Epistémologie des modèles et demandes d'explicabilité de l'apprentissage machine

Franck Varenne
Maître de conférence (HDR) en philosophie des sciences
Université de Rouen, ERIAC (EA 4705) & IHPST (UMR 8590)

Séminaire ARISTOTE - 17 octobre 2019
Ecole Polytechnique

Un travail en collaboration…

…avec Christophe Denis

- Maître de conférences (HDR) à Paris 6
  membre du LIP6 – équipe ACASA

CONSTATS

Une demande d'interprétabilité et d'explicabilité de l'AM aux sources multiples

- Cette demande est forte car elle vient à la fois:
  - de préoccupations traditionnelles d'éthique venant des usagers des sciences et techniques
  - de préoccupations des ingénieurs en R/D eux-mêmes (développeurs, modélisateurs) : acceptabilité de techniques efficaces ponctuellement, mais mal comprises, peu robustes et/ou peu éprouvées
  - de préoccupations plus théoriques et académiques: caractère limité et contestable du « prédire sans comprendre » pour l'évolution efficace du savoir

Choix, méthode, thèse principale

- **Choix :**
  - Se concentrer sur une analyse conceptuelle des notions d'interprétation et d'explication
- **Méthode :**
  - S'inspirer de résultats issus de l'épistémologie des modèles en abordant les modèles à AM au titre de modèles prédictifs
- **Thèse principale :**
  - L'absence de représentation d'une causalité dans les modèles à AM reste la cause majeure de leur déficit d'explicabilité

Source : C. Denis, F. Varenne, « Interprétabilité et explicabilité pour l'apprentissage machine : entre modèles descriptifs, modèles prédictifs et modèles causaux. Une nécessaire clarification épistémologique » , *Actes de la CNIA 2019*, pp. 60-68.

Plan de l'exposé

- I. Interprétation et explication : choix des définitions
- II. Quatre fonctions des modèles : analyse de données, description, prédiction, explication
- III. Modèle expliquant, modèle expliqué en physique et en IA symbolique
- IV. Difficile explicabilité des modèles à AM

Conclusions

I. INTERPRÉTATION ET EXPLICATION : CHOIX DES DÉFINITIONS

Définitions courantes

Synthèses de Mittelstadt et al., 2019: « Explaining Explanations in AI »

- Mittelstadt, 2019, p . 2: *« L'**interprétabilité** réfère au degré de compréhensibilité humaine d'un modèle de type boîte noire ou d'une décision »* (Lisboa, 2013; Miller, 2017).

Définitions courantes

Synthèses de Mittelstadt et al., 2019: « Explaining Explanations in AI »

- Mittelstadt, 2019, p . 2: *« L'**interprétabilité** réfère au degré de compréhensibilité humaine d'un modèle de type boîte noire ou d'une décision »* (Lisboa, 2013; Miller, 2017).

  - Notez qu'une telle définition du terme « interprétation » mobilise la notion de **compréhension** (non définie), ce qui dilue le sens originel du terme et inverse à mon avis le rapport habituel de détermination

Définitions courantes

- Pour Mittelstadt, 2019: selon (Lepri, 2017; Lipton, 2016, Montavon, 2017, Mittelstadt, 2019) la **« transparence** porte sur la manière dont un modèle **fonctionne** intérieurement »
  - La **transparence** peut plus spécifiquement viser:
    - *« une compréhension mécanistique du fonctionnement du modèle (simulabilité)*
    - *les composants individuels (décomposabilité)*
    - *La transparence de l'algorithme »*

Définitions courantes

- Pour Mittelstadt, 2019: selon (Lepri, 2017; Lipton, 2016, Montavon, 2017, Mittelstadt, 2019) la **« transparence** porte sur la manière dont un modèle **fonctionne** intérieurement »
  - La **transparence** peut plus spécifiquement viser:
    - *« une compréhension mécanistique du fonctionnement du modèle (simulabilité)*
    - *les composants individuels (décomposabilité)*
    - *La transparence de l'algorithme »*

- « Les **interprétations [explications interprétables] post-hoc** concernent la manière dont le modèle **se comporte** » (ibid.). Elles prennent la forme
    - D'explications verbales
    - De visualisations ou d'interfaçages interactifs
    - D'explications locales et d'approximations
    - D'explications à base de cas particuliers

Définitions courantes

- **« L'explication »** est plus interactive. Elle consiste génériquement « en le fait d'échanger des informations au sujet d'un phénomène » (Mittelstad, 2019)

Définitions courantes

- **« L'explication »** est plus interactive. Elle consiste génériquement « en le fait d'échanger des informations au sujet d'un phénomène » (Mittelstad, 2019)
  - avec **différentes fonctions :**
    - Expliquer que le modèle se conforme bien à une législation
    - Vérifier et améliorer les fonctionnalités du modèle (débugger)
    - Aider les développeurs à apprendre quelque chose du système
    - Améliorer la confiance en le modèle et en ses décisions
  - vers **différentes audiences :**
    - Les développeurs experts
    - Les utilisateurs du modèle
    - Les êtres humains non spécialistes mais affectés par la décision. L'explication a alors un rôle :
      - Pédagogique
      - De persuasion (de bonne foi)
      - De persuasion de mauvaise foi (manipulation, idéologie, biais accepté)

