Conférence Nationale d'Intelligence Artificielle Année 2024

Emmanuel Adam, Thomas Guyet, Benoit Le Blanc, Dominique Longin, Nadia Abchiche-Mimouni, Ghislain Atemezing, Nathalie Aussenac-Gilles, Jean-Guy Mailly, Catherine Roussey, François Schwarzentruber, et al.

Conférence Nationale d'Intelligence Artificielle Année 2024

Sections Spéciales
Nadia Abchiche-Mimouni
Ghislain Atemezing
Nathalie Aussenac-Gilles
Jean-Guy Mailly
Catherine Roussey
Francois Schwarzentruber
Anaëlle Wilczynski
Haifa Zargayouna

Coordination AFIA
Emmanuel Adam
Thomas Guyet
Benoit Le Blanc
Dominique Longin

AfIA
Association française pour l'Intelligence Artificielle

Conférence Nationale d'Intelligence Artificielle Année 2024

Actes CNIA 2024

Conférence Nationale d'Intelligence Artificielle

Éditeurs : Président(e)s des Comités de Programme de PFIA 2024

Tables des matières

La Rochelle, au cœur de l'Intelligence Artificielle de demain

Karell Bertet
Professeur, L3I, Université de La Rochelle
Président du Comité d'Organisation de la Plate-Forme IA 2024

Thomas Guyet, **Davy Monticolo**, **Ahmed Samet**
Chercheur INRIA Lyon, Professeur Université de Lorraine, Maître de Conférence INSA Strasbourg,
Membres du Conseil d'Administration de l'AFIA,
Co-Présidents du Comité de Programmation de la Plate-Forme IA 2024

L'AFIA et le laboratoire L3I s'associent pour organiser du 1 au 5 juillet 2024 la 17e Plate-Forme Intelligence Artificielle – PFIA 2024. Les précédentes éditions se sont tenues à Palaiseau (1999), Grenoble (2001), Laval (2003), Nice (2005), Grenoble (2007), Hammamet (2009), Chambéry (2011), Lille (2013), Rennes (2015), Caen (2017), Nancy (2018), Toulouse (2019), Angers (2020), Bordeaux (2021), Saint-Étienne (2022) et Strasbourg (2023).

La plate-forme IA constitue un point de rencontre unique pour la communauté IA permettant de rapprocher les différentes disciplines qui la composent et d'établir des passerelles entre elles. À cette intention, la Plate-forme IA s'adresse à l'ensemble de la communauté francophone en IA pour aborder des problématiques communes.

La Rochelle compte des acteurs de premier plan dans le domaine de l'Intelligence Artificielle, avec de réelles compétences tant dans son cœur scientifique et technique que dans son intégration à d'autres technologies ou dans ses applications dans différents domaines.

Pour son édition 2024, la Plate-Forme IA héberge les 5 conférences suivantes :

APIA	10e conférence nationale sur les Applications Pratiques de l'IA	G. Atemezing et C. Roussey
CNIA	27e Conférence Nationale d'Intelligence Artificielle	N. Aussenac-Gilles
IC	35es journées francophones d'Ingénierie des Connaissances	H. Zargayouna
JIAF	18es Journées d'Intelligence Artificielle Fondamentale	A. Wilczynski, J.-G. Mailly et F. Schwarzentruber
RJCIA	22es Rencontres des Jeunes Chercheurs en Intelligence Artificielle	N. Abchiche-Mimouni

les 3 journées suivantes :

Agents & IA	journée « Agent et IA », en partenariat avec le groupe ACIA et JFSMA
Santé & IA	8e journée « Informatique médicale et IA », en partenariat avec AIM
Société & IA	journée « Société et IA », en partenariat avec l'ACE du GDR RADIA

7 ateliers thématiques (Défense & IA, Jeux & IA, MAFTEC, SOSEM, CÉCILIA, IA en Nouvelle-Aquitaine, GDR RADIA),

ainsi que 8 tutoriels hébergés sur les thèmes :

- Reservoir Computing : théorie, intuitions et applications avec ReservoirPy (X. Hinaut, P. Bernard) ;
- Comment appréhender la problématique des biais avec les LLM (A. Névéol, M. Roche, R. Decoupes) ;
- Introduction au Physics Informed machine learning – Ajout de connaissance physique dans des modèles d'apprentissage machine (S. Ghidalia) ;
- Prise de décision précoce (A. Cornuéjols, A. Bondu, A. Renault, V. Lemaire) ;
- Machine Learning Meets Program Synthesis (N. Fijalkow) ;
- Some new directions for explainable AI (C. Hudelot, T. Fel, W. Ouerdane, A. Poché) ;
- An Introduction to Symbolic Explainability (J. Marques-Silva) ;
- Prédiction Conforme (J. Dalmau, M. Mendil).

Six collègues ont également été invités pour des conférences en début de chaque demi-journée :

Hedi Karray	Université de Technologie Tarbes Occitanie Pyrénées	France
Enrico Motta	UK's Open University	Royaume-Uni
Nathalie Nevejans	Université d'Artois	France
Simon Parsons	University of London	Royaume-Uni
Samuel Tronçon	Résurgences R&D et philosophe	France
Pierre Zweigenbaum	Université Paris-Saclay	France

plus les 3 lauréats du prix de thèse de l'AFIA :

Virginie Do	Université Paris Dauphine-PSL	France
Pierre Marion	Sorbonne Université	France
Yuan Yin	Sorbonne Université	France

Nous remercions les présidents et membres de comités de programme des conférences et journées hébergées, les orateurs, les membres des comités de programmation et d'organisation, nos partenaires institutionnels et industriels, ainsi que tous les participants, pour leurs contributions précieuses à la réussite de cette plate-forme.

L'ensemble des informations sur ces événements est disponible à partir du site de la Plate-Forme IA 2024 (https://pfia2024.univ-lr.fr).

Comité de programmation

Conseil d'Administration de l'Association Française pour l'Intelligence Artificielle

Président

— Thomas GUYET, AIstroSight, Inria, Centre de Lyon

Membres

— Davy Monticolo, Université de Lorraine, ERPI ;
— Ahmed Samet, INSA Strasbourg, ICube.

Comités de programme

Conférence nationale sur les Applications Pratiques de l'IA (APIA)

Présidence

— Ghislain Atemezing (ERA, Valenciennes) ;
— Catherine Roussey (MISTEA INRAE, Montpellier).

Membres

— Florence Amardeilh (Elzeard, Bordeaux) ;
— Fabien Amarger (Logilab, Toulouse) ;
— Nicolas Audebert (CNAM CEDRIC, Paris) ;
— Nathalie Aussenac-Gilles (IRIT, Toulouse) ;
— Alain Berger (Ardans, Montigny-le-Bretonneux) ;
— Sandra Bringay (LIRMM Université Paul Valéry Montpellier, Montpellier) ;
— Xavier Briottet (ONERA, Toulouse) ;
— Stéphan Brunessaux (Sensei Consult, Louviers) ;
— Patrice Buche (IATE INRAE, Montpellier) ;
— Davide Buscaldi (LIPN Université Sorbonne Paris Nord, Villetaneuse) ;
— Bruno Carron (Airbus Defence and Space, Grand Paris) ;
— Laurent Cervoni (Talan, Grand Paris) ;
— Caroline Chopinaud (Hub France IA, Paris) ;
— Gaël de Chalendar (CEA LIST, Saclay) ;
— Yves Demazeau (LIG CNRS, Grenoble) ;
— Sylvie Despres (LIMICS Université Sorbonne Paris Nord, Bobigny) ;
— Gayo Diallo (AHeaD ISPED Université de Bordeaux, Bordeaux) ;
— Valentina Dragos (Onera, Palaiseau) ;
— Guillaume Dubuisson Duplessis (EDF, Paris) ;
— Catherine Faron (I3S Université Côte d'Azur, Sophia Antipolis) ;
— Bernard Georges (Société Générale, Paris) ;
— Céline Hudelot (CentraleSupélec MICS, Gif-sur-Yvette) ;
— Dino Ienco (TETIS INRAE, Montpellier) ;

— Arnaud Lallouet (Huawei Technologies Ltd, Boulogne-Billancourt) ;
— Christine Largouët (IRISA, Rennes) ;
— Christelle Launois (Société Générale, Paris) ;
— Mustapha Lebbah (DAVID Université Paris-Saclay, Versailles) ;
— Dominique Lenne (HEUDIASYC Université de Technologie de Compiègne, Compiègne) ;
— Sylvain Mahé (EDF Recherche et Développement, Chatou) ;
— Gauthier Picard (ONERA) ;
— Céline Rouveirol (LIPN, Université Sorbonne Paris Nord) ;
— Françoise Soulié-Fogelman (Hub France IA, Paris) ;
— Élodie Thiéblin (Logilab, Toulouse) ;
— Brigitte Trousse (INRIA, Sophia Antipolis).

Conférence Nationale en Intelligence Artificielle (CNIA)

Présidence

— Sandra Bringay, Université Paul-Valéry Montpellier.

Membres

— Jérome Azé, Université de Montpellier, LIRMM ;
— Isabelle Bloch, Sorbonne Université, LTCI ;
— Olivier Boissier, Mines Saint-Etienne, LIMOS ;
— Robert Bossy, INRAE Centre de Jouy en Josas, MaIAGE ;
— Armelle Brun, Université de Lorraine, LORIA ;
— Cécile Capponi, Université Aix-Marseille, LIS ;
— Sylvie Coste-Marquis, Université d'Artois, CRIL ;
— Benjamin Dalmas, Centre de Recherche Informatique de Montréal ;
— Yves Demazeau, CNRS, LIG ;
— Sébastien Destercke, HDS, Université Technologique de Compiègnes, Heudiasyc ;
— Arnaud Doniec, IMT Lille Douai ;
— Jérôme Euzenat, INRIA Alpes, LIG ;
— Jean-Gabriel Ganascia, Sorbonne Université, LIP6 ;
— Éric Gaussier, Université Grenoble Alpes et IUF, LIG ;
— Guillaume Gravier, CNRS, IRISA ;
— Andreas Herzig, CNRS, IRIT ;
— Nathalie Hernandez, Université Toulouse 2, IRIT ;
— Céline Hudelot, Ecole Centrale de Paris ;
— Camille Kurtz, Université Paris Cité ;
— Nicolas Lachiche, Université de Strasbourg ;
— Frederique Laforest, INSA Lyon, LIRIS ;
— Florence Le Ber, École Nationale du Génie de l'Eau et de l'Environnement de Strasbourg, ICUBE ;
— Philippe Lenca, IMT Atlantique ;
— Marie-Jeanne Lesot, Sorbonne Université, LIP6 ;
— Pascal Poncelet, Université de Montpellier, LIRMM ;
— Catherine Roussey, INRAE Centre Occitanie-Montpellier, MISTEA ;
— Pascale Sébillot, INSA Rennes, IRISA ;
— Nazha Selmaoui, Université de la Nouvelle Calédonie, ISEA ;

— Laurent Vercouter, INSA Rouen Normandie, LITIS ;
— Bruno Zanuttini, Université de Caen Normandie, GREYC.

Journées francophones d'Ingénierie des Connaissances (IC)

Présidence

— Haïfa Zargayouna LIPN, Université Sorbonne Paris Nord.

Membres

— Marie-Hélène Abel HEUDIASYC, Université de Technologie de Compiègne ;
— Nathalie Abadie LASTIG, Univ. Gustave Eiffel, IGN-ENSG ;
— Xavier Aimé Cogsonomy / LIMICS UMRS 1142 Inserm ;
— Yamine Ait-Ameur IRIT, Université de Toulouse, Toulouse INP ;
— Nathalie Aussenac-Gilles IRIT, Université de Toulouse, CNRS ;
— Bruno Bachimont Sorbonne Université ;
— Nathalie Bricon-Souf IRIT, Université de Toulouse, UT3 ;
— Sandra Bringay LIRMM ;
— Patrice Buche INRA ;
— Davide Buscaldi LIPN, Université Sorbonne Paris Nord ;
— Sylvie Calabretto LIRIS ;
— Pierre-Antoine Champin LIRIS, Université Claude Bernard Lyon1 ;
— Jean Charlet AP-HP & INSERM UMRS 1142 ;
— Victor Charpenay Mines Saint-Etienne, Univ Clermont Auvergne, INP Clermont Auvergne, CNRS, UMR 6158 LIMOS ;
— Jérôme David INRIA ;
— Sylvie Despres Laboratoire d'Informatique Médicale et de BIOinformatique (LIM&BIO) ;
— Gayo Diallo ISPED & LABRI, University of Bordeaux ;
— Gilles Falquet University of Geneva ;
— Catherine Faron Université Côte d'Azur ;
— Béatrice Fuchs LIRIS, université de Lyon ;
— Frédéric Fürst MIS - Université de Picardie - Jules Verne ;
— Jean-Gabriel Ganascia Pierre and Marie Curie University - LIP6 ;
— Mounira Harzallah LS2N, University of Nantes, France ;
— Nathalie Hernandez IRIT, Université de Toulouse, UT2 ;
— Liliana Ibanescu Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France ;
— Sébastien Iksal LIUM - Le Mans Université, France ;
— Antoine Isaac Europeana & VU University Amsterdam ;
— Khadija Jradeh IRIT, Université Toulouse Jean Jaurès ;
— Clément Jonquet MISTEA (INRAE) and LIRMM (U. Montpellier) ;
— Gilles Kassel University of Picardie Jules Verne ;
— Michel Leclère University of Montpellier (LIRMM/INRIA), France ;
— Maxime Lefrançois MINES Saint-Etienne ;
— Dominique Lenne Heudiasyc, Université de Technologie de Compiègne ;
— Jérôme Nobécourt LIMICS ;

— Nathalie Pernelle LIPN, Université Sorbonne Paris Nord ;
— Yannick Prié LINA - University of Nantes ;
— Cédric Pruski Luxembourg Institute of Science and Technology ;
— Sylvie Ranwez LGI2P / Ecole des mines d'Alès ;
— Catherine Roussey INRAE ;
— Fatiha Saïs, LISN, CNRS & Université Paris Saclay
— Karim Sehaba LIRIS CNRS ;
— Konstantin Todorov LIRMM / University of Montpellier ;
— Cassia Trojahn dos Santos IRIT, Université Toulouse Jean Jaurès ;
— Raphaël Troncy EURECOM.

Journées d'Intelligence Artificielle Fondamentale & Journées Francophones sur la Planification, la Décision et l'Apprentissage (JIAF-JFPDA)

Présidence

— Jean-Guy Mailly (IRIT, Université de Toulouse, UT Capitole) ;
— François Schwarzentruber (IRISA, ENS Rennes, Université de Rennes) ;
— Anaëlle Wilczynski (MICS, CentraleSupélec, Université Paris-Saclay).

Membres

— Francesco Belardinelli (Imperial College London) ;
— Aurélie Beynier (LIP6, Sorbonne Université) ;
— Elise Bonzon (LIPADE, Université Paris Cité) ;
— Olivier Buffet (INRIA / LORIA) ;
— Martin Cooper (IRIT, Université de Toulouse, UT3) ;
— Célia da Costa Pereira (Université Côte d'Azur) ;
— Sylvie Coste-Marquis (CRIL, Université d'Artois) ;
— Tiago de Lima (Université d'Artois, CRIL CNRS) ;
— Jilles Dibangoye (INSA Lyon, CITI lab, INRIA) ;
— Sylvie Doutre (IRIT, Université de Toulouse, UT Capitole) ;
— Florence Dupin de Saint-Cyr (IRIT, Université de Toulouse, UT3) ;
— Alain Dutech (Loria - Inria) ;
— Hugo Gilbert (Lamsade - Université Paris Dauphine) ;
— Andreas Herzig (IRIT, Université de Toulouse, CNRS) ;
— Sébastien Konieczny (CRIL - CNRS) ;
— Raida Ktari (Aix-Marseille Université) ;
— Jérôme Lang (CNRS, LAMSADE, Université Paris-Dauphine) ;
— Daniel Le Berre (CRIL, Université d'Artois) ;
— Jean Lieber (LORIA - INRIA Lorraine) ;
— Pierre Marquis (CRIL, U. Artois & CNRS - Institut Universitaire de France) ;
— Amedeo Napoli (LORIA Nancy, CNRS - Inria - Université de Lorraine) ;
— Arianna Novaro (Centre d'Economie de la Sorbonne (CES), Université Paris 1 Panthéon-Sorbonne) ;
— Damien Pellier (Laboratoire d'Informatique de Grenoble) ;

— Laurent Perrussel (IRIT, Université de Toulouse, UT Capitole) ;
— Sophie Pinchinat (IRISA Rennes) ;
— Julien Rossit (Université Paris Cité, LIPADE) ;
— Stephanie Roussel (ONERA) ;
— Régis Sabbadin (INRA-UBIA) ;
— Vincent Thomas (LORIA) ;
— Bruno Zanuttini (GREYC, Université de Normandie).

Rencontres des Jeunes Chercheurs en Intelligence Artificielle (RJCIA)

Présidence

— Nadia Abchiche-Mimouni, I3S, CNRS/UNS - Université côte d'Azur.

Membres

— Amel Bouzeroub, professeur, Institut Polytechnique de Paris - Telecom SudParis ;
— Feng Chu, professeur IBISC, univ. Evry - université Paris-Saclay ;
— Victor David, Researcher, INRIA Sophia Antipolis ;
— Maxime Devanne, Maître de conférences, IRIMAS - Université de Haute-Alsace ;
— Catherine Faron, professeur, I3S - université Côte d'Azur ;
— Arnaud Ferre, Chargé de Recherche, MaIAGE, INRAE - Université Paris-Saclay ;
— Maxime Folschette, Maître de conférences, CRIStAL - Centrale Lille Institut ;
— Fatima Ghedjati, Maitre de conférences, LICIIS - université Université de Reims Champagne-Ardenne ;
— Zahia Guessoum, Maître de conférences HDR, CReSTIC - université Université de Reims Champagne-Ardenne ;
— Camille Guinaudeau, Maitre de conférences, Japanese-French Laboratory for Informatics CNRS - Université Paris-Saclay ;
— Guillaume Lozenguez, Maitre de conférences, IMT Nord-Europe ;
— Jean-Guy Mailly, professeur junior, IRIT, Université de Toulouse, UT Capitole ;
— Mohamed-Lamine Messai, Maître de conférences, ERIC - Université Lyon 2 ;
— Pierre Monnin, Junior Fellow, Université Côte d'Azur, Inria, CNRS, I3S ;
— Charlotte Pelletier, Maître de conférences, IRISA - université Bretagne Sud ;
— Brian Ravenet, Maître de conférences, LISN - université Paris-Saclay ;
— Yasmina Sadi, Maître de conférences, IBISC - univ. Evry - université Paris-Saclay ;
— Nicolas Verstaevel, Maître de conférences, IRIT - université Toulouse Capitole ;
— Genane Youness, Maitre de conférences, CESI - LINEACT ;
— Farida Zehraoui, Maître de conférences, IBISC, univ. Evry - université Paris-Saclay.

