (D, w), resulting in a **different allocation**.

Shapley values and proportional values in linear regression

These allocations have been used in statistics for a very long time (without the knowledge that they came from cooperative game theory).

In the context of linear regression, the question is to redistribute the R^2 (the cake) among the players (the covariates).

Shapley values and proportional values in linear regression

These allocations have been used in statistics for a very long time (without the knowledge that they came from cooperative game theory).

In the context of linear regression, the question is to redistribute the R^2 (the cake) among the players (the covariates).

Let $X_1, \dots, X_d$ be the features (vectors of n observations), and Y be the target. Denote $R^2(A)$ be the R^2 of the linear model:

$$Y = \beta_0 + \sum_{i \in A} \beta_i X_i + \epsilon.$$

Shapley values and proportional values in linear regression

These allocations have been used in statistics for a very long time (without the knowledge that they came from cooperative game theory).

In the context of linear regression, the question is to redistribute the R^2 (the cake) among the players (the covariates).

Let $X_1, \ldots, X_d$ be the features (vectors of n observations), and Y be the target. Denote $R^2(A)$ be the R^2 of the linear model:

$$Y = \beta_0 + \sum_{i \in A} \beta_i X_i + \epsilon.$$

In this case, the cake is $v(D) = R^2(D)$,the R^2 of the linear regression of Y by every feature $X_1, \ldots, X_d$.

And for any coalition of features $A \subseteq D$, $v(A) = R^2(A)$, is the R^2 of the **nested** linear regression of Y by only the coalition of features X_A.

(D, R^2) forms a cooperative game !

LMG and PMVD as importance measures

In order to have a **good** importance measure $\phi_i, i = 1, \ldots, d$, in the context of linear regression (with correlated features), it must satisfy the four following criteria

- **Non-negativity:** $\forall i \in D, \phi(i) \geq 0$;
- **Proper exclusion:** If $\beta_i = 0$, then $\phi_i = 0$;
- **Proper inclusion:** If $\beta_i \neq 0$, then $\phi_i > 0$;
- **Efficiency:** $\sum_{i=1}^{d} \phi_i = R^2(D)$.

Lindeman, Merenda, and Gold (1980) proposed indices called LMG, which are nothing more than **the Shapley values** of (D, R^2). However they violates the proper exclusion criterion.

Feldman (2005) proposed indices called PMVD, which are nothing more than **the proportional values** of the dual of (D, R^2). They respect all four criterion.

Towards sensitivity analysis

But what if the model is non-linear, and the features are correlated ?

Then, whenever, for a non-linear model G, such that:

$$Y = G(X),$$

the cake is $\mathbb{V}(Y) = \mathbb{V}(G(X))$ (notice that there is no ϵ). It leads to the **Sobol' cooperative game** where:

$$v(A) = \mathbb{V}(\mathbb{E}[G(X) \mid X_A]).$$

Its Shapley values are known as the **Shapley effects** (Owen 2014), and the proportional values of its dual are known as the **proportional marginal effects** (Herin et al. 2022).

Main distinction for correlated **exogenous** features:

- Their Shapley effects can be non-zero.
- Their proportional marginal effects are zero (this has been proven).

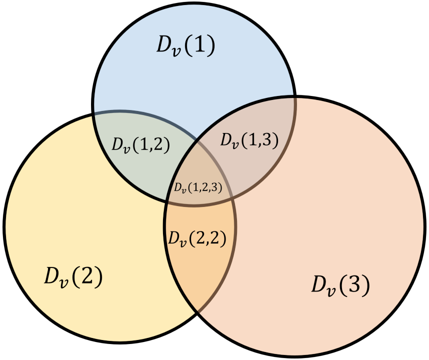
Bottom line: If the **conundrum** is about detecting exogenous inputs, the **Shapley effects may fail**, whereas the **proportional marginal effects do answer the question**.

Interpreting the Shapley values: Harsanyi dividends

Another equivalent enlightening representation of the Shapley values can be done using **Harsanyi dividends** (Harsanyi 1963).