Thèse de Mittelstadt et al. 2019

- Cet article se concentre ensuite sur **“l’explication interprétable [compréhensible] post-hoc”** et constate que les explications post-hoc par **modélisation compréhensive (simplifiante)** de modèle à AM ne sont pas des explications fiables:

  “Explainable AI generates approximate simple models and calls them ‘explanations’, suggesting reliable knowledge of how a complex model functions” (ibid., p. 3)

Problème et Suggestions

- Le sens du terme explication est devenu **trop flottant dans ce débat,** il finit par désigner trop de choses

- Ce faisant, on pense découvrir un problème réel, mais il y a aussi des **problèmes de mots :**

  1) On **confond** interprétation et explication

  2) En outre, le terme **interprétation** lui-même est vague : cela est dû au fait que l'on conditionne dès le début toute interprétation à une **compréhension humaine** alors que c'est l'inverse qui est le plus vraisemblable

  3) Il en ressort qu'**une interprétation n'a pas besoin d'une compréhension** préalable, ni d'une explication

Définitions alternatives

(Denis & Varenne, CNIA 2019)

- **Interprétabilité** d'un modèle: *« propriété qu'a un modèle de se voir composé d'éléments (signes, symboles, figures, concepts, données, etc.) qui ont chacun un sens [c'est-à-dire un référent possible] pour un sujet humain » (Denis, Varenne, p. 61).*

Définitions alternatives

(Denis & Varenne, CNIA 2019)

- **Interprétabilité** d'un modèle: *« propriété qu'a un modèle de se voir composé d'éléments (signes, symboles, figures, concepts, données, etc.) qui ont chacun un sens [c'est-à-dire un référent possible] pour un sujet humain » (Denis, Varenne, p. 61).*
  - *Remarque :*
    - *C'est une définition sémantique liée à une ontologie possible : une sémantique cognitive réelle + une sémantique référentielle possible*
    - *Donc pas liée dès le départ à une compréhension*

Définitions alternatives

- **Explicabilité (de l'algorithme à AM ou des sorties de l'AM) : «** *capacité de déploiement et d'explicitation de cet algorithme ou de ses sorties en séries d'étapes reliées entre elles par ce qu'un être humain peut interpréter sensément comme des causes ou des raisons »* (Denis, Varenne, p. 61).

Définitions alternatives

- **Explicabilité (de l'algorithme à AM ou des sorties de l'AM) : «** *capacité de déploiement et d'explicitation de cet algorithme ou de ses sorties en séries d'étapes reliées entre elles par ce qu'un être humain peut interpréter sensément comme des causes ou des raisons »* (Denis, Varenne, p. 61).

  - *Remarque : une explication nécessite une interprétation:*
    - *1. des éléments*
    - *2. des relations causales entre eux*

Définitions alternatives

- **Compréhension d'un phénomène, ou d'un calcul (*cum-prehendere*)** : il y a compréhension d'un phénomène quand notre esprit dispose de la possibilité d'en **unifier** les manifestations successives ou diverses sous une représentation unique et aisée à concevoir (Varenne, 2013; 2018)

II. QUATRE FONCTIONS DES MODÈLES : ANALYSE DE DONNÉES, DESCRIPTION, PRÉDICTION, EXPLICATION

Définition large du terme « Modèle »

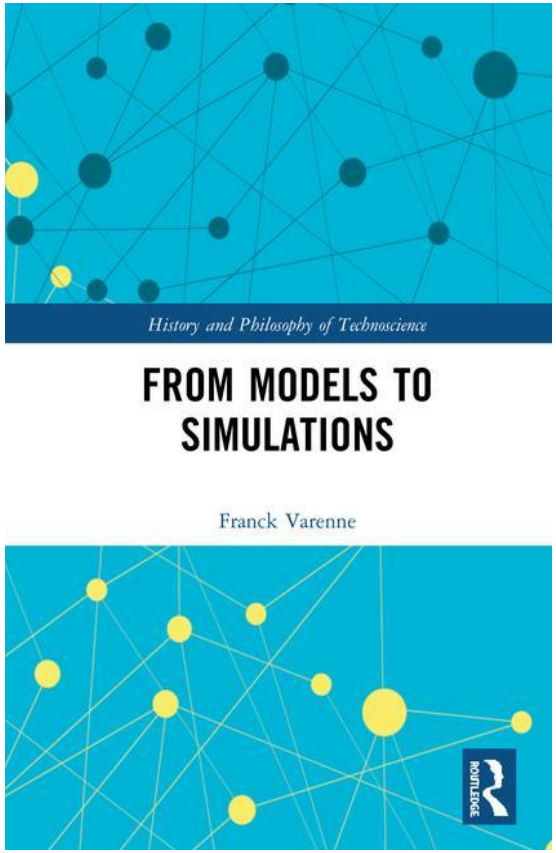
- « *Pour un observateur B, un objet A* est un modèle d'un objet A dans la mesure où B peut utiliser A* pour répondre à des questions qui l'intéressent au sujet de A » (Minsky, 1967)

Définition large du terme « Modèle »