Comité d'organisation

ICube, Université de Strasbourg

Présidence

— Karell Bertet.

Pilotage

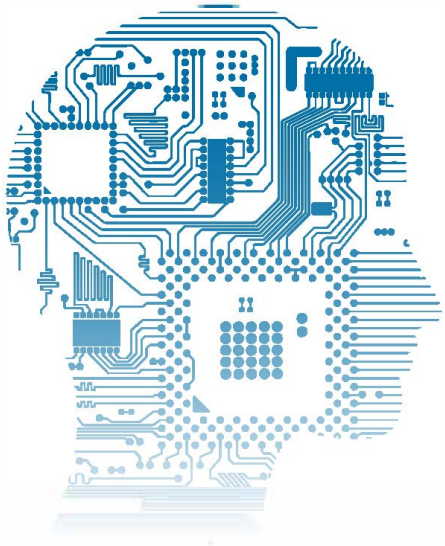
— Karell Bertet, responsable de la restauration et de l'accueil ;
— Christophe Demko, responsable du site web ;
— Cyril Faucher, responsable du budget et des partenaires ;
— Annick Lassus, responsable de l'atelier IA en Nouvelle-Aquitaine ;
— Damien Mondou, responsable de la communication et de la programmation.

Membres

— Laëtitia Barreau ;
— Thomas Chambon ;
— Mickaël Coustaty ;
— Salah Eddine Elgharbi ;
— Patrick Franco ;
— Petra Gomez ;
— Chloé Goudounesque ;
— Jean-Loup Guillaume ;
— Ahmed Hamdi ;
— Marwa Hamdi ;
— Axel Jean-Caurant ;
— Dominique Limousin ;
— Mélanie Malinaud ;
— Mourad Rabah ;
— Jérémy Richard ;
— Rémy Rondet ;
— Guillaume Savarit ;
— Anaïs Schmitt ;
— Cyrille Suire ;
— Martin Waffo Kemgne ;
— Rouaa Wannous.

PFIA 2024

Plate-forme Intelligence Artificielle



AfIA
Association française pour l'Intelligence Artificielle

Informatique Image Interaction
La Rochelle Univ.

La Rochelle Université

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

MAIF

ENGIE

ARDANS
Knowledge Consulting & Software Solutions

SOCIETE GENERALE

DARVA

néosoft

SERLI
Innovation - Ingénierie - Formation

GDR RADIA

ACQUIRE DECIDE

Cat-Amania
CONNECTONS NOS TALENTS

Communauté d'Agglomération de La Rochelle

odrive

Udi
L'INSTITUT Littoral Urbain Durable Intelligent

it
LA ROCHELLE

dihnamic

SemOntoMap : une méthode hybride pour l'annotation sémantique de textes cliniques en psychiatrie

O. Aouina[1], J. Hilbey[1,2], J. Charlet[1,2]

[1] Sorbonne Université, Inserm, Université Sorbonne Paris-Nord, Laboratoire d'informatique médicale et d'ingénierie des connaissances en e-santé, LIMICS, Paris, France

[2] Assistance Publique-Hôpitaux de Paris, Paris, France

correspondant : ons.aouina@etu.sorbonne-universite.fr

Résumé

Les descriptions en texte libre contenues dans les dossiers patients informatisés (DPI) revêtent un intérêt significatif pour la recherche clinique et l'optimisation des soins. Toutefois, la capacité des ordinateurs à interpréter directement ce texte libre est limitée, réduisant ainsi sa valeur potentielle. Bien que l'annotation sémantique offre une solution pour rendre le texte libre des DPI interprétable par les machines, elle rencontre des obstacles majeurs lorsqu'elle est appliquée aux ontologies de domaine spécifiques, particulièrement en français. Ces difficultés sont encore plus marquées dans le domaine psychiatrique où l'on cherche non seulement à extraire les concepts du domaine mais à les normaliser et à extraire les relations de textes décrivant longuement l'histoire d'une maladie d'un patient et ses ascendants. Face à ces enjeux, nous proposons un système fondé sur des techniques d'apprentissage non supervisé pour extraire les entités et leurs interrelations en utilisant une ontologie de domaine. Ce système est évalué dans le cadre du projet PsyCARE sur un échantillon de 60 comptes rendus analysés par deux évaluateurs.

Mots-clés

annotation sémantique, ontologie, plongement de l'ontologie, apprentissage automatique non supervisé, TALN, BERT, Word2Vec

Abstract

The free-text descriptions contained in Electronic Health Records (EHR) hold significant interest for clinical research and the optimization of care. However, computers' ability to directly interpret this free text is limited, thereby reducing its potential value. While semantic annotation offers a solution to make the free text of EHRs machine-interpretable, it faces major obstacles when applied to specific domain ontologies, particularly in French. These difficulties are even more pronounced in the psychiatric field, where there is an attempt to extract domain concepts relations from texts that extensively describe a patient's disease history and their ancestors. Faced with these challenges, we propose a system based on unsupervised learning techniques to extract entities and their interrelations using a domain ontology. This system is evaluated within the framework of the PsyCARE project on a sample of 60 reports analyzed by two evaluators.

Keywords

Semantic annotation, ontology embedding, unsupervised machine learning, Ontology, NLP, BERT, Word2Vec

1 Introduction

Dans le domaine de la recherche biomédicale et des soins aux patients, les documents techniques, et plus spécifiquement les textes biomédicaux, sont cruciaux. Ces documents, sont indispensables pour faire avancer les pratiques cliniques et stimuler l'innovation en santé. La complexité et la richesse de ces textes ont mené à l'utilisation de diverses ontologies biomédicales, telles que l'Ontologie des Gènes (GO) et la Nomenclature Systématisée de Médecine – Termes Cliniques (SNOMED CT), marquant des efforts significatifs pour structurer cette information vitale et améliorer son accessibilité [34]. L'annotation sémantique, qui associe le texte à des balises significatives issues de ces ontologies, joue un rôle clé dans de nombreuses applications, renforçant l'interopérabilité et l'efficacité de la récupération d'informations [27].

L'effort d'annotation sémantique varie du manuel au totalement automatisé, exploitant les avancées en traitement automatique du langage naturel (TALN) [25]. Notre étude se concentre sur l'annotation automatique de textes cliniques, tâche rendue complexe par la nature du langage médical. Une attention particulière est portée aux sections narratives des dossiers cliniques, surtout dans les résumés de sortie en psychiatrie (dans notre cas des comptes rendus d'hospitalisation ou CRH), qui recèlent des informations sur les événements cliniquement significatifs affectant la trajectoire médicale du patient. Ces informations incluent les antécédents familiaux, l'historique de la maladie, les traitements prescrits, ainsi que les relations temporelles entre ces événements. Des questions comme « comment la maladie a évolué chez le patient ? » ne peuvent être interprétées et on ne peut y répondre que si l'on prend en compte le contexte complet des antécédents du patient et les relations temporelles entre les différents concepts repérés. Ce problème est

abordé en psychiatrie par le projet RHU PsyCARE [1] qui vise à améliorer l'intervention précoce dans la psychose en fournissant des outils pour faciliter l'accès aux soins et offrir des programmes de traitement personnalisés.
Notre travail vise donc à annoter sémantiquement des CRH, en capturant les segments textuels qui correspondent à des ontologies ou des terminologies standardisées mais aussi en déchiffrant les modalités, les relations temporelles et les informations détaillées sur les antécédents et l'évolution de la psychose. Dans cet article, nous proposons une méthode d'annotation sémantique des CRH fondée d'abord sur une ontologie développée dans le cadre de PsyCARE. Cette ontologie est combinée avec des modèles de langue et des algorithmes d'apprentissage pour construire un modèle formel précis du texte [15].

2 Contexte

Les CRH sont rédigés par des professionnels de la santé divers et aux styles d'écriture variés. Le traitement efficace de ces documents est essentiel pour ces derniers qui doivent parcourir d'importants volumes de dossiers médicaux électroniques pour dégager les informations clés. La normalisation des entités nommées joue un rôle crucial dans la réduction de l'ambiguïté des comptes rendus cliniques. Néanmoins, ces étapes peuvent aboutir à l'extraction de concepts redondants ou de faible valeur informative. Face à ce défi, le développement des méthodes d'extraction d'information (EI) non supervisées devient une évidence. Ces algorithmes permettent d'identifier des informations significatives sans dépendre des corpus préalablement annotés, offrant ainsi une réponse efficace aux contraintes des approches traditionnelles [22].
Parallèlement, la normalisation des entités, également connue sous les termes de désambiguïsation ou de liaison d'entités, joue un rôle crucial dans l'extraction d'informations. Cette démarche consiste à associer les mentions d'entités présentes dans le texte avec des catégories ou des concepts issus d'un vocabulaire de référence [28] ou d'une ontologie spécifique, ce qui permet d'uniformiser la représentation de ces mentions. Pour améliorer l'efficacité de la normalisation des entités, certaines recherches proposent d'intégrer des données concernant la structure des graphes de connaissances [24], tandis que d'autres études mettent en avant les bénéfices de combiner les plongements de mots et d'entités afin de créer des connexions significatives entre les entités. Ces approches visent à renforcer les performances de cette tâche en exploitant les relations sémantiques profondes [24]. Dans ce contexte, l'*embedding* (ou plongement) d'ontologies représente un domaine de recherche prometteur. P. Devkota *et al.* [9] montre que l'intégration d'informations issues du plongement d'ontologies peut significativement affiner la détection des concepts ontologiques dans la littérature scientifique, renforçant ainsi la concordance sémantique entre les informations textuelles et les structures ontologiques [3].
Dans cette section, nous explorons les techniques d'EI qui, dans notre contexte, concernent l'extraction de syntagmes, y compris les syntagmes nominaux (SN) et les syntagmes verbaux (SV) ainsi que le plongement d'ontologies pour structurer et enrichir les connaissances extraites.

1. https://psy-care.fr/

2.1 Extraction de syntagmes

Pour l'extraction de syntagmes, nous adoptons l'hypothèse selon laquelle les syntagmes correspondent à une liste de N-grammes, soit des séquences de n mots manifestant une structuration grammaticale particulière. Cette tâche consiste à déterminer un ensemble de séquences de mots qui encapsulent les thèmes centraux ou les idées présentées dans un document, offrant un aperçu de son contenu le plus critique. Ces algorithmes sont classés en méthodes supervisées [32] et non supervisées [31]. Compte tenu de la polyvalence et de l'applicabilité générale des méthodes non supervisées, se concentrant sur les attributs inhérents du texte pour l'extraction de syntagmes, notre proposition se concentre sur l'extraction non supervisée. Trois méthodes principales se distinguent dans ce domaine :

Méthodes fondées sur les Graphes. Ces méthodes convertissent le document en un graphe et classent les phrases candidates dans le graphe [21, 31] . Les nœuds correspondent à des éléments textuels tels que les mots ou les phrases, et les arêtes reflètent les liens entre eux, par exemple la co-occurrence ou la similarité sémantique. Cette approche permet d'évaluer l'importance des phrases candidates en fonction de leur position et de leurs connexions au sein du graphe. En exploitant les relations contextuelles entre les éléments textuels, ces méthodes se distinguent par leur capacité à identifier avec précision les syntagmes clés qui sont directement liés aux thèmes centraux du document. Ainsi, elles améliorent la pertinence et l'efficacité des systèmes de récupération d'information en facilitant l'identification de syntagmes essentiels qui récapitulent de manière efficace le contenu central du texte.

Méthodes Statistiques. Ces méthodes, telles que YAKE [5], sont fondées sur TF-IDF (Fréquence du Terme - Inverse de la Fréquence des Documents), TextRank [21] ou SingleRank [37]. Ces méthodes analysent les propriétés de distribution des mots et des phrases dans un texte par rapport au corpus de travail pour identifier des phrases clés et mettre en évidence des termes qui sont spécifiques et informatifs du contenu du document. L'importance de combiner des analyses statistiques avec des informations contextuelles est spécialement mise en avant par YAKE qui se distingue par son utilisation de métriques statistiques avancées pour saisir le contexte et la dispersion des termes à travers le document. Mais bien que ces méthodes soient efficaces d'un point de vue computationnel et simples à mettre en œuvre, elles ne capturent pas toujours la richesse sémantique du texte, limitant potentiellement leur efficacité dans certains contextes de récupération d'informations.

Méthodes fondées sur l'apprentissage profond. Ces méthodes exploitent les réseaux neuronaux pour apprendre des représentations du texte qui capturent des relations et des motifs sémantiques. Des approches non supervisées

d'apprentissage profond telles que les auto-encodeurs ou les modèles fondés sur les transformateurs comme BERT [10], peuvent modéliser implicitement l'importance des syntagmes. Parmi celles-ci, KeyBERT [12] tire parti des modèles tels BERT, pour identifier de manière efficace les syntagmes clés dans les textes. KeyBERT combine la capacité des transformateurs à comprendre le contexte profond du texte avec une approche ciblée pour l'extraction des phrases clés, permettant ainsi une identification précise et contextuellement riche des informations clés contenues dans les documents. PatternRank [26] s'appuie sur des modèles de langage et des parties de discours (PoS) pré-entraînés pour l'extraction non supervisée de phrases-clés à partir de documents uniques. Cet algorithme représente l'état de l'art dans l'extraction de phrases clés, grâce à son intégration de modèles de partie du discours pour la sélection des phrases candidates, permettant ainsi son adaptation à divers domaines. Cette approche permet une granularité et une précision accrues dans l'extraction des phrases clés, en se basant sur des critères syntaxiques spécifiques pour identifier les éléments les plus informatifs du texte. Dans notre approche, nous avons adapté PatternRank pour améliorer sa capacité à capturer des syntagmes nominaux et verbaux, en exploitant le *part of speech*, augmentant la précision de notre méthode.

2.2 Plongement des ontologies

Les modèles de plongement de graphe de connaissances (Knowledge Graph Embedding ou KGE) sont utilisés pour la transformation des vastes réseaux complexes d'entités et de relations au sein d'un graphe de connaissances en des espaces vectoriels de faible dimension, ainsi gérables [8]. L'essence du KGE réside dans sa capacité à transformer des informations complexes et de haute dimension d'un graphe de connaissances – comprenant diverses entités et les relations à facettes multiples entre elles – en une forme à la fois efficace sur le plan du calcul et sémantiquement riche.

Plusieurs modèles pour KGE, tels que DistMult [39] et RotatE [33], ont été proposés pour relever ces défis, montrant de bons résultats sur des ensembles de données de graphes de connaissances à usage général comme FB15K-237 [35]. Cependant, leur efficacité dans des domaines spécialisés, tels que la médecine, peut ne pas être aussi satisfaisante en raison de difficultés liées à la représentation et au raisonnement autour des entités et relations médicales [11]. Les méthodes existantes ne capturent pas adéquatement les relations complexes, les structures hiérarchiques, et l'hétérogénéité des entités médicales, ni n'abordent les problèmes de données bruyantes, incomplètes et la haute dimensionnalité souvent rencontrés dans les graphes de connaissances médicales.

Dans ce contexte, le plongement d'ontologies se présente comme une approche prometteuse, complétant le KGE. Axée sur la modélisation des relations directes entre entités, le plongement d'ontologies utilise la richesse sémantique et la structure logique des ontologies [19] . Cette méthode permet de capturer non seulement les relations entre entités mais aussi les concepts abstraits, les hiérarchies de classes et les axiomes qui structurent les connaissances dans un domaine spécifique.

L'intégration des techniques de prédiction par apprentissage automatique et d'analyse statistique des ontologies gagne en popularité et des méthodes pour plonger la sémantique des ontologies OWL commencent à émerger dans la littérature. Contrairement aux graphes de connaissances, les ontologies OWL ne se limitent pas à une structure graphique mais incorporent également des constructeurs logiques, et les entités sont souvent enrichies d'informations lexicales détaillées, spécifiées via $rdfs{:}label$, $rdfs{:}comment$ et de nombreuses autres propriétés d'annotation personnalisées ou intégrées. Dans cette approche, le but du plongement d'ontologie OWL est de représenter chaque entité nommée OWL (classe, instance ou propriété) par un vecteur, de manière à conserver dans l'espace vectoriel les relations inter-entités indiquées par les informations mentionnées ci-dessus et à maximiser la performance des tâches en aval où les vecteurs d'entrée peuvent être considérés comme des caractéristiques apprises.

EL Embedding [18] et Quantum Embedding [16] sont deux algorithmes de plongement d'ontologie OWL. Ils élaborent des fonctions de score et des fonctions de perte spécifiques pour les axiomes logiques issus respectivement d'EL++ et d'ALC, en transformant les relations logiques en relations géométriques. Cela encode la sémantique des constructeurs logiques mais néglige la sémantique supplémentaire apportée par les informations lexicales de l'ontologie. De plus, bien que la structure graphique soit explorée en considérant les axiomes de sous-classement et d'appartenance à une classe, l'exploration reste incomplète car elle se limite uniquement aux arêtes $rdfs{:}subClassOf$ et $rdf{:}type$ et ignore les arêtes impliquant d'autres relations.

Onto2Vec [29] et OPA2Vec [30] sont deux algorithmes de plongement d'ontologie utilisant le paradigme du plongement de mots, fondés sur l'architecture skip-gram ou CBOW. Onto2Vec utilise les axiomes d'une ontologie comme corpus pour l'entraînement, tandis qu'OPA2Vec enrichit le corpus d'Onto2Vec avec les informations lexicales fournies par, par exemple, $rdfs{:}comment$. Les deux méthodes adoptent la fermeture déductive d'une ontologie avec un raisonnement par inférence. Les deux méthodes traitent chaque axiome comme une phrase, ce qui signifie qu'elles ne peuvent pas explorer la corrélation entre les axiomes. Cela rend difficile l'exploration complète du graphe et de la relation logique entre les axiomes, et peut également conduire à un problème de pénurie de corpus pour les ontologies de petite à moyenne échelle.