Let (D, v) be a cooperative game, and for any $A \subseteq D$, let the **Harsanyi dividend** of the coalition A be:

$$D_v(A) = \sum_{B \subseteq A} (-1)^{|A|-|B|} v(A).$$



The Harsanyi dividends can be interpreted as the **surplus (or shortfall)** that a coalition generates:

$$D_v(1) = v(1), \quad D_v(2) = v(2),$$

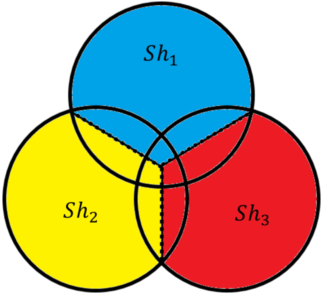
$$D_v(1,2) = v(1,2) - v(1) - v(2).$$

Interpreting the Shapley values: Harsanyi dividends

The Shapley values are then defined as:

$$\mathrm{Sh}_i = \sum_{A \subseteq D : i \in A} \frac{D_v(A)}{|A|},$$

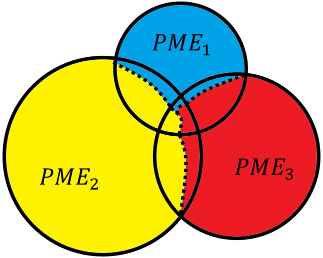
or, in other words, each dividend of a coalition is **equally** redistributed between the players that composes it.



Quick example: Eve and John are two developers, Eve produces 10.000 lines of code, John produces 8.000 lines of code.
However, John really likes to play babyfoot, but Eve is a hard-worker.
When working together, they only produce 10.000 lines of code.
This means that the dividend of their coalition is -8.000.
Is it fair to attribute Eve -4.000 **lines of code, even if she did all the work ?**

Interpreting the proportional values: Harsanyi dividends

The proportional values can also be interpreted in terms of Harsanyi dividends. It represents a **proportional** redistribution of the dividends of coalitions between the players that composes it.



Quick example: In this case, most of the -8.000 lines of code would be given to John, but Eve would also be receive a hit.

Is it still fair ? Is it cooperative ?

Mythbuster

A lot of things are said in the literature about cooperative game theory, and the Shapley values, which are discutable:

- An allocation is a Shapley value.

Mythbuster

A lot of things are said in the literature about cooperative game theory, and the Shapley values, which are discutable:

- An allocation is a Shapley value. **False. Cooperative game theory is a rich field with many different solution concepts.**

Mythbuster

A lot of things are said in the literature about cooperative game theory, and the Shapley values, which are discutable:

- An allocation is a Shapley value. **False.** **Cooperative game theory is a rich field with many different solution concepts.**
- Cooperative game theory deals with cooperation.

Mythbuster

A lot of things are said in the literature about cooperative game theory, and the Shapley values, which are discutable:

- An allocation is a Shapley value. **False. Cooperative game theory is a rich field with many different solution concepts.**
- Cooperative game theory deals with cooperation. **False. It deals with ways coalitions of players can interact between each other. Example of a non-cooperative allocation: Proportional values.**

Mythbuster

A lot of things are said in the literature about cooperative game theory, and the Shapley values, which are discutable:

- An allocation is a Shapley value. **False. Cooperative game theory is a rich field with many different solution concepts.**
- Cooperative game theory deals with cooperation. **False. It deals with ways coalitions of players can interact between each other. Example of a non-cooperative allocation: Proportional values.**
- Shapley values are fair.

Mythbuster

A lot of things are said in the literature about cooperative game theory, and the Shapley values, which are discutable:

- An allocation is a Shapley value. **False. Cooperative game theory is a rich field with many different solution concepts.**
- Cooperative game theory deals with cooperation. **False. It deals with ways coalitions of players can interact between each other. Example of a non-cooperative allocation: Proportional values.**
- Shapley values are fair. **False. They are an egalitarian way of redistributing dividends.**

Mythbuster

A lot of things are said in the literature about cooperative game theory, and the Shapley values, which are discutable:

- An allocation is a Shapley value. **False. Cooperative game theory is a rich field with many different solution concepts.**
- Cooperative game theory deals with cooperation. **False. It deals with ways coalitions of players can interact between each other. Example of a non-cooperative allocation: Proportional values.**
- Shapley values are fair. **False. They are an egalitarian way of redistributing dividends.**
- Explanations based on Shapley values are always suitable.

Mythbuster

A lot of things are said in the literature about cooperative game theory, and the Shapley values, which are discutable:

- An allocation is a Shapley value. **False. Cooperative game theory is a rich field with many different solution concepts.**
- Cooperative game theory deals with cooperation. **False. It deals with ways coalitions of players can interact between each other. Example of a non-cooperative allocation: Proportional values.**
- Shapley values are fair. **False. They are an egalitarian way of redistributing dividends.**
- Explanations based on Shapley values are always suitable. **False. They answer a specific conundrum, which is yet to be fully understood (but we're working on it !).**

Conclusion

- Shapley values **can provide explanations**, but they are not **always suitable**.