- « *Pour un observateur B, un objet A* est un modèle d'un objet A dans la mesure où B peut utiliser A* pour répondre à des questions qui l'intéressent au sujet de A » (Minsky, 1967)*
- C'est un objet qui assure une **médiation facilitante** dans le cadre d'une questionnement
- Il y a de nombreux types de facilitation

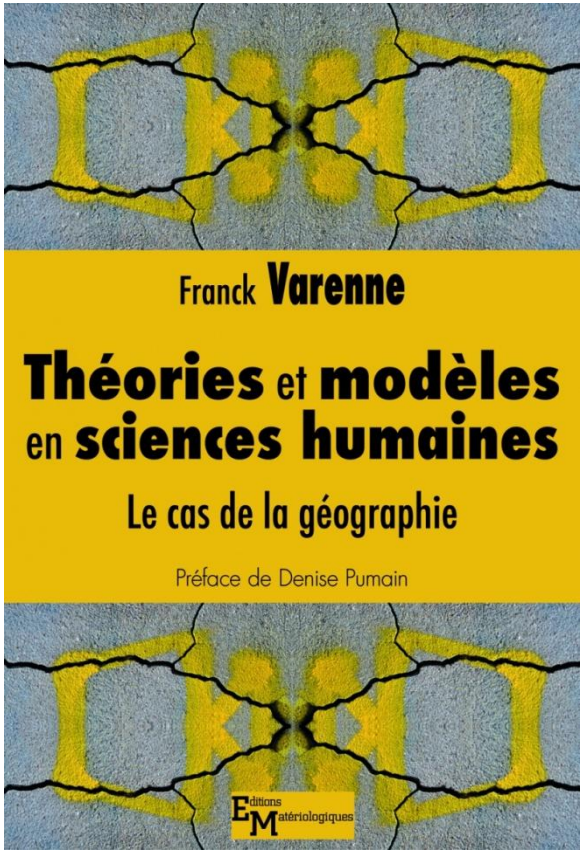
Sur les fonctions de connaissance des modèles scientifiques

(21 fonctions répertoriées)



From Models to Simulations,

Routledge, 2018, 224p.



Théories et modèles en sciences humaines. Le cas de la géographie, Paris, Matériologiques, 2017, 644p.

Tableau récapitulatif des fonctions des modèles		
Grandes fonctions	**Fonctions spécifiques**	**Exemples**[1]
I Faciliter l'appréhension sensible	1. Rendre perceptibles certaines propriétés sur un substitut	Écorchés de cire, maquettes de molécules avec des billes…
	2. Rendre perceptibles certains rapports sur un substitut	Diagrammes, cartes, organismes modèles, maquettes de bateaux…
	3. Faciliter la mémorisation par une représentation ordonnée	Contines, images, théâtres mentaux, systèmes architecturaux, mémoire locale…
	4. Condenser l'information pour faciliter l'accès et le rappel à volonté	Systèmes d'axes de symétrie, moments statistiques (moyenne, variance, etc.), paramètres de modèles statistiques analytiques…
II Faciliter la formulation intelligible	5. Faciliter la compression de données pour préparer la conceptualisation	Modèles de données, modèles statistiques descriptifs ou synthétiques, enveloppe statistique…
	6. Faciliter une sélection de types d'entités ou de propriétés	Modèles conceptuels, modèles de connaissance, classifications, hiérarchies, ontologies…
	7. Faciliter la reproduction ou production de structures de données par des moyens intelligibles déductifs ou de calcul	Modèles phénoménologiques (à base de données), modèles descriptifs et/ou prédictifs, modèles de conception (ingénierie), modèles de synthèse par analyse spectrale de données…
	8. Faciliter une explication	Modèle d'interaction, modèle de séquence finie d'interactions, modèle de mécanismes…
	9. Faciliter une compréhension	Modèle d'optimisation, modèle à principe variationnel valant à échelle agrégée, modèle axiologique à l'échelle des individus (rationalité en valeur), geste mental, idéal-type…

Source : Varenne, ***Théories et modèles en sciences humaines***, 2017, p. 140.

III Faciliter la théorisation	10. Faciliter une ébauche de théorie : modèle théorique	*Homo economicus*, modèles de la rationalité (utilitariste, limitée ou ordinaire), « théories de l'acteur rationnel »…
	11. Faciliter une interprétation de théorie : modèle de théorie	Images mentales, modèles physiques de théories mathématiques…
	12. Faciliter une illustration de théorie : modèle pour la théorie	Modèle des courants de fluide pour la circulation électrique, modèle d'oscillateurs électriques pour la théorie des dynamiques de populations…
	13. Faciliter un test de cohérence interne de la théorie	Modèle sémantique, modèle concret (*i. e.* se référant à des objets) d'une théorie formelle, modèle des valeurs de vérité en théorie logique des propositions, modèle euclidien pour une géométrie…
	14. Faciliter l'applicabilité de la théorie	Modèle sémantique approché ou incomplet, sous-structures empiriques[2]…
	15. Faciliter la calculabilité d'une théorie	Modèle partiellement phénoménologique ou approché du fonctionnement de la théorie mathématique, modèle numérique, *computational template*[3] ou gabarit computationnel, modèle de simulation de type 2[4]…
	16. Faciliter une hybridation et une cocalculabilité de plusieurs théories	Modèle mixte polyphase, modèle *ad hoc*, modèle asymptotique, modèle multi-échelles…

Source : Varenne, ***Théories et modèles en sciences humaines***, 2017, p. 141.