OWL2Vec* [6] propose une solution aux limitations des approches précédentes en enrichissant leur corpus d'axiomes avec des données générées par des parcours sur des graphes RDF issus de la transformation des ontologies OWL. Cette approche prend en compte à la fois le graphe et les constructeurs logiques de l'ontologie. En outre, OWL2Vec* maximise l'exploitation des informations lexicales en créant des plongements non seulement pour les entités de l'ontologie mais également pour les termes lexicaux. Ainsi, OWL2Vec* condense efficacement les informations séman-

tiques et structurelles d'une ontologie dans un espace vectoriel compact, facilitant l'utilisation de ces données par des algorithmes d'apprentissage automatique pour des tâches en aval.
Le cadre d'OWL2Vec* est structuré autour de deux étapes clés comme illustré dans la figure 1 : (i) l'extraction d'un corpus à partir de l'ontologie, et (ii) l'entraînement d'un modèle de plongement de mots avec ce corpus. Ce corpus se compose de trois documents distincts : un document de structure, un document lexical, et un document combiné. Les deux premiers documents sont conçus pour explorer la structure de l'ontologie, ses constructeurs logiques et ses informations lexicales, permettant ainsi l'activation du raisonnement par inférence. Le troisième document vise à maintenir la corrélation entre les entités (IRIs) et leurs étiquettes lexicales (mots), en utilisant le premier document comme base tout en intégrant les informations lexicales disponibles de l'ontologie.

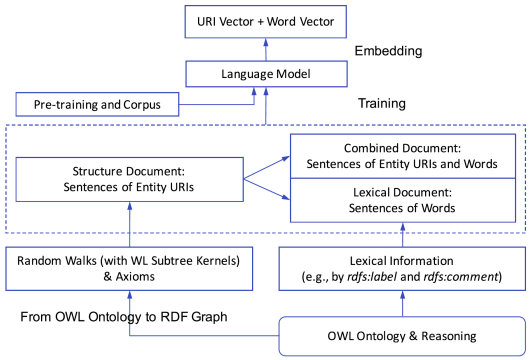


FIGURE 1 – Le contexte général d'OWL2Vec* Source : [6]

3 Système d'annotation

Dans cette section, nous présentons l'architecture du système, SemOntoMap, conçu pour enrichir les CRH de psychiatrie avec des annotations sémantiques. Notre approche s'appuie sur un corpus de textes de psychiatrie non annotés et une ontologie dédiée à ce domaine, visant à structurer ces documents.
Comme le montre la figure 2, la tâche d'annotation sémantique se déroule en trois grandes étapes : le prétraitement des textes, l'identification et la normalisation des entités nommées et l'extraction des relations entre ces entités.

3.1 Jeu de données en psychiatrie

Les documents cliniques exploités dans cette étude sont une compilation de près de 8000 CRH s'étendant sur une période de dix ans, totalisant environ 3,5 millions de mots. Ces CRH ont été collectés au sein du Groupe Hospitalier Universitaire Psychiatrie et Neurosciences de Paris. Ils sont semi-standardisés, en format Word et ont été pseudo-anonymisés au préalable, en remplaçant tous les noms, dates, lieux, etc. Chaque document se conclut par un diagnostic formulé selon la Classification Internationale des Maladies, version 10 (CIM-10 [2]). Rédigés en français, ces comptes rendus fournissent un aperçu détaillé de l'historique et du contexte social des patients, des prescriptions médicamenteuses, des circonstances d'admission à l'hôpital ainsi que les diagnostics psychiatriques actuels et antérieurs. Pour les besoins de l'annotation sémantique, une sélection de 30 comptes rendus a été annotée d'une façon aléatoire, en se basant sur le code CIM-10. Cette méthode de sélection vise à garantir une diversité dans les cas cliniques étudiés, couvrant un large éventail de diagnostics psychiatriques.

3.2 Ontologie de domaine

L'ontologie utilisée dans le processus d'annotation, appelée par la suite OntoPSY est une version fusionnée des modules ontoDOPSY, ontoMEDPSY, ontoDOME et ontoPOF de l'ontologie développée dans le cadre de PsyCARE [3] pour l'intégration des données et leur annotation sémantique. Ces modules contiennent les branches d'intérêt tels que les aspects cliniques psychiatriques (signes, symptômes, troubles psychiatriques), les médicaments identifiés par leur code ATC, des éléments relatifs à l'imagerie ainsi qu'une dimension temporelle pour représenter les connaissances médicales de manière adéquate. À partir de cette base, un schéma d'annotation est construit. Une branche de l'ontologie dédiée à la structure des CRH est ajoutée pour lier les concepts à leur contexte d'apparition dans le document, c'est-à-dire la section dans laquelle ils sont repérés [13] (« Histoire de la maladie », « Traitement de sortie », etc.) . En plus de décrire les aspects cliniques et les entités médicales, l'ontologie détaille les relations entre les différents concepts, enrichissant ainsi notre compréhension des interactions et des liens au sein des données cliniques.

3.3 Prétraitement des données textuelles

Ce processus implique le traitement du format du document et l'extraction de segments textuels pertinents à partir du document source, tout en écartant les balises et les éléments non pertinents. À ce stade, une analyse TALN de base est réalisée, incluant la tokenisation, la normalisation, et le marquage morphosyntaxique (*part-of-speech tagging*, POS). Le produit de cette phase est un texte brut enrichi de certaines annotations. Les algorithmes employés lors de cette étape ont été décrits dans un article antérieur [1].

3.4 Reconnaissance d'entités et normalisation

Dans cette section, nous détaillons les différentes étapes consacrées à l'extraction des candidats pour la reconnaissance des entités nommées (REN) ainsi qu'à leur normalisation en concordance avec les concepts de l'ontologie.

3.4.1 Extraction de syntagmes

L'importance des SN dans l'analyse des textes médicaux et psychiatriques est soulignée par les travaux de chercheurs comme Liu et al. [14] qui mettent en évidence

2. https://icd.who.int/browse10/2019/en
3. Cette ontologie sert à plus de processus que le seul TALN ; elle sert en particulier de modèle d'interopérabilité général pour le projet [13].

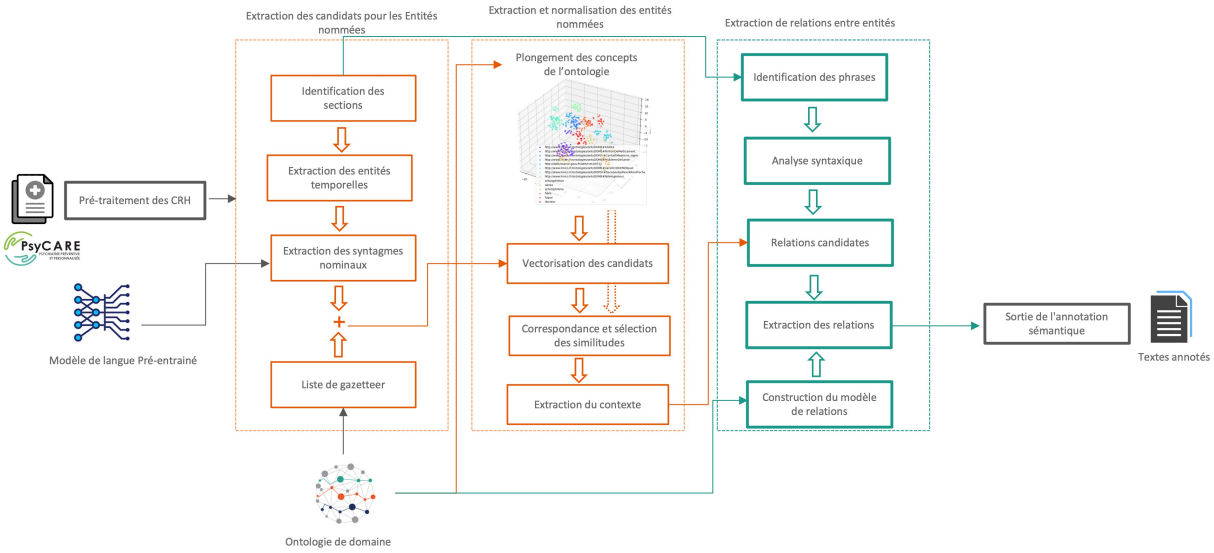


FIGURE 2 – Architecture du système d'annotation sémantique proposé.

la valeur de l'identification précise des termes médicaux pour améliorer l'accès à l'information dans les documents cliniques. Comme mentionné précédement, nous modifions PatternRank (Cf. sec. 2.1) pour l'adapter à la complexité narrative de ces documents. Les étapes sont détaillées dans la figure 3. Cette approche implique la segmentation du texte, le marquage syntaxique (POS), et la sélection des syntagmes qui répondent à des critères spécifiques (e.g., <(NN.|JJ|ADV)+><NN.> pour identifier des séquences commençant par un nom, adjectif, ou adverbe suivis de noms). Par la suite, les similarités cosinus entre la représentation du document et les représentations des syntagmes candidats sont calculées et ces derniers sont classées en ordre décroissant en fonction des scores trouvés. Les candidats sont à la fin transformés en vecteurs et classés par similarité cosinus avec le document pour extraire les termes les plus significatifs.

Dans nos expériences, nous utilisons le modèle de langage pré-entraîné Sentence-CamemBERT-Large développé par La Javaness [4]. Il s'agit d'un modèle SBERT qui a démontré sa capacité à produire de bonnes représentations textuelles pour des tâches de similarité sémantique.

3.4.2 Extraction d'information temporelles, medicaments et dosage

L'extraction des informations temporelles, médicamenteuses et de dosages est essentielle pour le suivi clinique et l'évaluation des traitements des patients. De nombreux travaux se sont concentrés sur ces extractions, notamment en ce qui concerne la temporalité et dans le domaine plus large du TALN médical [4]. Nous avons adopté la solution proposée par Aumiller Dennis [2] pour la reconnaissance et la normalisation des expressions temporelles, en combinant les capacités des librairies HeidelTime et SUTime pour l'identification complète des expressions temporelles dans les textes, car elles couvrent dates, heures, fréquences, et durées, et permettent l'ajustement de la date de référence pour une interprétation contextuelle. HeidelTime est utilisée pour son efficacité dans l'extraction temporelle de narrations non cliniques, adapté ici aux contextes cliniques. SUTime, en complément, offre la flexibilité d'une date de référence, utile pour notre analyse documentaire. Nous intégrons également le Temporal Tagger Service pour une détection précise de ces informations temporelles. Pour l'extraction des mentions de médicaments dans les comptes rendus hospitaliers en psychiatrie, EDS-NLP [36], développé par l'AP-HP, est employé pour sa spécialisation dans le traitement des données de santé en français. Enfin, le *GATE Tagger* [5] est utilisé pour identifier dosages et unités, facilitant l'interprétation des prescriptions.

3.5 Correspondance entre les informations extraites et les concepts de l'ontologie

Plongement de OntoPSY. Afin de produire le plongement sémantique de l'ontologie OntoPSY, nous avons mis en œuvre l'outil OWL2Vec* (voir Section 2.2). Cet outil a été configuré pour se servir d'un modèle Word2Vec préalablement entraîné sur un corpus diversifié, comprenant des articles de Wikipédia en français, des textes biomédicaux, ainsi que des corpus spécialisés [17]. Le modèle a ensuite été finement ajusté pour s'aligner avec les spécificités de l'ontologie, dont le prétraitement a été détaillé dans une publication antérieure [1]. Le réglage fin du modèle avec le corpus de l'ontologie a été réalisé à travers des marches aléatoires d'une profondeur de trois, permettant une exploration approfondie de la granularité de l'ontologie. Réalisée sur une série de 100 itérations, ce réglage a utilisé la même stratégie de marche aléatoire pour garantir une compréhen-

4. https://huggingface.co/dangvantuan/sentence-camembert-large

5. https://github.com/GateNLP/gateplugin-Tagger_Measurements

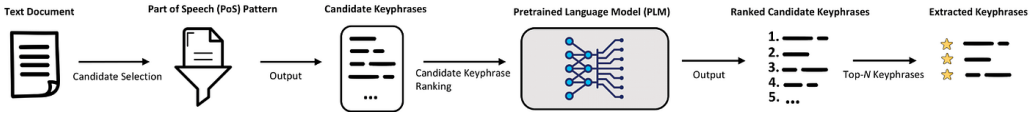


FIGURE 3 – Schéma du processus d'extraction non supervisée des syntagmes en utilisant PatternRank. Source : [26].

sion complète et adéquate de l'ontologie.

Vectorisation des candidats. Chaque syntagme extrait est transformé en un vecteur en utilisant la sortie de OWL2vec*. Cette étape de vectorisation permet leur représentation dans le même espace vectoriel que les axiomes de l'ontologie, facilitant ainsi les comparaisons sémantiques directes entre ces syntagmes et les concepts de l'ontologie.

Correspondance et sélection des similarités. Pour chaque vecteur de syntagmes, nous déterminons les dix concepts ontologiques les plus proches, classés selon leurs scores de similarité sémantique. Pour affiner cette sélection initiale et identifier avec précision le concept adéquat parmi les premiers candidats, nous employons un module de reclassement décrit plus en détail dans cet article [16], lequel se fonde sur une analyse syntaxique poussée. Ce module, exploite l'analyse syntaxique pour distinguer le concept le plus pertinent parmi les options pré-sélectionnées, en se fondant sur les scores de similarité issus de notre modèle de plongement. Le cœur de notre innovation réside dans l'utilisation de l'analyseur syntaxique de SpaCy [6], un outil conçu pour isoler le mot ou le syntagme le plus significatif au sein d'une phrase. En analysant la structure grammaticale du syntagme, le module de reclassement peut identifier avec précision l'entité principale, permettant ainsi une correspondance plus exacte entre le syntagme analysé et le concept de l'ontologie pertinent.

3.6 Extraction non supervisée de relations

Dans le contexte de l'analyse des CRH, l'extraction de relations (ER) constitue une étape cruciale du TALN. Cette tâche vise à identifier et à définir les liens sémantiques existant entre les entités nommées détectées dans le texte. Il existe deux principales techniques d'extraction de relations entre entités : les méthodes fondées sur des règles de modèles (*template rule-based*) et les méthodes fondées sur des vecteurs propres (*eigenvector-based*).

Dans la méthode fondée sur des règles, les caractéristiques linguistiques des relations entre entités sont d'abord organisées par des linguistes. Ensuite, les règles sont compilées [23], enfin, les relations entre entités sont extraites à travers une correspondance fondée sur ces règles.

Les méthodes fondées sur des vecteurs propres peuvent être divisées en deux types : l'apprentissage automatique traditionnel et l'apprentissage profond [38]. Pour répondre à nos besoins, nous utilisons une combinaison des méthodes décrites ci-dessus. Nous combinons l'analyse syntaxique des dépendances et la struture de l'ontologie pour identifier et classer les relations entre les entités.

6. https://spacy.io/models/fr

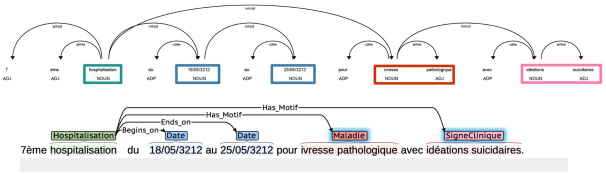


FIGURE 4 – Analyse Syntaxique et Structuration de Texte : Panneau Supérieur - Arbre d'Analyse Syntaxique avec Positions de Mots et Types de Dépendances ; Panneau Inférieur - Regroupement en SN et Identification de hospitalisation/date et hospitalisation/Maladie/SigneClinique.

3.6.1 Définition de la Tâche

À ce stade du processus d'annotation, les informations extraites sont liées aux concepts de l'ontologie OntoPSY. La figure 2 décrit l'architecture du processus d'extraction de relations (représenté par le rectangle vert en pointillés). Notre approche se décompose en quatre phases principales que nous détaillons dans cette section. Nous ciblons spécifiquement l'extraction de 14 relations que nous regroupons en cinq catégories : a) les relations temporelles, qui articulent un ordre ou une séquence d'événements ou d'épisodes de soin ; ensuite, b) les associations « a pour motif » qui relient des événements ou des épisodes de soin à leurs causes sous-jacentes, telles que des maladies ou des symptômes ; puis, c) la relation « participe », qui lie des individus ou des médicaments à l'épisode de soin auquel ils sont associés ; quatrièmement, d) les relations de qualification offrent des précisions sur les entités, en se référant à des attributs comme le niveau scolaire ou la texture pour les individus, ou pour caractériser des soins et événements. Finalement e) les relations de dosage qui concernent la connexion entre des médicaments ou substances chimiques et leurs dosages, modes d'administration, et fréquences d'administration. Cette approche est illustrée dans la figure 4 où nous présentons un exemple de ces relations.

3.6.2 Solution proposée

Dans un cadre non supervisé, le principal défi est l'absence d'échantillons étiquetés indiquant la relation spécifique r pour chaque paire d'entités (e_h, e_t) dans la phrase avec e_h est l'entité tête et e_t et l'entité queue. Par conséquent, l'ensemble des relations $\mathcal{R}$ (ensemble de r_i) est explicitement défini grace aux relations de l'ontologie et la tâche repose sur l'identification de motifs, de dépendances syntaxiques et d'indices sémantiques au sein de la phrase x pour inférer la relation potentielle r.

Phase 1 - Selection des relations candidates. La sélection des relations est initialement guidée par la structure et les

relations présentes dans l'ontologie. Par exemple, si entre les entités *Maladie* et *Signe Clinique* il n'existe pas de relation dans l'ontologie, aucune relation candidate n'est considérée entre ces entités dans notre système. Cette approche nous permet de restreindre l'ensemble des relations possibles à celles qui sont soutenues par des connaissances ontologiques, garantissant une cohérence initiale dans le processus de sélection. Dans une premiere étape, nous identifions les phrases contenant plusieurs entités nommées et établissons toutes les combinaisons possibles de relations entre elles. En tenant compte de l'ordre qui est déterminant dans notre analyse, nous considérons les entités et les relations potentielles illustrées dans notre corpus de données. Considérons la phrase type issue de notre corpus, illustrée dans la figure 4. Nous avons les entités *hospitalisation*, *ivresse pathologique*, *idéations suicidaires* ainsi que les dates spécifiques *18/05/3212* et *25/05/3212*.
Nous examinons alors les combinaisons suivantes :

- $r1(e_h, e_t)$ où e_h est *hospitalisation* et e_t est *ivresse pathologique*, la relation $r1$ pouvant être interprétée comme *a pour motif*;
- $r2(e_h, e_t)$ où e_h est *hospitalisation* et e_t est *idéations suicidaires*, la relation $r2$ étant également *a pour motif*;
- pour les dates qui peuvent être associées respectivement aux entités *hospitalisation*, *ivresse pathologique* et *idéations suicidaires* considérées comme e_h, plusieurs relations candidates sont envisageables en fonction des informations contextuelles et de l'ontologie :
  - $r3(e_h, e_t)$ pour *a pour date de début*,
  - $r4(e_h, e_t)$ pour *a pour date de fin*,
  - $r5(e_h, e_t)$ pour *a pour date*, utilisée si on ne fait pas la distinction entre la date de début et de fin,

  avec, dans les 3 cas, $e_h \in$ {*hospitalisation, ivresse pathologique, idéations suicidaires*}.