Conclusion

- Shapley values **can provide explanations**, but they are not **always suitable**.
- They rely on **egalitarian principles**, which is **not synonym to fairness**.

Conclusion

- Shapley values **can provide explanations**, but they are not **always suitable**.
- They rely on **egalitarian principles**, which is **not synonym to fairness**.
- **Explainability** and **ML interpretability** is not restricted to Shapley values, allocations, or even **prediction decomposition**.

Conclusion

- Shapley values **can provide explanations**, but they are not **always suitable**.
- They rely on **egalitarian principles**, which is **not synonym to fairness**.
- **Explainability** and **ML interpretability** is not restricted to Shapley values, allocations, or even **prediction decomposition**.
- It is not because an interpretation method gives a **good result**, that it is **always suitable**.

To go further (and open questions)

- Consequences between a **player-centric** perspective (cooperative game theory) vs. a **model-centric** perspective (statistics/ML) ?

To go further (and open questions)

- Consequences between a **player-centric** perspective (cooperative game theory) vs. a **model-centric** perspective (statistics/ML) ?
- Same cake as SHAP, but for another allocation ? Can we ensure nice properties ?

To go further (and open questions)

- Consequences between a **player-centric** perspective (cooperative game theory) vs. a **model-centric** perspective (statistics/ML) ?
- Same cake as SHAP, but for another allocation ? Can we ensure nice properties ?
- Do the choice of **value function** matter in the diagnostic, and the subsequent explanation ?

To go further (and open questions)

- Consequences between a **player-centric** perspective (cooperative game theory) vs. a **model-centric** perspective (statistics/ML) ?
- Same cake as SHAP, but for another allocation ? Can we ensure nice properties ?
- Do the choice of **value function** matter in the diagnostic, and the subsequent explanation ?
- Other forms of interpretability methods, for different types of conundrums ? (Robustness, stability, behavior guarantees...).

To go further (and open questions)

- Consequences between a **player-centric** perspective (cooperative game theory) vs. a **model-centric** perspective (statistics/ML) ?
- Same cake as SHAP, but for another allocation ? Can we ensure nice properties ?
- Do the choice of **value function** matter in the diagnostic, and the subsequent explanation ?
- Other forms of interpretability methods, for different types of conundrums ? (Robustness, stability, behavior guarantees...).
- Come-up with, and improve **estimation strategies**.

References i

Asher, N., L. De Lara, S. Paul, and C. Russell. 2022. "Counterfactual Models for Fair and Adequate Explanations" [in en]. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, *Machine Learning and Knowledge Extraction* 4, no. 2 (June): 316–349. ISSN: 2504-4990, accessed November 28, 2022. https://doi.org/10.3390/make4020014. https://www.mdpi.com/2504-4990/4/2/14.

Barredo Arrieta, A., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, et al. 2020. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI" [in en]. *Information Fusion* 58 (June): 82–115. ISSN: 1566-2535. https://doi.org/10.1016/j.inffus.2019.12.012. https://www.sciencedirect.com/science/article/pii/S1566253519308103.

Feldman, B. E. 2005. "Relative Importance and Value" [in English]. *SSRN Electronic Journal,* ISSN: 1556-5068. https://doi.org/10.2139/ssrn.2255827. http://www.ssrn.com/abstract=2255827.

Harsanyi, J. C. 1963. "A Simplified Bargaining Model for the n-Person Cooperative Game." *International Economic Review* 4 (2): 194–220. ISSN: 0020-6598.

Herin, Margot, Marouane Il Idrissi, Vincent Chabridon, and Bertrand Iooss. 2022. "Proportional marginal effects for global sensitivity analysis." Working paper or preprint, October. https://hal.archives-ouvertes.fr/hal-03825935.

Lindeman, R. H., P. F. Merenda, and R. Z. Gold. 1980. *Introduction to bivariate and multivariate analysis.* Glenview, IL: Scott Foresman / Company.

Owen, A. B. 2014. "Sobol' Indices and Shapley Value" [in English]. *SIAM/ASA Journal on Uncertainty Quantification* 2, no. 1 (January): 245–251. ISSN: 2166-2525, accessed December 2, 2020. https://doi.org/10.1137/130936233.

Shapley, L. S. 2016. "17. A Value for n-Person Games." In *Contributions to the Theory of Games* (*AM-28*)*, Volume II,* edited by Harold William Kuhn and Albert William Tucker, 307–318. Princeton University Press. https://doi.org/doi:10.1515/9781400881970-018.

Thank you for your attention!

Any questions?