IV Faciliter la coconstruction des savoirs	17. Faciliter une communication entre acteurs scientifiques	Base de données, ontologie explicite et ouverte, modèle de vulgarisation…
	18. Faciliter la délibération et la concertation entre parties prenantes	Modèle multi-aspectuel pour la concertation, modèle d'exploration de scénarios concertés…
	19. Faciliter la coconstruction de représentations et de modes de contrôle de système mixtes (humains/ non-humains)	Modèle en recherche-action, modèle participatif, modélisation d'accompagnement…
V Faciliter la décision et l'action	20. Faciliter l'intervention *sur* un système mixte et hétérogène	Modèle de décision, arbres de décision, modèle de crise, heuristique…
	21. Faciliter une décision d'action *dans* un système principalement notionnel	Modèle d'anticipation de marché, modèle de produits dérivés en finance…

Source : Varenne, ***Théories et modèles en sciences humaines***, 2017, p. 142.

Fonctions les plus fréquentes

- Les plus fréquentes sont les fonctions 5, 6, 7 et 8 :

  – l’analyse ou la réduction de données
  – la description
  – la prédiction
  – l’explication

MODÈLES D’ANALYSE DE DONNÉES

Les modèles d'analyse de données

- Ils opèrent sur la **seule structure informationnelle** des données du système cible, mais pas directement sur la structure des propriétés intrinsèques du système cible ni des relations mutuelles entre ces propriétés
- Ils sont **faiblement prescriptifs ontologiquement**.
- Ils **préparent** l'utilisation d'autres modèles : les modèles à fonction de description ou d'explication du système cible.
- C'est parce qu'ils traitent les données comme des ***signaux*** mais pas comme des ***signes***.

Approche « signal » *Vs.* Approche « signe »

- Un **signal** indique, qualifie ou quantifie une interaction.
- Un signal est le résultat de la détection ou de la mesure par capteur physique d'un phénomène d'**interaction** entre l'objet cible et son environnement physique.
- Cet **environnement** est au **minimum** son cadre spatial, temporel ou spatio-temporel.
- Les propriétés physiques du système cible sont certes supposées exister mais leur nature peut demeurer largement **inconnue**.

- Un **signe** désigne, qualifie ou quantifie une propriété, en lui donnant une unité de mesure significative, c'est-à-dire interprétable.
- Dans **l'approche « signe »**, on entend rendre compte d'une propriété du système cible et de sa valeur.
- On recourt pour cela à cet autre type de médiateurs que sont les **instruments de mesure**.

MODÈLES DESCRIPTIFS

Les modèles descriptifs

- Ils reproduisent et structurent des données qui, séparément, ont déjà un **sens minimal**, c'est-à-dire qui sont pour certains interprétables en termes de signes référant à des propriétés au regard de la connaissance que l'on a, par ailleurs, du système cible

Les modèles descriptifs

- Mais la **structure** que le modèle descriptif propose pour ces propriétés (leur relation mutuelle représentée dans le modèle) peut être complètement phénoménologique, c'est-à-dire ne pas se fonder elle-même sur une propriété profonde de structure du système cible.
- Ainsi, la structure représentée peut n'avoir pas de sens, pas d'**interprétation**, mais certains éléments qui la composent **doivent** en avoir
- C'est la différence essentielle avec les modèles d'analyse ou de réduction de données

MODÈLES PRÉDICTIFS

Les modèles prédictifs

- Ce sont des cas particuliers de **modèles descriptifs dynamiques** (i.e. avec état initial et état final)
- Ils **décrivent** le système dynamiquement à travers au minimum deux types de données qui le représentent partiellement sans encore l'expliquer :
  - des **données prédictives** (nommées parfois « explicatives » de manière trompeuse en statistique inférentielle) utilisées par l'algorithme ou le modèle
  - et des **donnés comportementales ou prédites** qui servent à évaluer la qualité de la prédiction, donc la qualité du modèle

Les modèles prédictifs

- Plus précisément, c'est lorsque cette dynamique reproduite permet non seulement …
  - 1. de **décrire** correctement le **comportement observable** du système dans les cas connus d'entrées/sorties
  - 2. mais aussi **d'interpoler ou d'extrapoler** correctement une description du comportement observable du système cible à partir de données qui n'ont pas été utilisées pour calibrer le modèle (données nouvelles, période de temps non encore testée),
- …que le modèle descriptif (de régression, à réseau de neurones, de classification, etc.) se trouve être également un **modèle prédictif**

Deux grands types de modèles prédictifs en AM

- **<u>à régression</u>** au sens large : ceux qui servent à prédire des **variables quantitatives**
- **<u>de classification</u>**, i.e. servant à prédire une **variable qualitative** ou, plus largement, à estimer la **probabilité d'un événement**

Deux types de modèles prédictifs de classification

- **les modèles génératifs** se fondant sur l'hypothèse de l'existence d'une distribution de probabilité conjointe précise [P(prédictive **et** observée)] et permettant d'**engendrer** les probabilités conditionnelles

