La “vérité dans la fiction” conçue comme un problème d’apprentissage non-supervisé

Louis Rouillé
louis.rouille@netc.eu

Dans cette présentation je défends l'idée que le problème dit de la "vérité dans la fiction" est conceptuellement identique à un problème d'apprentissage non-supervisé. Si cette idée est juste, alors il faudrait que les philosophes de la fiction et les chercheur.euses en apprentissage statistique commencent à collaborer étroitement.

Le problème de la vérité dans la fiction consiste à expliquer le contraste entre les deux énoncés suivants:

(1) Hamlet est un humain.

(2) Hamlet est un crocodile.

(1) est *vrai* dans *Hamlet*; tandis que (2) est *faux* dans *Hamlet*.[1] L'existence de ce contraste n'est pas controversé, cependant son interprétation est très controversée. Certains comme (Lewis 1978) affirment que le contraste se situe au niveau des *conditions de vérité* de ces énoncés, d'autres comme (Walton 1990) pensent que c'est une fausse piste.[2]

Quel que soit le résultat de ce débat, on peut avec profit caractériser le problème de manière un peu plus abstraite. En réalité, le contraste ci-dessus est le résultat d'un raisonnement. En effet, ni (1) ni la négation de (2) ne sont des phrases du texte de Shakespeare. Cependant, Shakespeare a écrit un énoncé très pertinent: Hamlet est un prince. (C'est dans le titre: *La tragédie de Hamlet, prince du Danemark*.) On raisonne donc ainsi: dans *Hamlet*, Hamlet est un prince; les princes sont des humains; donc, (1) est vrai dans *Hamlet*. (2) est faux dans *Hamlet*, puisque (1) est vrai dans *Hamlet*. Le type de contraste qui nous intéresse repose donc sur les inférences que l'on fait en lisant des textes de fiction.

Et c'est là que le bât blesse (comme dirait Hamlet) car dans la fiction, les inférences font parfois des choses bizarres. En effet, dans certaines fictions, les princes sont des crocodiles. Dans de telles fictions, on conclurait que (2) est vrai... Naturellement, *Hamlet* n'est pas une fiction de ce genre. Mais comment le savons-nous?

Le problème de la fiction, en toute généralité, consiste donc à définir la relation d'inférence propre à chaque fiction, en prenant en compte toute une série d'information sur le type de fiction auquel on a affaire. De manière abstraite, on peut concevoir une telle relation d'inférence comme un méchanisme cognitif permettant d'extraire l'ensemble des "vérités dans la fiction" à partir de l'ensemble des phrases qui constituent le texte.

Une distinction relative à la complexité des problèmes à résoudre en Intelligence Artificielle a émergé avec force récemment.[3] Les problèmes d'apprentissage supervisés (typiquement des problèmes de classification d'images) sont des problèmes pour lesquels on dispose de données annotées qui vont servir à entraîner des machines (typiquement des réseaux de neurones convolutifs).[4] Concrètement, les machines extrapolent ce qu'elles ont "appris" en phase d'entraînement pour résoudre le problème sur des données non-annotées. On obtient de très bon résultats avec cette méthode dite d'apprentissage profond.

L'apprentissage non-supervisé, en revanche, désigne les problèmes pour lesquels on ne peut pas utiliser des données annotées comme base d'entraînement. Les raisons pour lesquelles on n'a pas de données annotées sont diverses. La raison principale, cependant, est qu'il est extrêmement difficile de définir rigoureusement ce qui compte comme une "bonne" réponse au problème (par opposition à une "mauvaise").

Je montrerai que le problème de la vérité dans la fiction, défini abstraitement ci-dessus, correspond à un problème d'apprentissage non-supervisé.

[1] Je considère que *Hamlet* est une fiction. Si vous doutez de cela, vous pourrez adapter l'exemple en choisissant votre fiction préférée.

[2] Voir (Woodward 2011) pour un panorama de cette controverse structurante en philosophie contemporaine de la fiction.

[3] Voir en particulier Yann Le Cun, Collège de France, cours du 15 avril 2016.

[4] Voir Yann Le Cun 2016 et Stéphane Mallat 2018 au Collège de France.

Références

Lewis, David (1978). "Truth in fiction". In: *American philosophical quarterly* 15.1, pp. 37–46.
Walton, Kendall (1990). *Mimesis as Make-believe: On the Foundations of the Representational Arts*. Harvard University Press.
Woodward, Richard (2011). "Truth in fiction". In: *Philosophy Compass* 6.3, pp. 158–167.

Courte bio-bibliographie

Louis Rouillé est entré à l'école normale supérieure rue d'Ulm en 2012 par le concours A/L option philosophie, après des années de prépas scientifique (MPSI) puis littéraire (HK/KH) à Rennes. Durant sa scolarité à l'ENS, il a obtenu un master de logique à Paris 1 (LoPhiSC) en 2014, le master de sciences cognitives ENS/EHESS/P5 (Cogmaster) en 2016, ainsi que l'agrégation de philosophie en 2015. Il se passionne très vite pour les questions théoriques sur la fiction. Après avoir consacré deux Masters recherche sur des problèmes logiques et linguistiques liés à la notion de fiction, il prépare une thèse de philosophie sur la fiction à l'ENS, sous la direction de François Récanati et Paul Égré à l'institut Jean Nicod. Sa thèse, débutée en septembre 2016 et soutenue en décembre 2019, s'intitule "Disagreeing about fiction". Il contribue dans cette thèse à des débats contemporains sur la vérité dans la fiction, les désaccords fictionnels et la référence dans la fiction. Les résultats de ses travaux de recherches doctorales font l'objet de publications à venir ou en cours. Il travaille notamment avec Guillaume Schuppert à une traduction de *Mimesis as Make-Believe* de Kendall Walton (Harvard University Press), éditrice du projet: Françoise Lavocat. Louis Rouillé est depuis janvier 2020 professeur de philosophie à Rouen, au lycée Blaise Pascal, et il enseigne la logique à l'université Paris-Ouest Nanterre.

Articles acceptés pour publication

(* examen du manuscrit à l'aveugle)

2020 **From fictional disagreements to thought experiments** (22 pages, 7900 mots)* – *Argumenta*, special issue "Fiction and Imagination as Grounds for Counterfactual Reasoning, Scientific Modeling, and Thought Experiments".

2020 **Anti-realism about fictional names at work: a new theory for metafictional sentences** (23 pages, 8600 mots)* – *Organon F*, special issue "Names and Fictions".

2018 **Notice sur** *Les Fondements de l'arithmétique* **de Frege** (9 pages, 4600 mots) – *Encyclopédie des Œuvres philosophiques* (Éditions ellipses), éditeurs: Elsa Ballanfat, Audrey Benoit, Clotilde Nouët et Jean-François Suratteau.

2018 **Notice sur** *De la pluralité des mondes* **de Lewis** (7 pages, 3600 mots) – *Encyclopédie des Œuvres philosophiques* (Éditions ellipses), éditeurs: Elsa Ballanfat, Audrey Benoit, Clotilde Nouët et Jean-François Suratteau.

2016 **Derrière les noms dans la fiction** (8 pages, 5700 mots) – *Le Pardaillan, N.1-Fictions Populaires*, éditrice: Luce Roudier.