Phase 2 - Analyse Syntaxique. Nous utilisons, ensuite, un composant d'analyseur de dépendances fondé sur les transitions de Spacy. Ce dernier est fondé sur le modèle Transformer, notamment camembert-base [20] avec une précision de 0.95 [7]. Les phrases, une fois étiquetées avec des tags de parties du discours (POS), sont analysées par cet outil. Il génère alors un arbre de dépendances pour chaque phrase, illustré dans la figure 4 (panneau supérieur), et assigne une fonction syntaxique à chaque mot.

Phase 3 - Identification des relations. Cette phase exploite l'analyse des chemins syntaxiques au sein de l'arbre de dépendance par mot, traçant le parcours depuis un point de départ e_h, généralement l'effecteur, vers les entités cibles e_t. Cette analyse syntaxique révèle une absence de connexion syntaxique directe entre *ivresse pathologique* et *idéations suicidaires* et les dates, signifiant qu'aucun chemin de dépendance n'indique une relation temporelle explicite avec ces entités. Toutefois, partant de *hospitalisation*, il y a une dépendance syntaxique tant avec les entités temporelles qu'avec *Maladie* et *Signe Clinique*, indice d'une relation de modification ou de causalité. Par conséquent, nous validons les relations $r1$ et $r2$ qui relient *hospitalisation* à *ivresse pathologique* et *idéations suicidaires* via *a pour motif*, marquant un lien causal où l'hospitalisation résulte de ces conditions. Les relations $r3$, $r4$, et $r5$ associant *hospitalisation* aux dates spécifiques restent des candidats et sont sujettes à une analyse plus approfondie dans la phase suivante.

Phase 4 - Application des règles. Cette étape finale du processus d'extraction de relations tire parti de règles spécifiques centrées sur les mots situés avant les entités temporelles pour déterminer la nature précise de la relation temporelle. En s'appuyant sur des indicateurs lexicaux clairs, tels que des mots ou des expressions indiquant un commencement (« début », « commencement », « à partir du ») ou une conclusion (« fin », « jusqu'au », « termine le »), nous pouvons affiner notre compréhension des relations $r3$, $r4$, et $r5$ restantes entre *hospitalisation* et les instances temporelles. La présence de ces indicateurs dans le contexte immédiat avant une entité temporelle nous permet d'attribuer avec précision la relation la plus adéquate. Par exemple, si un indicateur de début ou de fin précède une entité temporelle associée à *hospitalisation*, la relation $r3$ ou $r4$ sont validées. Si le contexte ne spécifie pas clairement un début ou une fin, ou que les indicateurs sont ambigus ou absents, la relation générale $r5$ *a pour date* est considérée comme appropriée. En résultat, illustré dans la figure 4 (panneau Inférieur), les relations finales sélectionnées pour *hospitalisation* sont directement influencées par ces indicateurs lexicaux, renforçant l'exactitude sémantique et contextuelle de notre modèle relationnel.

4 Analyse et résultats

4.1 Analyse des performances du système

Notre approche a été évaluée de manière distincte sur trois composantes clés : l'extraction des entités nommées, la normalisation des entités, et l'extraction des relations. Pour chaque composante, nous avons réalisé une analyse manuelle approfondie en utilisant un schéma d'annotation dédié conçu pour mesurer précisément les performances de notre pipeline. L'outil d'annotation BRAT a été utilisé pour sa facilité d'usage et sa capacité à répondre à nos critères spécifiques. Dans le cadre de l'optimisation de l'évaluation manuelle, nous avons classifié les entités extraites en 17 concepts uniques de haut niveau de l'ontologie OntoPSY. Les annotations dans BRAT incluaient les URI correspondant à chaque concept de l'ontologie, donnant ainsi aux annotateurs la possibilité de corriger les annotations au besoin. Pour les relations, nous avons retenu 14 types de relation. Il faut noter la différence d'approche d'évaluation entre les entités et les relations : les entités correspondent pour une grande majorité à des concepts médicaux et sociétaux : ils peuvent être appréhendés par les évaluateurs qui ont 17 concepts de haut niveau à leur disposition et peuvent préciser les concepts repérés à l'envi en balayant l'ontologie. les relations décrites dans l'ontologie sont très précises en raison du rôle tenu par icelles – modèle de données de

7. https://spacy.io/models/fr#fr_dep_news_trf

la plateforme gérant les données cliniques – dans le projet PsyCARE. Les relations telles qu'elles sont appréhendées par les experts sont plus proches de dépendances syntaxiques visibles dans les phrases : c'est pour cela qu'on en a retenu 14 (synthétisées en 5 types, Cf.3.6.1) et que nous sollicitons les experts dessus sans leur demander d'approfondir les URI.

Notre corpus d'évaluation est composé de 60 CRH extraits aléatoirement de l'ensemble de données (5120 phrases, 10013 concepts ontologiques non uniques annotés). Deux personnes ont évalué l'annotation sémantique et leur contexte, notamment le repérage de la négation, l'hypothétique, la temporalité et la personne impliquée (p. ex. le patient vs un membre de sa famille). Dans les résultats, pour les tâches REN et ER, nous indiquons les scores de précision, de rappel et de F1.

Dans le processus de normalisation des entités, où il est possible d'attribuer à chaque entité détectée plusieurs URIs candidats issus de l'ontologie, l'utilisation de métriques fondées sur le classement s'avère essentielle pour évaluer avec précision les correspondances établies. Ainsi, nous mettons en œuvre des mesures largement utilisées dans ce domaine : Hits@1, Hits@5, ainsi que le rang réciproque moyen MRR. Hits@1 et Hits@5 évaluent le rappel en mesurant la présence des correspondances correctes parmi les 1 et 5 premiers résultats proposés par notre système de normalisation. Le MRR, quant à lui, offre une perspective quant à la qualité du classement en calculant la moyenne des inverses des positions attribuées aux correspondances correctes. Hits@1, en particulier, permet de déterminer dans quelle mesure l'URI le mieux classé par notre système coïncide avec une correspondance vérifiée. Le MRR complète cette analyse en appréciant de manière globale l'exactitude du classement des URIs candidats, grâce à l'agrégation des positions relatives des correspondances avérées.

4.2 Résultat

Nous avons évalué l'accord inter-annotateurs à travers les contributions des deux annotateurs sur l'ensemble des tâches. Cet accord s'est avéré être de 0,79, indiquant une cohérence significative dans les annotations fournies. En conséquence, nous avons procédé à la consolidation de tous les comptes rendus annotés. La REN et l'extraction du contexte des concepts ont démontré une précision globale de 0.9610, un rappel de 0.9248 et un score F1 de 0.9425. Les résultats détaillés sont présentés dans le Tableau 1.

Dans le cadre de l'évaluation de la ER, nous avons distingué 14 types de relations différents. Ces derniers ont été synthétisés dans le Tableau 2, représentant un ensemble de 3473 annotations. Les performances globales atteintes pour cette tâche sont résumées par une précision de 0.92, un rappel de 0.81, et un score F1 de 0.86.

Dans le cadre de notre analyse des performances de normalisation des entités, les résultats obtenus pour les métriques clés sont particulièrement révélateurs. Pour Hits@1, nous avons atteint un taux de 84.8%. Cette performance souligne l'efficacité du système à déterminer l'URI le plus pertinent pour chaque entité dès la première proposition.

TABLE 1 – Résultats quantitatifs de l'évaluation de la reconnaissance d'entités nommées.

Entité nommée		Total	Precision	Recall	F1
Age		164	0.977	0.904	0.939
Substance	Name	1034	0.8200	0.9805	0.8822
	Dosage	608	0.99	0.94	0.97
	DrugForm	14	0.857	0.857	0.857
Temporal Inf.	Date	1208	0.9942	0.9709	0.9824
	Duration	221	0.7481	0.9619	0.8417
	Frequency	511	0.9954	0.7688	0.9819
	Time	88	0.8750	0.9459	0.9091
EpisodeDeSoin		400	0.975	0.9485	0.8273
EvenementVecu		343	0.9589	0.9333	0.9459
ExamenClinique		308	0.9799	0.8811	0.8875
Hospitalisation		520	0.9954	0.9688	0.9819
Individu		540	0.991	0.7774	0.8747
Maladie		1166	0.9894	0.9852	0.9843
PartieDuCorps		22	0.98	0.6667	0.8000
Qualifier		600	0.9882	0.8802	0.9342
SigneClinique		1250	0.9870	0.9775	0.9823
AnnotationsToAdd		721	0.9256	0.8077	0.8626

TABLE 2 – Résultats quantitatifs des évaluations de l'extraction de relations.

Relation	Total	Precision	Rappel	F1
Relations Temporelles	860	0.9130	0.8077	0.8571
A pour motif	1050	0.9750	0.9070	0.9398
Participe	286	0.9831	0.5800	0.7296
Qualifie	756	0.9211	0.7368	0.8188
Relations Medicaments dosage	521	0.9046	0.8333	0.8678

En élargissant notre évaluation aux cinq premières propositions avec Hits@5, le taux s'améliore pour atteindre 90.4%, démontrant ainsi la capacité du système à inclure la correspondance exacte parmi les choix les plus privilégiés. Cette métrique confirme que même si la correspondance idéale n'est pas toujours première, elle figure presque toujours parmi les premières propositions.

Quant au MRR, qui offre une vue d'ensemble de la performance du système en prenant en compte le rang de la bonne réponse, le score obtenu est de 85%.

5 Discussion

Les résultats témoignent des performances obtenues au sein de notre étude. Notre démarche méthodique, adaptée tant à l'analyse des entités qu'à celle des relations, a été fructueuse et a mis en lumière les défis spécifiques liés au traitement de textes complexes, en particulier ceux du domaine de la psychiatrie.

Ces résultats encourageants s'expliquent principalement par deux facteurs. Premièrement, la richesse du vocabulaire de l'ontologie, notamment dans les domaines des signes, symptômes psychiatriques, des maladies, et des événements vécus, ce qui contribue directement à la qualité du plongements ontologique ainsi qu'à celle de la normalisation. En outre, nous avons réalisé des expérimentations sur la qualité du plongement générés par OWL2Vec*, qui ont

révélé notre aptitude à distinguer efficacement les classes de premier niveau de l'ontologie OntoPSY, cette analyse est disponible sur un notebook Jupyter GitHub [8].
Le second facteur déterminant est l'intégration de la structuration spécifique des CRH dans le processus d'extraction des relations, notamment à travers l'ajout de règles de filtrage. Cette adaptation améliore considérablement la précision de notre système, bien que cela puisse représenter un défi pour la généralisation de cette approche à d'autres types de documents.
De plus, il est essentiel de souligner que la performance globale de notre système est étroitement liée à l'efficacité du modèle d'extraction des syntagmes. Les imperfections inhérentes à ce processus ne se limitent pas à leur occurrence initiale ; elles se propagent à travers le système, impactant chaque étape subséquente de l'analyse. Cette interdépendance souligne la nécessité d'une extraction précise des syntagmes dès les premiers stades, étant donné que toute erreur générée peut être amplifiée et influencer l'ensemble des résultats obtenus. Cette limitation nécessite une étude d'ablation pour comprendre l'impact de chaque étape sur les résultats finaux du système d'annotation. En outre, l'analyse de dépendance influence également la précision et le rappel des tâches d'extraction de relations. Cette interdépendance met en exergue l'importance vitale d'une extraction précise des groupes nominaux dès les premiers instants, puisque les erreurs initiales peuvent être exacerbées, affectant de manière significative la qualité totale des résultats. Face à cette contrainte, il s'avère indispensable de recourir à une méthode d'ablation pour identifier avec exactitude l'impact de chaque élément sur la performance globale du système d'annotation.
L'utilisation de méthodes d'apprentissage non supervisé, intégrant une ontologie spécifique au domaine pour affiner la précision de l'apprentissage, pourrait diminuer le besoin d'annotations manuelles. La performance globale de l'annotation représente la somme des performances des différents composants. Suite à plusieurs améliorations apportées à chaque élément, comme l'intégration de règles spécifiques ou l'emploi de plongements pré-entraînés sur un corpus médical, nous avons atteint un niveau de performance jugé satisfaisant. Néanmoins, des opportunités d'amélioration de la performance de chaque composant du système proposé subsistent et feront l'objet de recherches approfondies dans nos travaux futurs.

6 Conclusion

L'objectif principal de ce travail est de reconstruire les données structurées des patients à partir de leurs CRH afin d'enrichir les données du projet PsyCARE. À cette fin, nous avons combiné l'utilisation de méthodes d'apprentissage non supervisées avec OntoPSY, une ontologie spécifique à la psychiatrie, pour récupérer et normaliser les entités biomédicales et identifier les relations entre ces paires d'entités dans le texte.
Initialement, pour l'extraction d'information, nous avons adapté l'algorithme PatternRank d'extraction de syntagmes clés. Nous avons ensuite exploité le plongement de l'ontologie dans un espace vectoriel avec OWL2Vec* pour associer ces informations aux concepts correspondants de l'ontologie. Enfin, en nous appuyant sur la structure de l'ontologie et l'analyse des dépendances syntaxiques, nous avons pu extraire les relations entre les entités.
Ce travail se distingue par l'exploitation des technologies du web sémantique combinées à l'apprentissage profond pour créer automatiquement des documents annotés dans le domaine de psychiatrie. Les performances de notre système sont prometteuses et ouvrent la voie à de nombreuses améliorations en termes de performances. Cette initiative a mis en lumière l'apport des plongements d'ontologie dans le contexte d'ontologies biomédicales variées et interconnectées, renforçant l'efficacité de l'annotation sémantique. Bien que cet article se concentre sur le domaine de la psychiatrie, des tests préliminaires dans le champ de la néphrologie avec une ontologie dédiée ont également révélé des perspectives encourageantes, bien que ces dernières ne soient pas l'objet principal de cette publication.
Les prochaines étapes de notre recherche incluent une analyse comparative entre notre méthode utilisant des plongements de mots non contextuels (word2vec) et les plongements sémantiques contextuels [7] pour l'annotation sémantique. Nous visons à améliorer le taux de rappel dans l'extraction de relations, en envisageant l'utilisation de notre base de données annotées et l'application d'apprentissage faiblement supervisé. Nous prévoyons également de tester l'efficacité de notre approche avec des données annotées en français disponibles publiquement.

8. https://github.com/AouinaOns/Semantic-Annotation

Remerciements

Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du Programme d'Investissements d'Avenir portant la référence PsyCARE ANR-18-RHUS- 0014.

Références

[1] Ons Aouina, Jacques Hilbey, and Jean Charlet. Ontology-Based Semantic Annotation of French Psychiatric Clinical Documents. *Studies in health technology and informatics*, 302 :793–797, May 2023.

[2] Dennis Aumiller et al. Online dateing : A web interface for temporal annotations. 07 2022.

[3] Antoine Bordes et al. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[4] Borui Cai et al. Temporal knowledge graph completion : A survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI-2023. International Joint Conferences on Artificial Intelligence Organization, August 2023.

[5] Ricardo Campos et al. YAKE ! Keyword extraction from single documents using multiple local features. *Information Sciences*, 509, 2020.

[6] Jiaoyan Chen et al. OWL2Vec* : Embedding of OWL Ontologies, January 2021. arXiv :2009.14654 [cs].

[7] Jiaoyan Chen et al. Contextual Semantic Embeddings for Ontology Subsumption Prediction, March 2023.

[8] Shivani Choudhary, Tarun Luthra, Ashima Mittal, and Rajat Singh. A Survey of Knowledge Graph Embedding and Their Applications, July 2021.

[9] Pratik Devkota et al. Using ontology embeddings with deep learning architectures to improve prediction of ontology concepts from literature. 2023.

[10] Jacob Devlin et al. BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. arXiv :1810.04805 [cs].

[11] Aryo Pradipta Gema et al. Knowledge graph embeddings in the biomedical domain : Are they useful ? a look at link prediction, rule learning, and downstream polypharmacy tasks, 2023.

[12] Maarten Grootendorst. Keybert : Minimal keyword extraction with bert., 2020.

[13] Jacques Hilbey, Xavier Aimé, and Jean Charlet. *Temporal Medical Knowledge Representation Using Ontologies*. May 2022.

[14] Ali Hur et al. A Survey on State-of-the-art Techniques for Knowledge Graphs Construction and Challenges ahead, December 2021.

[15] Lars Juhl Jensen et al. Literature mining for the biologist : from information retrieval to biological discovery. *Nature Reviews Genetics*, 7(2), 2006.

[16] İlknur Karadeniz et al. Linking entities through an ontology using word embeddings and syntactic reranking | BMC Bioinformatics 2019 | Full Text.

[17] Dongkwan Kim et al. Supervised Graph Attention Network for Semi-Supervised Node Classification. 2019.

[18] Maxat Kulmanov et al. El embeddings : Geometric construction of models for the description logic el ++.

[19] Xuexiang Li et al. Efficient Medical Knowledge Graph Embedding : Leveraging Adaptive Hierarchical Transformers and Model Compression. 12, 2023.

[20] Louis Martin et al. Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, volume abs/1911.03894, 2019.

[21] Rada Mihalcea et al. TextRank : Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

[22] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space, September 2013.

[23] Scott Miller, Heidi Fox, Lance Ramshaw, and Ralph Weischedel. A Novel Use of Statistical Parsing to Extract Information from Text. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*, 2000.

[24] Jose Moreno et al. Combining word and entity embeddings for entity linking. 2017.

[25] Dietrich Rebholz-Schuhmann et al. Text processing through Web services. *Bioinformatics*, 2008.

[26] Tim Schopf et al. PatternRank : Leveraging Pretrained Language Models and Part of Speech for Unsupervised Keyphrase Extraction. In *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, pages 243–248, 2022.

[27] N. Shadbolt, T. Berners-Lee, and W. Hall. The Semantic Web Revisited. *IEEE Intelligent Systems*, 21(3) :96–101, January 2006.

[28] Wei Shen et al. Entity linking with a knowledge base : Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27 :443–460, 2015.

[29] Fatima Zohra Smaili et al. Onto2Vec : joint vector-based representation of biological entities and their ontology-based annotations. *Bioinformatics*, 2018.

[30] Fatima Zohra Smaili et al. OPA2Vec : combining formal and informal content of biomedical ontologies to improve similarity-based prediction. 35, 11 2018.

[31] Chengyu Sun et al. A Review of Unsupervised Keyphrase Extraction Methods Using Within-Collection Resources. *Symmetry*, 12(11) :1864, November 2020.

[32] Si Sun, Zhenghao Liu, Chenyan Xiong, et al. Capturing Global Informativeness in Open Domain Keyphrase Extraction, September 2021.