- les **modèles discriminatifs (ou séparateurs)** qui ne partent pas d'une distribution de probabilité conjointe et définissent directement les probabilités conditionnelles [P(observée**/**prédictive)] : régression logistique, perceptron (RN à une couche), modèles à vecteurs de support (SVM)

Débat sur les classifieurs génératifs

- D'après S. Shalev-Schwartz et S. Ben-David, (*Understanding Machine Learning: From Theory to Algorithms,* 2014), comme ces modèles permettent l'adoption d'une approche bayésienne, si les *probabilités a priori* sont acceptées et se révèlent fécondes à l'usage, cela peut nous mener à l'idée que le **modèle prédictif est explicable parce qu'explicatif au regard d'une hypothèse informationnelle explicite sur le monde**.

Débat sur les classifieurs génératifs

- D'après S. Shalev-Schwartz et S. Ben-David, (*Understanding Machine Learning: From Theory to Algorithms,* 2014), comme ces modèles permettent l'adoption d'une approche bayésienne, si les *probabilités a priori* sont acceptées et se révèlent fécondes à l'usage, cela peut nous mener à l'idée que le **modèle prédictif est explicable parce qu'explicatif au regard d'une hypothèse informationnelle explicite sur le monde**.

- Mais j'objecterais que le fondement de l'explicabilité de tels modèles reste ***ad hoc*** et **seulement épistémique** (bayésienne) car n'entendant nullement reposer sur une hypothèse ontologique large de **lois de la nature et de causalité**.

MODÈLES EXPLICATIFS

Les modèles explicatifs

- **Mise au point sur « explication »**
  - En philosophie des sciences contemporaine, il n'existe pas de consensus sur la différence précise entre expliquer et comprendre.
  - Cependant, une grande partie des auteurs (cf. Varenne, 2018, p. 18) s'accorde sur le fait d'associer :
    - l'**explication à la causalité, plus précisément à des mécanismes**
    - et la **compréhension à l'unification** d'une diversité de phénomènes sous un **principe unique**

Les modèles explicatifs

- **Mise au point sur « causalité »**
  - Nous n'entendrons pas ici causalité au sens où l'entend la pratique de **l'inférence causale**
  - Car celle-ci reste **épistémique**, même si la sophistication récente de sa stratégie de **formalisation des propositions contrefactuelles** au moyen d'une approche **structurelle** se révèle **pragmatiquement efficace** : cf. Judea Pearl : *Models, Reasoning, and Inference*, 2009 ; *The Book of Why*, 2018.

Les modèles explicatifs

- On peut dire qu'un modèle mathématique ou algorithmique est **explicatif d'un système cible** lorsque :
  - Il est au moins **partiellement prédictif** pour ce système,
  - Il offre une **représentation interprétable**, c'est-à-dire **signifiante** et accessible à un esprit humain non aidé, **à la fois** des **éléments** dont il est composé et **des processus** élémentaires d'interaction qu'il met en œuvre **(« sémantique cognitive »)**
  - Ces éléments et processus élémentaires peuvent être supposés eux-mêmes référer plus ou moins iconiquement à des éléments et des processus d'interaction causale (ou mécanismes) intervenant **réellement** et **majoritairement** dans le système cible lui- même **(« sémantique référentielle »)**

Les modèles explicatifs en physique

- **Critères pour leur validation :**

  Un modèle explicatif ne devra pas être seulement validé dans sa **capacité à reproduire certains comportements** du système cible. Il faudra aussi évaluer sa **capacité à représenter pas à pas**, de manière correcte, *i.e.* approximativement réaliste, non seulement les **états successifs du système cible** mais aussi **chaque opération du processus** lui-même (appliquant chacun un mécanisme implémentant une loi physique).

III. MODÈLE EXPLIQUANT, MODÈLE EXPLIQUÉ EN PHYSIQUE ET EN IA SYMBOLIQUE

Question : *quid* de l'explication de modèle ?

- **Attention** :

  Explication du système cible par le modèle ≠ Explication du modèle lui-même et de son fonctionnement

- **Mais** il y a des liens entre les deux

Exemples de la détermination de l'une par l'autre

- **Exemple : un modèle d'interaction locale** entre deux astres reposant sur les lois de Newton, ou une interaction entre deux atomes en chimie reposant sur l'équation de Schrödinger, etc.
  - On peut alors modéliser le système cible en modélisant de manière fidèle la causalité même affectant ce système cible.
  - On le fait en représentant de manière iconique (*i.e.* au moins termes à termes) les principaux éléments en interaction et leurs principales interactions causales
  - Par là, le modèle est fidèle à la fois à l'individuation des éléments naturels réels comme une planète, un atome, et à la réalité des interactions causales entre ces individus

Exemples de la détermination de l'une par l'autre

- **Autre exemple : un système expert (IA symbolique) modélisant une décision médicale** en recourant à des bases de données et des règles de raisonnement sensées et appliquées pas à pas. Les calculs de ce modèle **expliquent** en même temps le processus même de la décision.
- Notons que, dans ce cas de décision humaine motivée, on peut considérer qu'une **raison interprétable** joue le même rôle qu'une **cause signifiante** dans un système physique soumis à des lois physiques

Exemples de la détermination de l'une par l'autre

- Dans ces deux cas favorables, le modèle est non seulement **explicatif** mais aussi **explicable**.
- Cela veut dire que le **processus de computation** suivi par le modèle implémenté dans le programme est également **interprétable et explicable** en lui-même.
- Il est interprétable car l'ontologie du modèle renvoie à des ensembles d'entités et de propriétés reconnues comme existant réellement dans le système cible (**sémantique référentielle**) auquel on a accès par ailleurs sous une forme interprétable (**sémantique cognitive**).

