# Générer des explications contrefactuelles à l'aide d'un autoencodeur supervisé

Victor Guyomard[*,**], Françoise Fessant[*]
Tassadit Bouadi[**] Thomas Guyet[***]

[*]Orange, Lannion, France
[**]Univ Rennes, Inria, CNRS, IRISA, Rennes, France
[***]Inria – Grenoble Rhône-Alpes, Villeurbanne, France

**Résumé.** Dans cet article nous proposons une manière d'améliorer l'interprétabilité des explications contrefactuelles. Une explication contrefactuelle se présente sous la forme d'une version modifiée de la donnée à expliquer qui répond à la question : que faudrait-il changer pour obtenir une prédiction différente ? La solution proposée consiste à introduire dans le processus de génération du contrefactuel un terme basé sur un auto-encodeur supervisé. Ce terme contraint les explications générées à être proches de la distribution des données et de leur classe cible. La qualité des contrefactuels produits est évaluée sur un jeu de données d'images par le biais de différentes métriques. Nous montrons que notre solution s'avère compétitive par rapport à une méthode de référence de l'état de l'art.

## 1 Introduction

L'apprentissage automatique est désormais massivement utilisé pour automatiser la prise de décision dans de nombreux domaines, et en particulier dans des domaines qui impactent notre vie quotidienne. Les modèles utilisés sont généralement complexes et opaques. C'est le phénomène de la « boite noire ». L'IA explicable (ou XAI) vise à limiter ce problème en fournissant un ensemble de méthodes pour qu'un utilisateur humain comprenne les facteurs qui ont motivé la décision d'un modèle. L'enjeu de l'explicabilité devient crucial que ce soit pour l'acceptation de l'IA ou le respect des réglementations. Par exemple, le règlement général sur la protection des données de l'union européenne, entré en application en 2018, introduit un droit à l'explication pour les individus lorsque la prise de décision automatisée les affecte significativement. De nombreuses approches d'explicabilité ont été développées récemment et plusieurs typologies existent pour les classifier (Molnar, 2019). On s'intéresse ici aux méthodes d'interprétabilité *post-hoc* locales, c'est à dire qui s'appliquent après l'apprentissage du modèle de classification et pour une prédiction donnée. Les explications contrefactuelles appartiennent à cette catégorie. Le principe est d'expliquer la décision du modèle de classification à l'aide d'un exemple, proche de l'exemple à expliquer, qui montre comment celui-ci devrait changer pour que sa prédiction change.

La plupart des méthodes d'explications contrefactuelles sont basées sur la perturbation de l'instance originale grâce à l'optimisation d'une fonction de coût (Wachter et al., 2018). Se-