[33] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. RotatE : Knowledge Graph Embedding by Relational Rotation in Complex Space, February 2019.

[34] The OBI Consortium, Barry Smith, et al. The OBO Foundry : coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11) :1251–1255, November 2007.

[35] Kristina Toutanova et al. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, April 2015.

[36] Perceval Wajsburt et al. Eds-nlp : efficient information extraction from french clinical notes.

[37] Xiaojun Wan et al. Single document keyphrase extraction using neighborhood knowledge. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08. AAAI Press, 2008.

[38] Rui Xing et al. BioRel : towards large-scale biomedical relation extraction. *BMC Bioinformatics 2020*.

[39] Bishan Yang et al. Embedding entities and relations for learning and inference in knowledge bases, 2015.

Estimation du Rendement du Mil Perlé (*Pennisetum glaucum*) par Machine Learning à l'aide d'Images Satellites

A. Chemchem[1], L. Mohimont[2], F. Alin[2], L.A. Steffenel[2]

[1] ATOS - Pôle Data Driven Intelligence
Rue du Mas de Verchant, 34000 Montpellier, France

[2] Université de Reims Champagne-Ardenne,
Laboratoire LICIIS - LRC CEA DIGIT

lamine.chemchem@atos.net, lucas.mohimont@univ-reims.fr ;
francois.alin@univ-reims.fr, luiz-angelo.steffenel@univ-reims.fr

Résumé

L'estimation du rendement agricole joue un rôle crucial dans la poursuite des objectifs de développement durable des Nations Unies, représentant ainsi un outil essentiel dans la prise de décisions concernant les systèmes d'approvisionnement. Dans ce travail, nous nous intéressons à la prédiction du rendement du Pennisetum glaucum, aussi connu comme "mil à chandelle" ou "mil perlé". Connaître le potentiel de production le plus tôt possible permet de prendre des mesures préventives et éviter des défauts d'approvisionnement pour la population. Pour ce faire, nous croisons les données historiques de rendement des parcelles au Sénégal avec des données satellitaires couvrant trois phases différentes du cycle de vie du mil, grâce à des méthodes d'apprentissage automatique. En comparant différentes méthodes, nous avons obtenu des estimations de rendement assez précises 1 mois avant la récolte, avec un taux d'erreur qui ne dépasse pas 140 kg/ha.

Mots-clés

Rendement agricole, Télédétection optique, Apprentissage automatique

Abstract

Agricultural yield estimation contributes to many of the United Nations' sustainable development goals, and can be considered as a decision-making tool for a supply system. In this work, we are interested in predicting the yield of Pennisetum glaucum, also known as "pearl millet". Knowing the production potential of this cereal as early as possible enables authorities to take preventive measures and avoid supply shortages for the population. In this work, we cross-reference historical plot yield data in Senegal with satellite data covering three different phases of the millet life cycle, using machine learning methods. By comparing different methods, we obtained fairly accurate yield estimates 1 month before harvest, with an error rate of no more than 140 kg/ha.

Keywords

Crop yield estimation, Optical remote sensing, Machine learning

1 Introduction

Le "mil à chandelle", "mil perlé" ou simplement **mil** (*Pennisetum glaucum*) est une espèce de plantes annuelles de la famille des Poaceae (Graminées). Elle est cultivée comme céréale pour ses graines comestibles et joue un rôle important en tant que culture vivrière en Inde et au Pakistan, ainsi que dans le Sahel africain et dans des zones semi-arides. Connaître le potentiel de production fait partie des mesures permettant de garantir la sécurité alimentaire de la population en cas de baisse de production.

Cette étude vise à estimer les rendements de mil par télédétection optique dans la petite agriculture familiale au Sénégal. Ce projet s'inscrit dans le cadre de la sécurité alimentaire et des moyens de subsistance des populations, et ses résultats pourraient représenter un outil d'aide à la décision pour les services d'approvisionnement de plusieurs ODD (Objectifs de Développement Durable) adaptés par les Nations Unies. En particulier pour l'ODD 2, qui vise à "Éliminer la faim, assurer la sécurité alimentaire et une meilleure nutrition et promouvoir l'agriculture durable".

Le site choisi pour cette étude est situé dans la zone de l'ancien bassin arachidier du Sénégal, comme le montre la figure 1. Il représente la principale zone de production agricole du pays. Le bassin est constitué de sols ferrugineux tropicaux permettant une production agricole principalement composée de céréales sèches (mil, sorgho, maïs) et de légumineuses (arachide, niébé) cultivées seules ou en association.

Le jeu de données de cette étude a été partagé dans le cadre d'un défi de science des données organisé par Acta [1]. Il se compose d'un total de 81 parcelles de mil réparties dans le site du bassin. Cette région a été choisie parce qu'il s'agit d'une zone d'intérêt de longue date pour plusieurs équipes de recherche, de sorte que nous disposons de connaissances

1. Instituts techniques agricoles `http://www.acta.asso.fr/`

sur le terrain, en plus de la base de données agronomique ou paysagère historique. Un autre avantage très important est la présence d'équipes de recherche sur le terrain pour assurer et coordonner la collecte de données.

Le climat de la région est unimodal avec une saison des pluies entre juillet et octobre. Cependant, la zone connaît une forte croissance démographique, à laquelle s'ajoutent une réduction des temps de jachère, un appauvrissement progressif des sols, ce qui conduit à une baisse des rendements observés en milieu rural, à une dégradation des ressources naturelles et à la perte de biodiversité [3].

Afin d'atteindre nos objectifs, nous mettons en oeuvre des méthodes d'apprentissage automatique avec deux axes : régression et classification. Le premier axe d'étude consiste à prédire le rendement quantitatif avec des méthodes de régression pour chaque parcelle. Dans le deuxième axe, nous adaptons les méthodes de classification supervisée afin d'affiner encore la précision des prédictions. Le résultat final de cette étude pourrait être un OAD (outil d'aide à la décision) de surveillance des risques d'approvisionnement.

Le reste du document est organisé comme suit : la section 2 passe en revue la littérature sur l'apprentissage automatique appliqués à l'agriculture intelligente. La section 3 décrit l'ensemble de données et le dispositif expérimental mis en place. La section 4 expose les résultats obtenus avec quelques analyses et discussions. Enfin, la section 5 conclut le document avec quelques perspectives.

2 État de l'Art

Les méthodes d'apprentissage automatique sont utilisées comme une nouvelle approche qui révolutionne les modèles classiques de prédiction du rendement. Toutefois, les approches basées sur l'apprentissage de données historiques nécessitent des données étiquetées pour fournir une modélisation précise. En effectuant une étape d'apprentissage sur les données étiquetées, ces méthodes sont capables de produire des modèles de prédiction quantitatifs appelés méthodes de régression. Par exemple, dans le cas de la prédiction du rendement, on veut prédire la quantité produite en tonnes par hectare, ou en kilogrammes par hectare. Il est aussi possible de construire un modèle de prédiction qualitatif via les méthodes de classification, permettant par exemple de prédire la classe de rendement attendue : faible, moyen ou élevé.

Dans la littérature, on peut citer le travail de [9], qui a comparé deux méthodes de régression : Perceptron multicouche (MLP) et Support Vector Regressor (SVR) pour la prédiction du rendement du maïs. En utilisant les données de l'indice de végétation amélioré (EVI) et des séries temporelles climatiques des dix dernières années, la méthode MLP mise en œuvre a atteint un score R^2 de 0,81. L'EVI et les données satellitaires en général sont utiles pour compenser le manque de données agricoles de détection sur le terrain. Malheureusement, nous n'avons pas pu accéder à la base de données de cette étude pour comparer nos méthodes, mais nous avons mis en œuvre les approches SVR et MLP sur notre ensemble de données.

Une autre recherche intéressante est celle de [14], dans laquelle les auteurs ont utilisé des séries temporelles de NDVI (Normalized Difference Vegetation Index) sur 10 jours avec des apports d'engrais chimiques comme caractéristiques d'apprentissage pour la prédiction du rendement du blé. Ils ont comparé différentes caractéristiques basées sur le NDVI, le NDVI cumulé et le NDVI cumulé à des dates significatives avec l'arbre de régression boosté (Boosted Regression Tree - BRT) et le SVR. Les deux techniques ont été utilisées pour la sélection des caractéristiques et la modélisation de la régression. Les meilleurs résultats ont été obtenus avec le BRT, avec une erreur calculée par RMSE (Root Means Square Error) inférieure à 0,2 tonne par hectare. Dans notre étude, nous avons utilisé l'indice NDVI avec cinq autres indices.

Dans [8], les auteurs ont utilisé des séries temporelles NDVI de 16 jours avec une résolution de 250 mètres et des observations climatiques SILO [2] pour la prévision du rendement du blé en Australie. Chaque échantillon de données est un pixel correspondant à un carré de 250m x 250m et différents modèles de base et d'ensemble ont été comparés sur la précision basée sur le pixel avec une validation croisée 5 fois. Les meilleurs résultats ont été obtenus par un SVR (Support vector Regressor) adapté utilisant le noyau RBF (radial basis function) qui atteint une erreur RMSE de 0,55t/ha et R^2 de 0,77.

Une autre étude de [12] se concentre sur l'axe de la classification supervisée afin de traiter les images satellites. Les auteurs ont utilisé les données NDVI, pédologiques et météorologiques pour prédire le rendement du maïs à l'intérieur d'un champ. Le champ a été divisé en 63 unités de traitement et la modélisation a été simplifiée à une classification binaire : classes de faible rendement et de rendement moyen à élevé. Cinq classificateurs supervisés différents ont été comparés avec une validation croisée 5 fois, et les meilleurs résultats ont été obtenus par le classificateur XG-Boost avec une précision de 95%. Cette étude nous a incités à réaliser une étude comparative des méthodes de classification les plus populaires.

Des méthodes basées uniquement sur l'historique de la production et les données météorologiques ont aussi montré des résultats prometteurs. Ainsi, [5] a pu obtenir des taux de précision supérieurs à 99% pour la culture du soja ou du riz, et de 98% pour la culture du maïs, en utilisant la méthode Random Forest sur des données issues de régions agricoles brésiliennes. Ce travail a toutefois bénéficié d'un historique sur plus de 20 ans de production agricole, ce qui n'est pas possible dans notre cas.

L'apprentissage profond peut également être utilisé pour prédire le rendement. Souvent, l'apprentissage profond nécessite de grands ensembles de données, mais dans certains cas, l'apprentissage par transfert peut être mis en œuvre pour compenser le manque de données en utilisant un modèle pré-entraîné. Les auteurs de [15] ont utilisé l'apprentissage par transfert pour former un modèle de mémoire à

2. SILO est une base de données du gouvernement du Queensland contenant des données climatiques quotidiennes continues pour l'Australie de 1889 à nos jours.

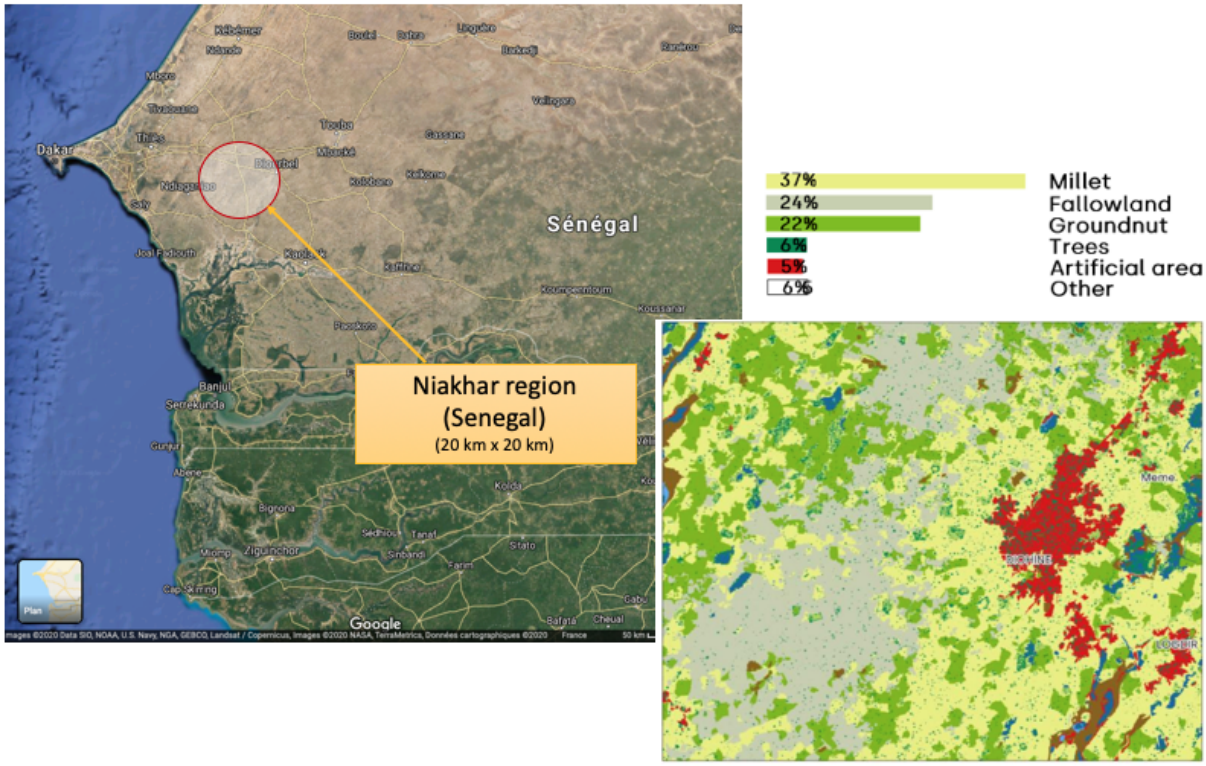


FIGURE 1 – Site étudié : région de Niakhar, ancien bassin producteur d'arachide, Sénégal [11].

long terme (LSTM) avec la réflectance MODIS [3] et les séries temporelles de températures pour la prédiction du rendement du maïs au Brésil. Le modèle a d'abord été entraîné sur un plus grand ensemble de données, avec 1837 échantillons de récoltes en Argentine, puis le modèle a été réentraîné sur 336 données de récoltes brésiliennes. La LSTM avec apprentissage par transfert a obtenu de meilleurs résultats moyens que la LSTM sans apprentissage par transfert. C'est pourquoi nous prévoyons d'explorer cette méthode sur notre ensemble de données dans un travail futur.

3 Matériaux & Méthodes

3.1 L'ensemble de données

L'ensemble de données utilisé dans cette étude est récolté à partir de 81 parcelles de mil situées dans la région de Niakhar, au Sénégal. Ce jeu de données comprend des données historiques collectées sur les années 2017 et 2018, représentant le rendement du mil en Kg/ha, ainsi que les données satellitaires des parcelles correspondantes, comme le montre la figure 2. Il faut noter que les parcelles étudiées sont trop petites, pour cette raison nous ne pouvons pas réaliser une étude individualisée (*intra-field*).

Les données satellitaires extraites contiennent les indices de végétation suivants : **NDVI, MSAVI2, NDWI, CIGreen, GDVI** et **PSRINIR**. Ces indices permettent d'estimer les paramètres biophysiques et sont calculés à partir de la réflectance de deux bandes spectrales, rouge (R) et proche infrarouge (NIR).

L'indice NDVI (Normalized Difference Vegetation Index) est le plus utilisé. Le calcul de cet indice est basé sur la réflectance de la chlorophylle dans le proche infrarouge et permet de suivre la biomasse intraparcellaire. L'indice SAVI (Soil-Adjusted Vegetation) est dérivé de cet indice et propose un ajustement avec une constante.

Plus tard, l'indice MSAVI2 (Modified Soil-Adjusted Vegetation Index) a été proposé par [13], il utilise une constante ajustée aux conditions locales.

Le NDWI (Normalized Difference Water Index) est basé sur le même principe que le NDVI et permet de surveiller l'état hydrique des cultures (Gao, 1996). Le NDWI est basé sur le pic d'absorption de l'eau dans une bande infrarouge de courte longueur d'onde.

Le CIGreen (Green Chlorophyll Index) est utilisé pour évaluer la teneur en chlorophylle des feuilles. Cet indice est sensible aux petites variations de chlorophylle.

Le GDVI (Generalized Difference Vegetation Index) est un indice dérivé du NDVI, particulièrement adapté aux zones arides où le couvert végétal est faible [16].

Finalement, le PSRINIR (Plant Senescence Reflectance Index Near Infra Red), proposé par [10], compare les caroténoïdes et les chlorophylles, identifiant ainsi la sénescence de la canopée (augmentation des caroténoïdes).

Les approches d'apprentissage automatique présentées dans cet article ont été mises en œuvre en utilisant le langage python avec la bibliothèque Scikit-Learn. L'optimi-

3. https://visibleearth.nasa.gov/images/54078/modis-surface-reflectance

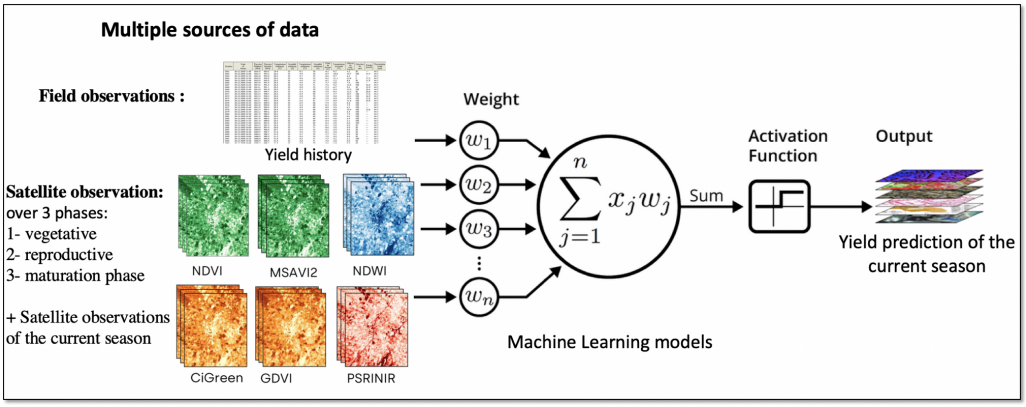


FIGURE 2 – Workflow général : De l'intégration des données à la prédiction du rendement.

sation des hyperparamètres a été accélérée grâce aux ressources du Centre de Calcul Régional ROMEO [4] de l'Université de Reims Champagne-Ardenne.

3.2 Méthodologie

Les observations par satellite sont collectées au cours de trois périodes de croissance différentes, conformément au calendrier de culture du mil à chandelle (*Pennisetum glaucum*) expliqué dans la figure 3.