Exemples de la détermination de l'une par l'autre

- Dans ce **cas particulier de modèle explicable**, on utilise donc la connaissance préalable que l'on a :
  - 1) de la **structuration réelle** du système cible (l'ontologie qu'on lui reconnaît),
  - 2) du fait que le **modèle utilise cette structuration** et n'utilise qu'elle dans ses processus,
  - 3) du fait que les **processus** du modèle sont également supposés **réalistes**,
  - 4) du fait que ce dépliement processuel pas à pas **converge** mathématiquement vers les résultats,
- pour, au final, décider que le modèle non seulement explique son système cible mais qu'il est également **interprétable et explicable** en lui-même.

IV. DIFFICILE EXPLICABILITÉ DES MODÈLES À AM

Cas des modèles à AM

- À la différence des cas précédents, l'explicabilité du modèle à AM n'est pas aussi facile à assurer déjà parce qu'elle ne peut pas être directement héritée, par transitivité, du fait que le modèle serait explicatif.

Cas des modèles à AM

- **L'explicabilité du modèle que l'on recherche doit être fondée autrement pour deux raisons.**
  - 1. Comme pour un modèle d'analyse de données standard, en AM, le modèle contrôlant les relations entrées/sorties **n'entend pas représenter**, même de manière seulement stylisée, **un scénario causal** d'interactions pas à pas opérant sous l'effet de lois ou de règles motivées, mais des corrélations.
  - 2. Mais la situation de l'AM est pire : **l'ontologie sous-jacente aux données et à leur structure peut en effet être maintenue complètement inconnue ou fictionnelle** (données mal ou non structurées)

Cas des modèles à AM

- Quand bien même une structure serait perceptible dans les données, une technique comme les RN par exemple met en œuvre **des modèles non linéaires** reliant les valeurs prédictives et les valeurs prédites: les valeurs prédictives interagissent fortement, donc on ne peut plus parler de simples corrélations.
- Certes, dans le cas d'un **modèle non linéaire à arbres de décision**, les étapes élémentaires restent interprétables une à une, mais le processus d'ensemble n'est pas pour autant aisément sensément résumable : il n'est pas compréhensible

« Pari » métaphysique en analyse des données

- Même en analyse des données classique, le modèle repose sur des **hypothèses globales et minimales** - qu'on peut dire métaphysiques - de **symétries temporelles ou spatiales des signaux** liées à l'environnement de captation des données (y compris dans les approches dites « non paramétriques »)

Quels sont les paris métaphysiques sous-jacents aux RN ?

- Ce sont ces hypothèses métaphysiques minimalistes qui ne sont même pas toujours possibles en AM.
- Voir le rapprochement récent entre l'analyse par ondelettes et les réseaux de neurones convolutionnels: Stéphane Mallat, "Understanding deep convolutional networks". *Phil. Trans. R. Soc*, 2016.
- Ce résultat incite à penser qu'un RN quelconque (non convolutionnel) est en général plus neutre encore et moins-disant d'un point de vue métaphysique et causal que les approches par analyse de données paramétriques ou non paramétriques (fondées sur des symétries)

Synthèse sur l'explicabilité *a priori* des modèles à AM

Synthèse sur l'explicabilité *a priori* des modèles à AM

- Les modèles à AM ne peuvent pas hériter directement leur interprétabilité et leur explicabilité du caractère réaliste et causal des interactions qu'ils modélisent dans leur calcul.

Synthèse sur l'explicabilité *a priori* des modèles à AM

- Les modèles à AM ne peuvent pas hériter directement leur interprétabilité et leur explicabilité du caractère réaliste et causal des interactions qu'ils modélisent dans leur calcul.
- Ce défaut fragilise les pratiques de vérification, de validation, mais aussi de diffusion et d'appropriation par les utilisateurs, d'où les demandes d'interprétabilité et d'explicabilité de ces modèles

Nature des demandes

- **La demande d’interprétabilité**
- **La demande d’explicabilité**
- **La demande de compréhensibilité**

Nature des demandes

- La **demande d'interprétabilité** d'un modèle d'AM revient finalement à demander la construction d'un **modèle descriptif de ce modèle** d'AM: i.e. donnant des significations à quelques uns de ses composants, et à quelques unes des étapes de sa dynamique

Nature des demandes

- La **demande d'explicabilité** d'un modèle d'AM, quant à elle, revient à demander d'en construire un **modèle explicatif** avec représentation au moins partielle d'une **causalité** possible.

Nature des demandes

- La **demande de compréhensibilité**
- Dans la demande d'explicabilité, il y entre en fait souvent aussi une **demande de compréhensibilité** du modèle. C'est cela qu'on appelle **XAI** : **eXplanable AI**
- On recherche alors, en plus de son explication, des grands principes unificateurs permettant de penser et représenter de manière unitaire le fonctionnement global, la logique globale du modèle : voir Mittelstad et al. (2019)
- Notons que la **légitimation et l'acceptabilité** plus large (technique, sociale,…) du modèle va souvent de pair avec sa **compréhensibilité**

Relativité de la compréhensibilité

- Plus encore que l'interprétabilité, la compréhensibilité est très sensible à l'arrière-plan et aux compétences intellectuelles de la personne à laquelle elle s'adresse

- C'est là qu'entre en jeu la **rhétorique** (Aristote): la rhétorique existe pour persuader et/ou mettre en confiance les personnes qui ne peuvent suivre de manière attentive de **longues** chaînes de causes ou de raisons

- Cette rhétorique prend la forme de modélisation simplifié du modèle de décision

Biais propre au format des donnés

- Comme les données pour un AM ne sont pas reliées d'entrée de jeu à une ontologie explicite ni à un scénario causal explicite (voir ce que l'on a dit précédemment)…
- …mais que l'on peut continuer à dire qu'on en propose une sorte d'interprétabilité puis une sorte d'explicabilité pour le modèle simplifiant qui les traite ensuite…
- …cette interprétabilité et cette explicabilité post hoc peuvent avoir pour effet de **masquer davantage** encore les biais dus aux choix de format et de représentation qui structurent implicitement les données initiales et leurs prétraitements.

Origine de la confusion

- Ce n'est pas parce qu'on a réussi à modéliser de manière explicative ou compréhensive un modèle par ailleurs purement prédictif que l'on a rendu ce modèle explicatif
- On a pu expliquer le fonctionnement interne de ce modèle prédictif mais pas le fonctionnement du système cible initial.
- Car l'explication simplifiée trouvée peut ne pas convenir au système cible lui-même
- On peut trouver une sémantique rendant compréhensible dans les grandes lignes un modèle, mais cette sémantique pourrait s'avérer être nullement réaliste
- Autrement dit : une explication valable dans un sémantique cognitive n'est pas assurée d'être aussi une explication valable dans une sémantique référentielle

Tout ce que l'on peut dire

- Ainsi, les structures implicites de données qui nourrissent l'AM sont certes à l'œuvre, mais sans que l'on sache dans quelle mesure ni à quel niveau c'est le cas, même quand le modèle prédictif remplit son office.
- Le succès d'un modèle de prédiction n'est pas une preuve mais seulement une **indication** que les éléments qu'ils postulent explicitement pour effectuer ses calculs (par exemple : des neurones très simplifiés) **pourraient** refléter finalement quelque chose qui serait réellement à l'œuvre dans le système cible.

Thèse biomimétiste et objection

- La croyance en le caractère non surprenant de cette coïncidence est ce qui fonde l'opinion philosophique **biomimétiste** de Yann Le Cun

- **Objection possible** : le théorème d'universalité concernant les RN peut aussi nous engager plutôt à penser qu'il s'agit simplement d'une autre forme, parmi d'autres, d'automate universel de calcul, simplement plus commode à utiliser en pratique pour certaines formes de données (celles dont on ignore la structure) et de questions qu'on leur adresse.

Conclusions 1/2

Conclusions 1/2

- Nous avons proposé de redéfinir les notions d'interprétation et d'explication de modèle essentiellement à partir de considérations de sémantique et de causalité

Conclusions 1/2

- Nous avons proposé de redéfinir les notions d'interprétation et d'explication de modèle essentiellement à partir de considérations de sémantique et de causalité

- Nous avons précisé la différence entre les fonctions de connaissance des modèles d'analyse de données, de description, de prédiction, d'explication

Conclusions 1/2

- Nous avons proposé de redéfinir les notions d'interprétation et d'explication de modèle essentiellement à partir de considérations de sémantique et de causalité

- Nous avons précisé la différence entre les fonctions de connaissance des modèles d'analyse de données, de description, de prédiction, d'explication

- Nous avons introduit la distinction entre approche « signal » et approche « signe »

Conclusions 1/2

- Nous avons proposé de redéfinir les notions d'interprétation et d'explication de modèle essentiellement à partir de considérations de sémantique et de causalité

- Nous avons précisé la différence entre les fonctions de connaissance des modèles d'analyse de données, de description, de prédiction, d'explication

- Nous avons introduit la distinction entre approche « signal » et approche « signe »

- Par leur approche « signal » et le faible rôle donné à la sémantique, les modèles prédictifs à AM ne prétendent pas représenter de causalité et s'apparentent aux modèles d'analyse de données

Conclusions 1/2

- Nous avons proposé de redéfinir les notions d'interprétation et d'explication de modèle essentiellement à partir de considérations de sémantique et de causalité

- Nous avons précisé la différence entre les fonctions de connaissance des modèles d'analyse de données, de description, de prédiction, d'explication

- Nous avons introduit la distinction entre approche « signal » et approche « signe »

- Par leur approche « signal » et le faible rôle donné à la sémantique, les modèles prédictifs à AM ne prétendent pas représenter de causalité et s'apparentent aux modèles d'analyse de données