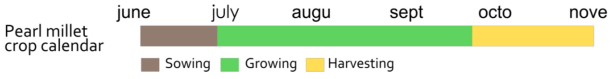


FIGURE 3 – Calendrier cultural du mil à chandelle dans l'ancien bassin arachidier, Sénégal [11].

Dans un premier moment, nous avons choisi d'estimer les rendements le plus tôt possible avant la récolte. Cela se fait en estimant le rendement uniquement avec des données de la phase végétative (5 mois avant la récolte), sans prendre en compte les données des phases reproductive et de maturation. Une deuxième expérience consiste à prédire les rendements 3 mois avant la récolte, utilisant cette fois-ci les données des phases végétative et reproductive. Finalement, la troisième expérience consiste à utiliser toutes les données disponibles, c'est-à-dire les données des trois phases végétation, reproduction et maturation, afin de prédire le rendement du mil environ 1 mois avant la récolte.
En outre, et pour tirer parti des méthodes d'apprentissage automatique, nous avons mis en œuvre des algorithmes pour estimer le rendement quantitatif, par le biais de méthodes de régression. Puis, dans un second temps, nous avons exploré les méthodes de classification supervisée afin de valider les résultats et d'être le plus précis possible. Pour chaque type d'approche d'apprentissage automatique, nous avons mis en place une étude comparative des algorithmes les plus prometteurs tels que décrits dans la section littérature. Il est à noter que, pour chaque approche d'apprentissage, les meilleurs hyperparamètres sont sélectionnés par validation croisée et recherche en grille (*GridSearch*).

4. http://romeo.univ-reims.fr

3.3 Méthodes pour la régression et leur évaluation

Dans cette partie, nous expliquons les étapes que nous avons suivies pour réaliser l'étude comparative des algorithmes de régression.

Prétraitement des données par mise à l'échelle des caractéristiques : cette étape est appliquée pour normaliser la plage des variables indépendantes de l'ensemble de données. Étant donné que la plage de valeurs des données brutes varie considérablement, les fonctions objectives ne fonctionneront pas correctement sans normalisation. Une autre raison pour laquelle la mise à l'échelle des caractéristiques est appliquée est que la descente de gradient converge beaucoup plus rapidement avec la mise à l'échelle des caractéristiques [7].
Dans notre implémentation, nous avons appliqué la normalisation min-max, qui est la méthode la plus simple et qui consiste à remettre à l'échelle la plage de caractéristiques dans un intervalle [0, 1] ou [-1,+1]. La sélection de la plage cible dépend de la nature des données, et puisque dans notre ensemble de données il n'y a pas de données négatives, nous les avons mises à l'échelle entre [0, 1].

Validation par répartition *train/test* : Fondamentalement, l'évaluation des approches d'apprentissage automatique se fait par la division de l'ensemble de données en deux ensembles, l'un appelé ensemble d'entraînement (*train*) et l'autre ensemble de test. Le premier contient les données avec les étiquettes utilisées pour construire le modèle, tandis que le second est utilisé pour tester les performances de ce modèle. Dans notre cas, comme nous ne disposons que de deux années d'historique de données, nous avons pris les données de 2017 pour former les modèles et celles de 2018 pour les tester.

Évaluation de la régression par R2-score et RMSE : nous avons évalué nos modèles de régression par les deux formules d'évaluation les plus connues : R^2-score et RMSE. Le Score R^2 (appelé aussi R-carré) est une mesure statistique de la proximité des données par rapport à la

droite de régression ajustée. Il est également connu sous le nom de coefficient de détermination ou de coefficient de détermination multiple pour la régression multiple [4]. En général, plus le R au carré est élevé, mieux le modèle s'adapte à vos données.
Le RMSE (Root Mean Squared Error) est la racine carrée de l'erreur quadratique moyenne (Mean Squared Error - MSE), qui est une fonction de risque correspondant à la valeur attendue de la perte d'erreur quadratique.

3.4 Méthodes de classification supervisée et leur évaluation

Après la mise en œuvre de l'étude comparative de régression, nous avons exploré les méthodes d'apprentissage automatique pour la classification supervisée, en transformant le problème de régression en un problème de classification multi-classes. Cette approche vise à affiner la précision de notre prédiction de rendement, de sorte qu'au lieu d'essayer de prédire une valeur quantitative, il nous suffit de prédire une classe parmi trois classes possibles : rendement faible, rendement moyen ou rendement élevé.

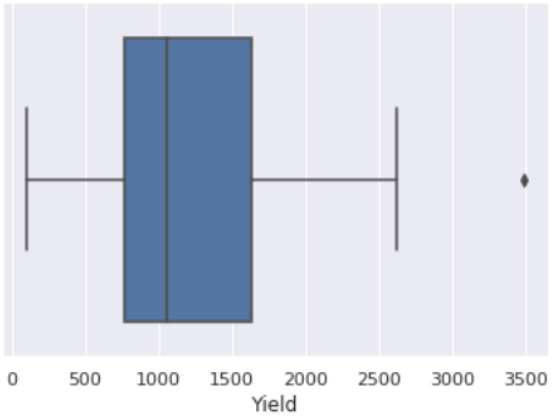


FIGURE 4 – Distribution des valeurs de rendement.

La distribution des valeurs de rendement se situe entre 107,9 kg/ha et 3488,9 kg/ha, comme le montre la figure 4. À partir de cette distribution, nous avons créé les trois classes de rendement suivantes :

— Classe "rendement bas" : si 700 kg/ha $>$ rendement ;
— Classe "rendement moyen" : si 700 kg/ha $\leq$ rendement $<$ 1600 kg/ha ;
— Classe "rendement haut" : si rendement $\geq$ 1600 kg/ha.

Le résultat de la distribution des classes obtenues est présenté dans la figure 5. Nous pouvons remarquer que les classes obtenues sont fortement déséquilibrées, ce qui nécessite une étape de prétraitement supplémentaire pour cet ensemble de données, avec la technique SMOTE (Synthetic Minority Over-sampling Technique). Dans SMOTE, les classes minoritaires sont suréchantillonnées en introduisant des instances synthétiques dans lesquelles chaque échantillon de classe minoritaire est prélevé. Les données générées sont insérées le long des segments de ligne reliant certains des k plus proches voisins de la classe minoritaire. Les voisins sont choisis au hasard parmi les k plus proches voisins en fonction de l'ampleur du suréchantillonnage nécessaire. Cinq voisins les plus proches sont actuellement utilisés dans la mise en œuvre de SMOTE [1] [2].

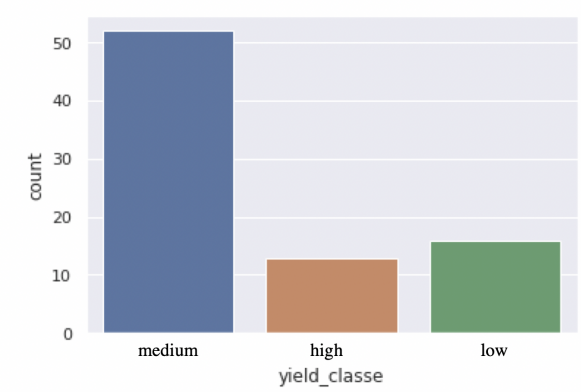


FIGURE 5 – Distribution des classes de rendement.

Afin de valider les résultats, nous avons utilisé les méthodes et métriques suivantes :

Validation des résultats par répartition *train/test* : De la même manière que pour l'étude comparative de régression, les données ont été réparties en deux groupes, l'un pour l'entraînement et l'autre pour le test. En raison de la faible quantité de données, l'ensemble d'entraînement couvre l'année 2017 tandis que l'ensemble de test correspond aux données de 2018.

Résultats Évaluation par F1-score & Accuracy : Nous avons évalué nos modèles de classification par les deux formules d'évaluation les plus connues : le score F et la précision (*accuracy*) de la classification. Le score F, également appelé mesure F, est basé sur les deux mesures principales : la précision (p*recision*) et le rappel (*recall*). La précision est la proportion de cas que le sujet a classés comme positifs et qui étaient vraiment positifs (TP - *true positive*). Elle est équivalente à la valeur prédictive positive. Le rappel est la proportion de cas vraiment positifs qui ont été classés comme positifs par le modèle. Il est équivalent à la sensibilité.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Où TP est le nombre de vrais positifs, TN est le nombre de vrais négatifs, FP est le nombre de faux positifs et FN est le nombre de faux négatifs.
Les deux métriques sont souvent combinées sous la forme de leur moyenne harmonique [6] appelé F-Score. La métrique F-score peut être utilisée pour équilibrer la contribution des faux négatifs en pondérant le rappel par un paramètre $\beta \geq 0$. Dans notre cas, β est fixé à 1, le score F1 est alors égal à :

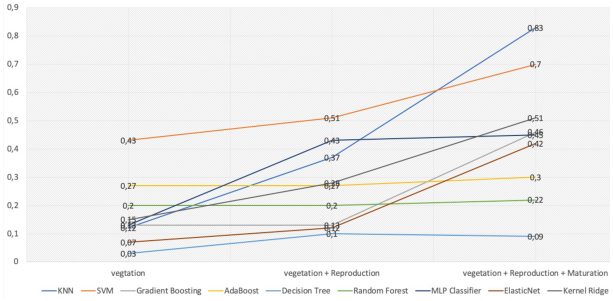


FIGURE 6 – Comparaison R^2-score.

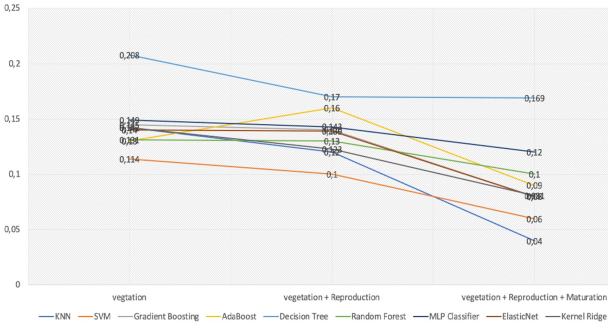


FIGURE 7 – Comparaison RMSE.

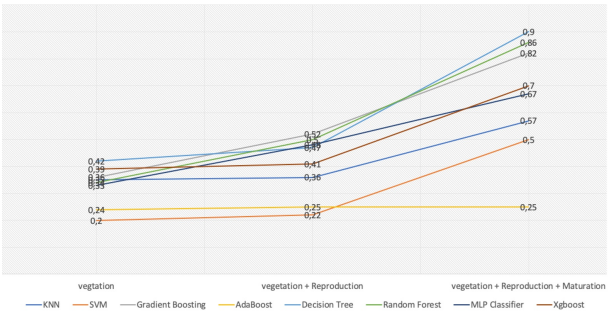


FIGURE 8 – Comparaison F-score.

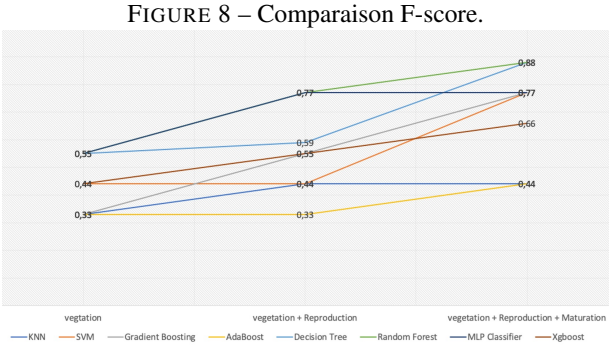


FIGURE 9 – Comparaison accuracy.

$$F1\_score = \frac{2 \times recall \times precision}{precision + recall}$$

Finalement, nous utilisons la métrique *accuracy* (traduite par exactitude ou justesse), l'un des critères les plus connus pour évaluer les modèles de classification. D'une manière non formelle, elle se réfère à la proportion de prédictions correctes faites par le modèle. Sa formule est la suivante :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

4 Résultats et Discussion

4.1 Le cas de la Régression

Après avoir effectué l'étape de prétraitement, les résultats obtenus à partir des méthodes de régression sont résumés dans le tableau 1.

En synthétisant les résultats obtenus et mentionnés dans le tableau, nous pouvons voir sur la figure 6 la comparaison du score R^2 des méthodes de régression au cours des trois phases de maturation du mil : végétative, reproductive et de maturation. De même manière, la figure 7 compare ces méthodes selon la métrique RMSE.

A partir de ces résultats, nous pouvons voir que les meilleurs scores pour les prédictions de rendement dans les phases végétative et reproductive sont donnés par la méthode SVM, alors que dans la phase de maturation les meilleurs scores sont donnés par la méthode du régresseur K plus proche voisin. Cela est en partie dû à la plus grande quantité de données disponibles lorsqu'on réduit le temps avant la récolte : plus nous ajoutons de nouvelles données aux cycles de vie du mil, plus la précision des méthodes de régression augmente.

De ce fait, l'erreur de prédiction diminue de RMSE = 0,11, équivalent à une erreur de 400 kg/ha si la prédiction n'inclut que des données de la phase végétative, à un RMSE de 0,04, correspondant à une erreur de prédiction de 140 kg/ha lors de la phase de maturation.

4.2 Le cas de la Classification

Après l'étape de prétraitement, les résultats obtenus par les méthodes mises en œuvre en utilisant les méthodes de validation et d'évaluation expliquées précédemment sont résumés dans le tableau 2. En schématisant les résultats obtenus et mentionnés dans le tableau, nous pouvons voir sur la figure 8 la comparaison du score R2 des méthodes de classification au cours des trois phases du cycle de vie du mil : végétative, reproductive et de maturation. De même, la figure 9 compare ces méthodes en évaluant leur précision.

D'après ces résultats, nous pouvons voir que les meilleurs scores pour la prédiction de la classe de rendement en utilisant uniquement les données végétatives (c'est-à-dire 5 mois avant la récolte) sont donnés par la méthode de l'arbre de décision, qui atteint une précision de 0,56. De même, dans la phase de maturation (lorsque l'on utilise tous les stades du cycle de vie du mil), les meilleures prédictions sont faites par le modèle d'arbre de décision avec 90% du score F et 88% de l'accuracy, suivi de près par la méthode Random Forest.

Comme nous nous y attendions, la précision de prédiction augmente de F-score = 0,42 pendant la phase végétative, à F-score = 0,90, ce qui signifie une erreur de prédiction de

Approche ML	Phase végétative		Végétative+reproduction		Toutes les trois phases	
	R2-score	RMSE	R2-score	RMSE	R2-score	RMSE
K-Nearest Neighbors	0.12	0.14	0.37	0.12	**0.83**	**0.04**
Support Vector Machine	**0.43**	**0.11**	**0.51**	**0.10**	0.70	0.06
Gradient Boosting	0.13	0.14	0.13	0.14	0.46	0.08
Ada Boosting	0.27	0.13	0.27	0.16	0.30	0.09
Decision Tree	0.03	0.20	0.10	0.17	0.09	0.16
Random Forest	0.20	0.13	0.20	0.13	0.22	0.10
ElasticNet	0.07	0.14	0.12	0.14	0.42	0.08
Kernel Ridge	0.15	0.14	0.28	0.12	0.51	0.08
Multi Layer Perceptron	0.13	0.19	0.43	0.11	0.45	0.09

TABLE 1 – Comparaison de performance pour les algorithmes de régression.

Approche ML	Phase végétative		Végétative+reproduction		Toutes les trois phases	
	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy
K-Nearest Neighbors	0.35	0.33	0.36	0.44	0.57	0.44
Support Vector Machine	0.20	0.44	0.22	0.44	0.50	0.77
Gradient Boosting	0.36	0.33	**0.52**	0.55	0.82	0.77
Ada Boosting	0.24	0.33	0.25	0.33	0.24	0.44
Decision Tree	**0.42**	**0.56**	0.47	0.59	**0.90**	**0.88**
Random Forest	0.34	0.55	0.50	**0.77**	0.86	**0.88**
Xgboost	0.39	0.44	0.41	0.55	0.70	0.66
Multi Layer Perceptron	0.20	0.55	0.50	**0.77**	0.72	0.77

TABLE 2 – Comparaison des performances des algorithmes de classification.

classe de 10% pendant la phase de maturation (c.-à-d. un mois avant la récolte).

4.3 Discussion

Les résultats obtenus dans cette étude se révèlent particulièrement satisfaisants, surtout lorsque l'on considère les limitations auxquelles nous avons dû faire face. Premièrement, notre ensemble de données d'entraînement était assez restreint, ne couvrant qu'une période de deux ans. Malgré cette contrainte temporelle, les performances de notre modèle ont été prometteuses, ce qui témoigne de son potentiel même avec des données limitées. De plus, un défi majeur auquel nous avons été confrontés était le manque de données météorologiques et de données de sol. Ces informations sont cruciales pour modéliser avec précision les rendements agricoles, mais malheureusement, leur disponibilité était limitée dans notre contexte. Malgré ces obstacles, les résultats que nous avons obtenus soulignent l'efficacité de notre approche méthodologique et suggèrent des possibilités futures pour améliorer encore davantage la prédiction des rendements agricoles dans des conditions de données similaires.

5 Conclusions et Perspectives

Cette étude montre l'application de méthodes d'apprentissage automatique en particulier afin d'améliorer l'estimation des rendements pour des paysages agricoles complexes, en utilisant des images satellites optiques à haute résolution spatiale et temporelle.

Dans un premier temps, nous avons exploré les méthodes de régression pour obtenir des estimations assez précises, avec un R2 score qui atteint 0.83 un mois avant la récolte. Dans un deuxième temps, et pour affiner encore la précision des prédictions, nous avons mis en œuvre des méthodes de classification supervisée. Grâce à cela, nous avons obtenu de bonnes prédictions de rendement des cultures avec une précision de 90% pour la classe de rendement un mois avant la récolte. Enfin, nous pouvons dire que les résultats obtenus par cette étude sont vraiment satisfaisants, car nous ne disposons pas d'un grand ensemble d'entraînement (seulement deux années de données).

Comme perspective, nous avons l'intention d'enrichir cette base de données par l'historique des années subséquentes, et aussi d'essayer de généraliser cette application pour d'autres cultures telles que le blé ou le maïs. Nous prévoyons également de croiser les cartes satellites avec les cartes météorologiques afin d'implémenter des réseaux de neurones profonds et d'étudier leurs comportements pour la prédiction des rendements.

Références

[1] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote : synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16 :321–357, 2002.

[2] Amine Chemchem, François Alin, and Michaël Krajecki. Combining smote sampling and machine learning for forecasting wheat yields in france. In *2019 IEEE Second International Conference on Artificial Intelli-*

gence and Knowledge Engineering (AIKE), pages 9–14. IEEE, 2019.