- Nous avons rappelé que les modèles à analyse de données classiques reposent sur des hypothèses métaphysiques minimales concernant (les contraintes pesant sur) la structure des signaux qu'ils prennent en compte

Conclusions 1/2

- Nous avons proposé de redéfinir les notions d'interprétation et d'explication de modèle essentiellement à partir de considérations de sémantique et de causalité

- Nous avons précisé la différence entre les fonctions de connaissance des modèles d'analyse de données, de description, de prédiction, d'explication

- Nous avons introduit la distinction entre approche « signal » et approche « signe »

- Par leur approche « signal » et le faible rôle donné à la sémantique, les modèles prédictifs à AM ne prétendent pas représenter de causalité et s'apparentent aux modèles d'analyse de données

- Nous avons rappelé que les modèles à analyse de données classiques reposent sur des hypothèses métaphysiques minimales concernant (les contraintes pesant sur) la structure des signaux qu'ils prennent en compte

- Nous avons suggéré que certaines techniques à apprentissage machine ne semblent pas reposer sur des hypothèses de structure du signal qui soient aussi claires ni interprétables pour nous à l'heure actuelle

Conclusions 2/2

Conclusions 2/2

- Thèse principale : l'absence de représentation d'une causalité reste à l'origine des points de fragilité de l'apprentissage machine déjà signalés dans la littérature

Conclusions 2/2

- Thèse principale : l'absence de représentation d'une causalité reste à l'origine des points de fragilité de l'apprentissage machine déjà signalés dans la littérature

- Nous avons montré que les modèles explicatifs sont d'emblée explicables, mais que les modèles prédictifs à AM ne le sont pas directement le plus souvent, bien qu'ils puissent être expliqués secondairement par d'autres modèles

Conclusions 2/2

- Thèse principale : l'absence de représentation d'une causalité reste à l'origine des points de fragilité de l'apprentissage machine déjà signalés dans la littérature

- Nous avons montré que les modèles explicatifs sont d'emblée explicables, mais que les modèles prédictifs à AM ne le sont pas directement le plus souvent, bien qu'ils puissent être expliqués secondairement par d'autres modèles

- Ces modèles de modèles d'AM peuvent les rendre localement ou approximativement « explicables » mais sans pour autant légitimer les ontologies hypothétiques (à sémantique seulement cognitive) mobilisées par ces modèles

Conclusions 2/2

- Thèse principale : l'absence de représentation d'une causalité reste à l'origine des points de fragilité de l'apprentissage machine déjà signalés dans la littérature

- Nous avons montré que les modèles explicatifs sont d'emblée explicables, mais que les modèles prédictifs à AM ne le sont pas directement le plus souvent, bien qu'ils puissent être expliqués secondairement par d'autres modèles

- Ces modèles de modèles d'AM peuvent les rendre localement ou approximativement « explicables » mais sans pour autant légitimer les ontologies hypothétiques (à sémantique seulement cognitive) mobilisées par ces modèles

- Les usages rhétoriques ou pédagogiques que l'on fait de ces modèles expliquant les modèles à AM ne doivent donc pas faire oublier les fragilités persistantes des modèles qu'ils modélisent.

MERCI DE VOTRE ATTENTION !

Quelques références

Denis, C., Varenne, F., « Interprétabilité et explicabilité pour l'apprentissage machine : entre modèles descriptifs, modèles prédictifs et modèles causaux. Une nécessaire clarification épistémologique », *Actes de la CNIA PFIA 2019*, Toulouse, AFIA, J. Lang (dir.), pp. 60-68. https://www.irit.fr/pfia2019/wp-content/uploads/2019/07/actes_CNIA_PFIA2019.pdf et https://hal.archives-ouvertes.fr/hal-02184519 (**source principale de l'exposé**)

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A. Specter, M., Kagal, L., "Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning", 2018, https://arxiv.org/abs/1806.00069

Herman, B., "The Promise and Peril of Human Evaluation for Model Interpretability", *Thirsty-first Conference on Neural Information Processing Systems*, 2017, https://arxiv.org/abs/1711.07414

Mittelstadt, B., Russell, C., Wachter, S., "Explaining Explanations in AI", *Proceedings of FAT* '19: Conference on Fairness, Accountability, and Transparency* (FAT* '19), January 29–31, 2019, Atlanta, GA, USA. ACM, New York, NY, USA, doi/10.1145/3287560.3287574. https://ssrn.com/abstract=3278331

Varenne, F., « Modèles et simulations dans l'enquête scientifique : variétés traditionnelles et mutations contemporaines », *in* F. Varenne, M. Silberstein, *Modéliser & simuler. Épistémologies et pratiques de la modélisation et de la simulation,* Tome 1, Paris, Matériologiques, 2013, pp. 9-47, https://www.academia.edu/16403372/Mod%C3%A8les_et_simulations_dans_lenqu%C3%AAte_scientifique_vari%C3%A9t%C3%A9s_traditionnelles_et_mutations_contemporaines et Modéliser & Simuler

Varenne, F., *From Models to Simulations,* London, Routledge, 2018. https://www.crcpress.com/From-Models-to-Simulations/Varenne/p/book/9781138065215