[3] Agricultural Research for Development Cirad. Diversité paysagère et securite alimentaire en Afrique. https://www.projects.igeo.fr/sites-d-etudes/, 2018. [Online; accessed 26-February-2020].

[4] Manal Fawzy, Mahmoud Nasr, Samar Adel, and Shacker Helmi. Regression model, artificial neural network, and cost estimation for biosorption of ni (ii)-ions from aqueous solutions by potamogeton pectinatus. *International journal of phytoremediation*, 20(4) :321–329, 2018.

[5] Lilian Hollard, Angelica Durigon, and Luiz Angelo Steffenel. Machine learning forecast of soybean yields on south brazil. In *Workshop on Edge AI for Smart Agriculture (EAISA 2022)*, 2022.

[6] George Hripcsak and Adam S Rothschild. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3) :296–298, 2005.

[7] Sergey Ioffe and Christian Szegedy. Batch normalization : Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv :1502.03167*, 2015.

[8] Elisa Kamir, François Waldner, and Zvi Hochman. Estimating wheat yields in australia using climate records, satellite image time series and machine learning methods. *ISPRS Journal of Photogrammetry and Remote Sensing*, 160 :124 – 135, 2020.

[9] K. Kuwata and R. Shibasaki. Estimating crop yields with deep learning and remotely sensed data. In *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 858–861, July 2015.

[10] Mark N Merzlyak, Anatoly A Gitelson, Olga B Chivkunova, and Victor YU Rakitin. Non-destructive optical detection of pigment changes during leaf senescence and fruit ripening. *Physiologia plantarum*, 106(1) :135–141, 1999.

[11] Babacar Ndao, Louise Leroux, Abdoul Aziz Diouf, Valerie Soti, and Bienvenu Sambou. A remote sensing based approach for optimizing the sampling strategies in crop monitoring and crop yield estimation studies. In Souleye Wade, editor, *Earth Observations and Geospatial Science in Service of Sustainable Development Goals*, pages 25–36, Cham, 2019. Springer International Publishing.

[12] A. Nyéki, C. Kerepesi, B. Daróczy, A. Benczúr, G. Milics, A.J. Kovács, and M. Neményi1. *Maize yield prediction based on artificial intelligence using spatio-temporal data*, chapter 124, pages 1011–1017. Wageningen Academic Publishers, 2019.

[13] J. Qi, A. Chehbouni, A.R. Huete, Y.H. Kerr, and S. Sorooshian. A modified soil adjusted vegetation index. *Remote Sensing of Environment*, 48(2) :119–126, 1994.

[14] M. Stas, J. Van Orshoven, Q. Dong, S. Heremans, and B. Zhang. A comparison of machine learning algorithms for regional wheat yield prediction using ndvi time series of spot-vgt. In *2016 Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*, pages 1–5, July 2016.

[15] Anna X. Wang, Caelin Tran, Nikhil Desai, David Lobell, and Stefano Ermon. Deep transfer learning for crop yield prediction with remote sensing data. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, COMPASS '18, New York, NY, USA, 2018. Association for Computing Machinery.

[16] Weicheng Wu. The generalized difference vegetation index (gdvi) for dryland characterization. *Remote Sensing*, 6(2) :1211–1233, 2014.

Performances et explicabilité de ViT et d'architectures CNN : une étude empirique utilisant LIME, SHAP et GradCam

M. Colin[1*] I. Chraibi Kaadoud[2,3]
[1] Cali Intelligences
[2] IMT Atlantique, Lab-STICC, UMR CNRS 6285, Brest, France
[3] Centre INRIA de l'Université de Bordeaux, France

melissa.colin0@proton.me ; ikram.chraibi-kaadoud@inria.fr

Résumé

Ces dernières années, l'IA explicable a été mise en avant comme la solution à plébisciter pour instaurer la confiance entre les utilisateurs et les systèmes d'IA. Pour étudier cette hypothèse, nous proposons une étude empirique sur le lien entre la performance et l'explicabilité de quatre algorithmes de vision par ordinateur : ViT, ResNet50, VGG16 et InceptionV3. Notre étude utilise trois méthodes d'explicabilité locale : LIME, SHAP et GradCam. Nous montrons que si l'IA explicable peut être un outil permettant de questionner la représentation artificielle d'un algorithme et son comportement, elle peut aussi présenter des problèmes de robustesse ou d'informations contradictoires susceptibles de miner la confiance. Les résultats de notre étude montrent que multiplier les outils d'explicabilité permet de vérifier la fiabilité des explications et des informations extraites.

Mots-clés

Explicabilité locale, Vision par ordinateur, Réseaux de neurones convolutifs, Vision Transformers, LIME, SHAP, GradCam

Abstract

In recent years, explainable AI has been presented as the main solution for building trust between users and AI systems. To investigate this hypothesis, we propose an empirical study on the link between the performance and explainability of four computer vision algorithms : ViT, ResNet50, VGG16 and InceptionV3. Our study uses three local explainability methods : LIME, SHAP and GradCam. We show that, while explainable AI can be a tool for challenging the artificial representation of an algorithm and its behavior, it can also present robustness problems or contradictory information that undermines trust. Our results show that by multiplying the use of explainable AI algorithms to explain one prediction, it is possible to verify the reliability of the explanations and extracted information.

Keywords

Local explainability, Computer vision, Convolutional neural networks, Vision Transformers, LIME, SHAP, GradCam

*Contact author

1 Introduction

1.1 Contexte

La transparence d'un système d'Intelligence Artificielle (IA) concerne, selon l'Union européenne, les données, le système en lui-même et les modèles économiques associés [1]. À ce titre, l'IA explicable, que nous noterons XAI [1] pour *explainable AI*, est devenue une exigence technique nécessaire pour établir la confiance entre les utilisateurs et les systèmes d'IA et ainsi atteindre une IA digne de confiance [1, 2]. Plus précisément, l'explicabilité et l'intelligibilité dans l'IA sont devenues un aspect important de la conception de systèmes d'IA acceptables [3], et ont été reconnues comme beaucoup plus importantes que la performance pure dans les systèmes d'IA [4, 5].

Or, depuis peu, la génération d'explication est l'objet d'étude, car : (i) Bien que diverses techniques d'XAI existent, il n'existe pas de solution universelle et le choix de l'approche d'XAI dépend de facteurs tels que la complexité du modèle, les données disponibles, le public cible et le domaine métier étudié [6], (ii) chaque algorithme d'explicabilité possède lui-même au niveau technique ses propres avantages et limites [7], (iii) les explications générées résultent elles-mêmes d'un compromis "complétude vs intelligibilité" qui peut introduire des biais d'interprétation ou de la désinformation [8].

Dans le domaine de la vision par ordinateur, les algorithmes d'explicabilité peuvent être utilisés pour : (i) exposer le fonctionnement interne du modèle expliqué au moyen de visualisations ou d'explications graphiques, ce qui peut aider les utilisateurs à comprendre comment le modèle traite les données et prend des décisions, (ii) identifier les caractéristiques d'une image qui contribuent le plus aux prédictions d'un modèle, en évaluant l'impact de chaque caractéristique sur les performances du modèle, (iii) mettre en évidence des caractéristiques spécifiques (pixels ou régions de l'image) qui ont conduit à une décision particulière afin d'expliquer des prédictions individuelles (répondre à la question "comment le modèle est-il parvenu à ce résul-

1. Nous utiliserons le sigle XAI dans la suite de cet article, car largement utilisé dans les communautés francophone et anglophone associées à ce domaine.

tat ?") [9]. Il s'agit alors d'explicabilité locale qui consiste à fournir une explication pour un résultat précis, i.e. pour une décision en particulier sur une échelle très réduite [10].

1.2 Positionnement et motivation

Nos travaux s'inscrivent dans ce dernier axe. Nous proposons ici de réaliser une étude comparative empirique sur les explications locales visuelles générées par 4 algorithmes différents de classifications d'image. Nos principales hypothèses sont les suivantes : (i) les explications visuelles issues de plusieurs algorithmes de classification d'image seront similaires pour une même classification ; (ii) les explications visuelles explicitent la représentation du monde de l'algorithme expliqué ; (iii) il est possible de générer des explications résultant de la fusion de plusieurs algorithmes d'XAI.

Nous focalisons notre étude sur deux familles d'algorithmes de classification d'images : les réseaux de neurones convolutifs (CNN pour *Convolutional Neural Networks*) et les *Vision Transformers* (ViT) [2]. Les expériences menées ont pour objectif de comparer les explications fournies par 3 algorithmes d'XAI appliqués à ces deux familles d'algorithmes de vision par ordinateur pour expliquer une prédiction, de tester les limites des différentes méthodes d'explicabilité, de comparer les performances et les ressources nécessaires pour l'ensemble des algorithmes, et de questionner l'utilisabilité des algorithmes d'XAI pour comprendre les différences entre les architectures convolutives et Transformers dans leurs façons de classer une image. Précisons ici que nos travaux se focalisent sur les explications des résultats des modèles et non les composants des modèles et leur rôle dans le comportement ou l'explicabilité de ces modèles.

1.3 Structure de l'article

Cet article est structuré de la manière suivante : La section 2 décrit la méthodologie utilisée. Les résultats expérimentaux sont présentés et discutés dans la section 3 au niveau des performances et de l'explicabilité, et une conclusion est donnée dans la section 4.

2 Méthodologie

L'objectif de notre étude est de comparer empiriquement les résultats d'algorithmes d'IA explicable locale sur deux familles d'algorithmes de vision par ordinateur, et non d'obtenir des résultats de pointe.

Par conséquent, nos expériences sont conçues pour que la configuration soit simple et qu'elles permettent une analyse comparative entre les résultats obtenus. Nous décrivons ci-dessous les données, les architectures et les algorithmes d'XAI utilisés pour permettre la reproductibilité des travaux présentés.

L'ensemble des expérimentations ont été effectuées sur la plateforme de calcul Google Colab [3] avec la configuration technique par défaut utilisant un CPU.

2. Nous utiliserons dans la suite de l'article les sigles CNN et ViT, car largement utilisés dans les communautés IA francophones et anglophones.
3. https ://colab.research.google.com/ ?hl=fr

2.1 Dataset

Le dataset utilisé dans notre étude est **un sous-ensemble de Asirra** (*Animal Species Image Recognition for Restricting Access*)[11] une base de données d'images de chiens et de chats annotées, soit deux classes d'objets [12] [4].

2.2 Algorithmes de classification d'images

Nous focalisons notre étude sur deux familles d'algorithmes populaires pour le traitement d'images : les CNN, qui sont historiquement les réseaux de neurones connus pour leur fiabilité en vision par ordinateur [13, 14], et les ViT, architectures plus récentes qui ont su démontrer de nombreuses fois de hautes performances [15].

2.2.1 Réseau de neurones convolutionnels

Les CNN reposent sur un empilement de couches de traitement [13]. Nous décrivons ici le fonctionnement des principales : la couche de convolution et la couche de pooling [5]

La couche de convolution repose sur le principe de convolution : un procédé de traitement d'image consistant en une opération de multiplication de deux matrices de tailles différentes mais de même dimensionnalité, correspondant respectivement à l'image en entrée et au *kernel* (le filtre ou fenêtre de convolution), afin d'en produire une nouvelle matrice, i.e. l'image filtrée, également de même dimensionnalité. Ce principe permet un traitement non coûteux pour la machine car il s'agit d'opérations simples (addition et multiplication).

Après chaque opération de convolution, une couche de pooling est généralement appliquée pour compresser l'information en réduisant la taille de l'image filtrée obtenue. Le pooling permet par ce procédé de regrouper les informations au sein de fenêtres de taille inférieure. Ce processus de convolution et de pooling est répété plusieurs fois en fonction de l'architecture du CNN, ce qui permet d'extraire les caractéristiques (*features*) importantes des images. Les données de sortie sont aplaties (*flattened*) pour être traitées par une couche de neurones entièrement connectés (*fully connected layer*). Traditionnellement, la sortie de cette couche est passée à une fonction softmax pour obtenir des probabilités de classe.

Différentes architectures reposant sur ces principes de convolution et de pooling existent. Dans cette étude, nous nous sommes intéressés particulièrement aux modèles suivants : ResNet50 [16], VGG16 [17], InceptionV3 [18]. Le lecteur pourra se référer à [19] pour une étude comparative des détails techniques de ces architectures.

2.2.2 Vision Transformers

L'architecture des ViT [15], schématisée dans la figure 1, repose sur le même principe que les Transformers utilisés dans le traitement du langage naturel [20].

Ce principe peut être décrit comme suit :

(i) Dans un premier temps, l'image est segmentée en plusieurs parties de taille fixe, nommées patchs, qui sont linéarisés (figure 1.a). Cette opération linéaire attribue un poids à chaque pixel du patch, représentant de façon

4. https ://zenodo.org/records/5226945 (Accès le 05/04/2024)
5. Pour plus de détails techniques, le lecteur pourra se référer à [13].

numérique chaque portion de l'image sous forme de jetons, ou *tokens*. Pour conserver l'information sur la localisation des patchs dans l'image, des embeddings de position sont ajoutés à chaque jeton, fournissant une indication précise de leur emplacement (figure 1.b).
(ii) Ensuite, le Transformer composé de n blocs (au nombre de 6 dans le modèle original) agit en tant qu'encodeur au travers de la répétition des 2 étapes suivantes [20] au sein de chaque bloc :
(ii.1) Utilisation d'un mécanisme d'attention multi-tête (*Multi-Head Attention, MHA*) [6] pour calculer le score d'attention entre chaque jeton par rapport aux autres (figure 1.d). Cela permet de déterminer l'importance relative de chaque jeton par rapport aux autres dans la représentation latente de l'image. Les jetons ayant les scores d'attention les plus élevés auront une contribution plus importante dans la génération de la représentation latente globale de l'image. Cette étape permet de mettre en évidence les parties de l'image les plus importantes pour la tâche en renforçant l'attention du modèle sur ces zones ;
(ii.2) Utilisation d'un réseau de neurones à propagation avant (*feedforward*) appliqué à la sortie du mécanisme d'attention multi-tête (figure 1.e). Ce réseau est constitué de transformations linéaires et non linéaires qui agissent sur les représentations des patchs. Cela permet au modèle de capturer des relations plus complexes entre les différentes régions de l'image et extraire des caractéristiques discriminantes des images, telles que les contours, les textures et les motifs significatifs.
(iii) Les sorties du Transformer sont finalement envoyées à un perceptron multicouche qui renvoie les probabilités finales pour chaque classe (figure 1.f).

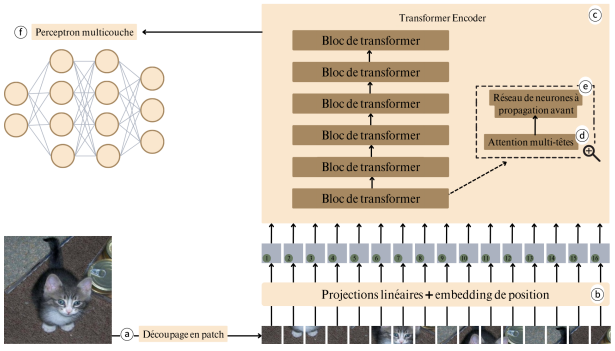


FIGURE 1 – Schéma de l'architecture ViT

2.3 Explicabilités d'XAI

Pour notre étude comparative, nous avons implémenté plusieurs méthodes d'explicabilité locale.

6. Le mécanisme d'attention Multi-tête est une extension du mécanisme d'auto-attention (*Self-Attention*), où au lieu de calculer une seule fonction d'attention, plusieurs fonctions d'attention sont calculées en parallèle, chacune avec sa propre matrice de poids.

2.3.1 LIME

LIME, pour *Local Interpretable Model-agnostic Explanations*, appartient à la famille des modèles de substitution locaux qui sont formés pour approximer les prédictions individuelles du modèle que l'on cherche à expliquer [21]. LIME peut être appliquée aux données tabulaires, aux textes et aux images. Pour ces dernières, LIME segmente l'image associée à la prédiction à expliquer en "superpixels" et active ou désactive ces derniers afin de créer des variations de cette image. Ensuite, LIME prédit la classe de chacun des points de données artificiels qui ont été générés à l'aide de notre modèle entraîné. Par superpixels, on entend des pixels interconnectés avec des couleurs similaires, qui peuvent être désactivés en remplaçant chaque pixel par une couleur définie par l'utilisateur, telle que le gris. L'utilisateur peut également spécifier une probabilité de désactivation d'un superpixel dans chaque permutation [22].
LIME permet d'utiliser différentes méthodes de segmentation telles que SLIC, Quickshift et Felzenszwalb [23] afin de diviser l'image à expliquer en superpixels ce qui facilite un traitement plus rapide des images.

2.3.2 GradCam

GradCam, pour *Gradient-weighted Class Activation Mapping*, est une technique permettant de produire des "cartes de chaleur" (*heatmaps*). Cet algorithme a été initialement créé pour les modèles CNN en utilisant les informations de gradient spécifiques à la classe qui circulent dans la dernière couche convolutionnelle du modèle [24]. Cependant, cette méthode a démontré des résultats tout aussi prometteurs sur les ViT en se focalisant sur le jeton de la dernière couche d'attention [25]. Quel que soit l'architecture sur laquelle GradCam est appliquée, cette méthode utilise les gradients pour peser l'importance de chaque neurone dans la couche en question, ce qui permet d'identifier les régions dans l'image d'entrée qui ont le plus contribué à la prédiction de la classe.

2.3.3 SHAP

SHAP pour *SHapley Additive exPlanations* permet d'expliquer les prédictions d'un modèle en attribuant à chaque caractéristique une valeur représentant son impact sur la prédiction [26]. Dans le contexte d'une image, une caractéristique peut être une valeur de pixel individuel, une texture, une forme, une couleur, une orientation, etc. Cette méthode est basée sur la théorie des jeux coopératifs et est capable de fournir des explications locales et globales pour les prédictions d'un modèle. Pour chaque classification, SHAP génère deux images qui correspondent aux masques expliquant l'implication des superpixels de l'image d'entrée dans les classes considérées pour la classification. Ces masques se présentent sous forme de superpixels allant de bleu à rouge, pour représenter respectivement ceux impactant négativement et ceux impactant positivement la classification. L'échelle de valeurs apporte un poids objectif à chaque contribution. Notons que pour les modèles complexes tels que les ViT, l'utilisation de SHAP peut nécessiter une grande quantité de mémoire RAM (au-delà de 13

Go) et de temps de calcul en raison du grand nombre de caractéristiques à traiter. Dans certains cas, cela peut rendre l'exécution de SHAP difficile ou même impossible.

2.4 Protocole expérimental

Notre travail se concentre sur la comparaison des perceptions des algorithmes d'IA appliquées au domaine de la vision par ordinateur au travers des explications générées. Pour ce faire, notre protocole expérimental est le suivant :

Étape (1) Préparation du dataset Asirra pour le fine-tuning : Cet ensemble de données a été segmenté en 348 images pour le dataset de test, 753 images pour le dataset d'entraînement et 273 images dans le dataset de validation avec 50% d'images de chats et 50% d'images de chien, soit un total de 1374 images utilisées pour notre étude. L'ensemble des images ont été redimensionnées avec la taille suivante : 224×224

Étape (2) Chargement des algorithmes de vision par ordinateur pré-entraînés : les 4 algorithmes pré-entrainés sur le dataset ImageNet ont été téléchargés et intégrés dans l'environnement d'expérimentation.

Étape (3) Fine-tuning et évaluation des algorithmes : Chaque algorithme est fine-tuné, i.e. son apprentissage est ajusté selon le nouveau dataset, pendant 10 epochs sur un batch de 8. Les performances des algorithmes sur les dataset de validation lors de l'entraînement sont sauvegardées pour connaître l'évolution de celui-ci. Et le modèle final est évalué à travers les données de tests pour en ressortir les matrices de confusion et les valeurs des *accuracy*.

Étape (4) Génération d'explications locales pour chaque algorithme : Pour chaque classification à expliquer :

• Pour LIME, nous avons créé une instance d'explication qui contient des détails sur les pixels qui ont contribué à la prédiction en générant : (i) une image, (ii) un masque et (iii) une combinaison de l'image et du masque qui peut être tracée pour voir quels pixels ont contribué à la prédiction. Parmi les différentes méthodes de segmentation, nous avons utilisé celle par défaut, i.e. Quickshift, et nous avons également calculé l'intersection entre les masques générés avec un coefficient de similarité de Jaccard.

• Pour SHAP, nous avons généré une explication avec (i) l'image, (ii) les pixels ayant contribué à la classification dans la classe "chien" et (iii) les pixels ayant contribué à la classification dans la classe "chat".

• Pour GradCam, nous avons généré la carte de chaleur qui explique la classification réalisée par l'algorithme de vision par ordinateur.

Notons ici que SHAP, par défaut, permet d'expliquer le rôle de chaque bloc de pixels dans la classification dans chacune des 2 classes considérées, alors que GradCam et LIME fournissent une explication liée à une classification réalisée au travers respectivement d'une heatmap et d'un masque. Précisons qu'un pixel qui contribue positivement à une classification implique qu'il joue un rôle positif dans le comportement de l'algorithme à classer la donnée en entrée dans la classe prédite. Inversement, un pixel qui contribue négativement à une classification "pousse" l'algorithme à ne pas classer la donnée en entrée dans la classe prédite. Tous nos algorithmes d'explicabilité locale offrent la possibilité d'expliquer visuellement les pixels qui contribuent positivement ou négativement à chaque classification.

3 Résultats et discussions

Dans cette section, nous présentons et comparons les performances des 4 algorithmes étudiés et les explications locales générées pour une classification en particulier, celle d'une image de chat présentée en figure 4, avant de les discuter en détail.

3.1 Comparaison des performances des algorithmes

Les performances des différents algorithmes ont été évaluées sur le dataset de test. Les mesures d'évaluation comprennent l'exactitude des prédictions, i.e. *accuracy*, le temps d'inférence, la durée du fine-tuning et le nombre d'epoch nécessaires pour atteindre l'exactitude maximale sur les données de validation avant 10 epoch.

La figure 2 représente les performances des algorithmes selon la métrique *accuracy*, le temps d'inférence en seconde nécessaire à réaliser une classification et le temps de Fine-Tuning en minute.

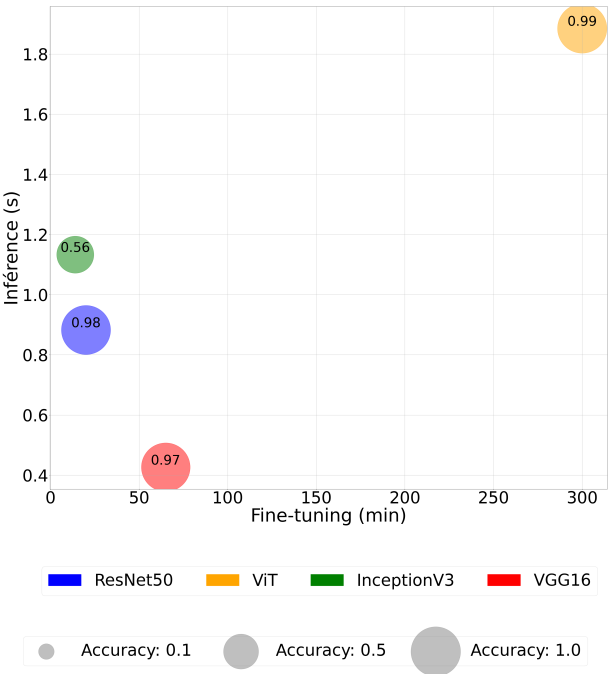


FIGURE 2 – Représentation des performances des algorithmes à travers la valeur de *accuracy*, selon le temps d'inférence en secondes nécessaire à réaliser une classification et le temps de Fine-Tuning en minute. Les valeurs des *accuracy* sont affichées dans les cercles.

De l'analyse des résultats présentés en figure 2, il ressort que :

• le ViT est l'architecture la plus performante avec une *accuracy* à 0,99, même si relativement proche du ResNet50

qui est à 0,98. Il reste important de noter que la différence de performance entre le ResNet50 et le ViT peut sembler minimale dans ce contexte, étant donné qu'ils tendent tous les deux vers 1.
• le ViT se démarque par son temps de fine-tuning de 300 minutes et d'inférence de 1,89 secondes qui sont plus importants que ceux des autres algorithmes qui sont tous fine-tunés en moins de 100 minutes et qui infèrent en moins de 1,2 secondes.
• le VGG16 se démarque des autres algorithmes en termes de temps d'inférence avec 0,43 secondes, cependant il présente une valeur *accuracy* similaire à celle du ResNet50 et au ViT avec une différence respective de 0,02 et 0,01.

Algorithme	**Temps mesuré pour le Fine-tuning sur 10 epoch en minutes**	**Temps estimé pour atteindre *accuracy* maximal en minutes**	**Epoch où *accuracy* est maximal**
ViT	300	45	1,5
ResNet50	20	10	5
InceptionV3	14	11,2	8
VGG16	60	30	5

TABLE 1 – Temps mesuré et estimé pour respectivement le fine-tuning des algorithmes, et l'atteinte de *accuracy* maximale lors de ce fine tuning avec les données de validation.

Le tableau 1 affiche le temps mesuré lors du fine-tuning sur 10 epoch, le nombre d'epoch nécessaire pour que chaque algorithme atteigne son *accuracy* maximale et le temps estimé pour atteindre cet epoch. Ce dernier temps est calculé à partir des deux précédentes informations. L'analyse de ces différents temps montre que : le ViT met le plus de temps à s'ajuster aux données lors de la phase de fine-tuning (300 min mesurée), alors qu'il est celui qui atteint son *accuracy* maximal avec un nombre d'epoch minimal (1,5 epoch) en seulement 45 min, ce qui représente 15% du temps total. Les autres architectures quant à elles nécessitent 50% à 80% du temps de fine-tuning pour atteindre leur *accuracy* maximale.

Les matrices de confusion de ViT, ResNet50, VGG16 et InceptionV3, figurant respectivement dans la figure 3.(a), (b), (c) et (d), permettent de comprendre plus en détail les différences de performances entre ces différents algorithmes de classifications. Elles permettent d'observer que ViT, ResNet50 et VGG16 rencontrent des difficultés sur la même classe : des images de "chat" sont classées en "chien" (4 images de chats sont mal classées pour ResNet50, contre 1 image pour ViT et 9 pour VGG16), ce qui suggère des points communs dans les éléments qui impactent positivement et négativement ces algorithmes. La matrice de confusion de InceptionV3 illustre bien que l'algorithme est globalement moins performant que les autres, mais également qu'il est plus performant sur la classe "chien" que "chat".

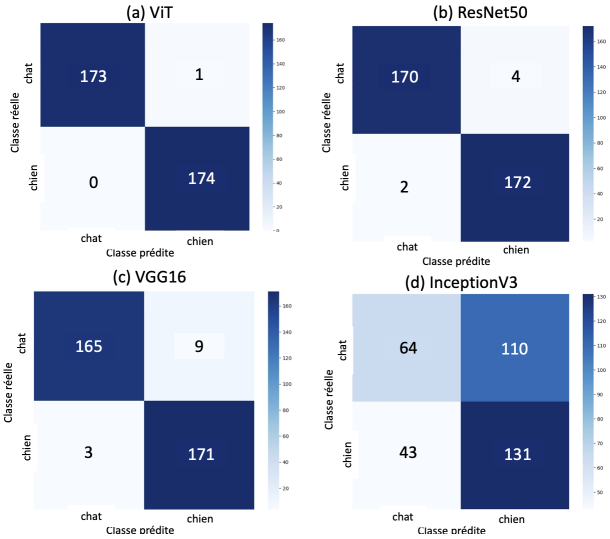


FIGURE 3 – Matrice de confusion : (a) ViT, (b) ResNet50, (c) VGG16 et (d) InceptionV3

3.2 Comparaison des résultats d'XAI

Nous présentons ici les explications locales générées sur les algorithmes étudiés. À des fins d'intelligibilité, nous présentons les résultats appliqués à une même donnée d'entrée (l'image de chat présentée en figure 4) afin de pouvoir comparer les explications générées par les différents algorithmes. Notons que :
(i) les algorithmes d'XAI LIME et GradCam ont pu être implémentés sur l'ensemble des 4 algorithmes, alors que SHAP n'a pu l'être que sur 3 (pas d'explicabilité SHAP pour ViT) à cause de limites techniques
(ii) une comparaison des explications a été réalisé quand la classification est correcte (i.e. l'image de chat est correctement classée dans la catégorie chat) pour les algorithmes ViT, ResNet50, et VGG16, et lorsque la classification est erronée avec l'algorithme InceptionV3.
(iii) un coefficient de similarité de Jaccard J a été calculé entre les masques générés par LIME par pour chacun des trois algorithmes ayant correctement classé l'image de chat (figure 4) : ViT, ResNet50 et VGG16. Il s'agit d'un coefficient utilisé pour comparer la similarité entre deux ensembles. Il est défini comme étant le rapport de la taille de l'intersection de deux ensembles et de la taille de l'union de ces mêmes ensembles.

FIGURE 4 – Image de chat utilisée dans notre étude et pour l'ensemble des explications générées.

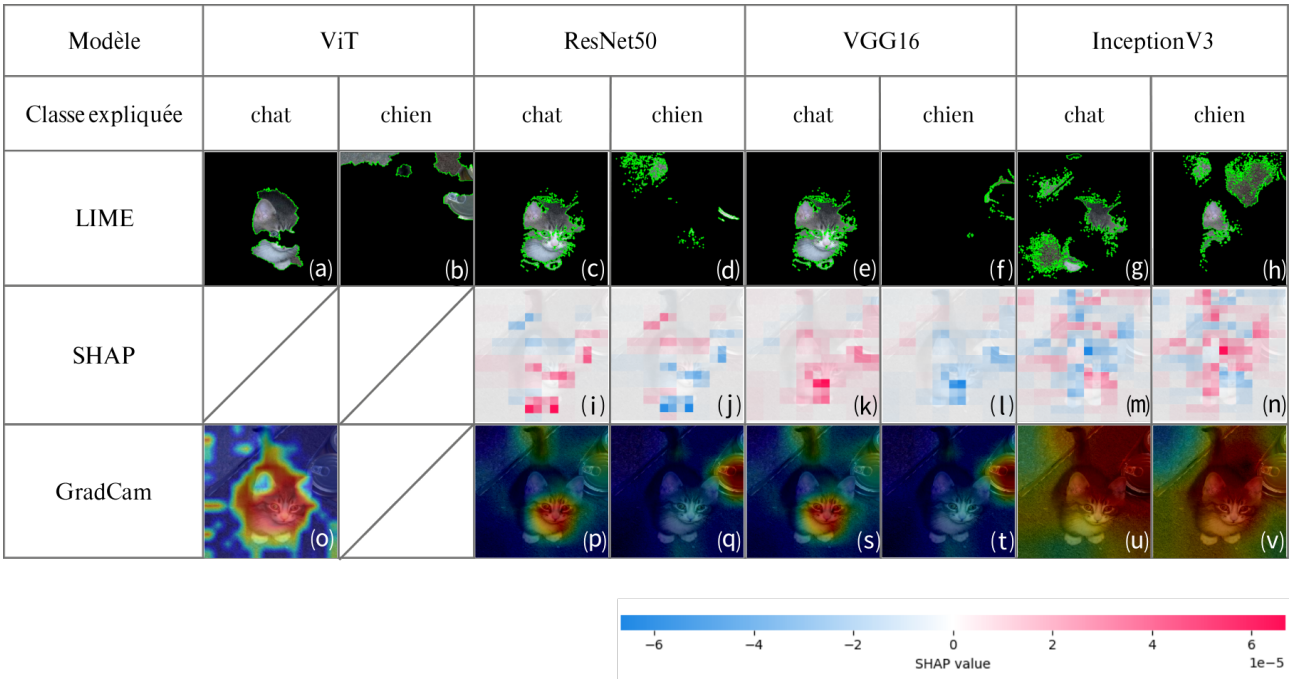


FIGURE 5 – Résultats des explications générées par LIME, SHAP et GradCam pour les classes "chat" et "chien" des modèles ViT, ResNet50, VGG16 et InceptionV3 sur l'image figure 4. Considérant ici pour LIME que les pixels ayant impacté négativement la prédiction faite par le modèle comme impactant positivement la classe non prédite.

3.2.1 Résultats LIME

L'analyse des explications fournies par LIME permet de mieux comprendre les différences entre les différents CNN et le ViT. En appliquant la méthode de segmentation Quickshift avec LIME sur l'image de chat présentée en figure 4, on observe des similitudes entre VGG16 et ResNet50. Sur l'ensemble des images testées, ces algorithmes sont expliqués par des masques identiques présentant les pixels impactant positivement la prédiction. Les masques présentés en figure 5.c et 5.e permettent de constater que les deux modèles se concentrent sur les mêmes zones de l'image pour prédire la classe correcte.

Les pixels impactant négativement les prédictions sont quant à eux différents entre ces deux modèles comme nous pouvons l'observer sur les masques 5.d et 5.f : les parties de l'image mises en évidence par les masques sont non seulement différentes mais tendent à signifier que l'arrière-plan est important dans ces deux cas de figure. En revanche, l'analyse des erreurs de prédiction du modèle InceptionV3 qui est moins performant, permet de constater (figure 5.h) que celui-ci se base souvent sur des parties du fond de l'image pour prendre sa décision. L'analyse des pixels ayant impacté négativement cette mauvaise prédiction, représentée figure 5.g, montre que l'algorithme InceptionV3 n'a pas pris en considération pour sa classification "chat" des parties essentielles du chat telles que le visage et le corps du chat (zone soulignée importante pour d'autres algorithmes, mais également qu'il a pris en compte des sections environnantes (l'arrière-plan à gauche et derrière le chat).

Lors de la comparaison entre ViT et l'ensemble des CNN, une différence significative émerge (figure 5.a et figure 5.b). Contrairement aux CNN qui analysent les pixels de manière unitaire, le ViT traite des parties d'images bien distinctes.

	Similarité de Jaccard J
ViT et ResNet50	≈ 0.577
ViT et VGG16	≈ 0.577
ResNet50 et VGG16	1.0
ViT, ResNet50 et VGG16	≈ 0.577

TABLE 2 – Similarité de Jaccard entre les masques des figures 5.a, figure 5.c et figure 5.e : LIME appliqué aux 3 algorithmes ResNet50, VGG16, ViT.

Pour étudier les similitudes entre les explications LIME de ViT, ResNet50 et VGG16, nous avons calculé le coefficient de similarité de Jaccard J sur les matrices représentant les masques associés aux bonnes prédictions de ces algorithmes. Les résultats, Table 2, présentent les valeurs obtenues pour différentes intersections de masque. L'analyse de ces valeurs a permis de confirmer que les masques générés par LIME pour Resnet50 et VGG16 étaient les mêmes (similarité de $J = 1$), et étaient eux-mêmes relativement similaires à celui généré pour ViT avec un coefficient à $J \approx 0.577$. L'intersection de l'ensemble de ces 3 masques est représentée visuellement dans la figure 6 et son ana-

lyse a mis en évidence que ces modèles utilisent une même partie de l'image de chat pour leur classification, à savoir, l'oreille gauche et un bout du museau. À l'inverse, aucun élément commun n'a été identifié dans ce qui impacte négativement la détection de ces 3 modèles avec un coefficient nul entre chacun des masques.

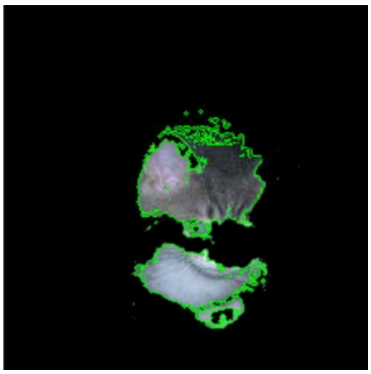


FIGURE 6 – Intersection des masques des figures 5.a, figure 5.c et figure 5.e : LIME appliqué aux 3 algorithmes ResNet50, VGG16, ViT

3.2.2 Résultats SHAP

SHAP a donné des résultats différents pour le ResNet50 notamment, comme le montre la figure 5.i et 5.j.
Des éléments similaires à ceux mis en avant par LIME ont été identifiés pour détecter la classe "chat", tels que les pattes et le visage, mais selon SHAP, l'oreille gauche impacte négativement la classification de l'image, et non positivement. De plus, SHAP met en évidence que ResNet50 prend aussi en compte une partie d'un objet (une conserve) présent dans l'image, alors que LIME non.
Cette explication se recoupe avec celle générée pour VGG16, à la différence que le masque généré par SHAP semble se focaliser plus sur le visage du chat et n'est que très peu impacté négativement par d'autres éléments de l'image.
Pour le modèle InceptionV3 qui s'est trompé dans la classification, les résultats de SHAP en figure 5.m et 5.n diffèrent de ceux de LIME en figure 5.g et 5.h : les caractéristiques déterminantes sont très parsemées et selon une grosse partie de l'image (en haut à droite de la tête du chat) a amené à classifier l'image comme "chien. Aucune explication du modèle ViT avec SHAP n'a pu être générée, car il nécessitait un trop grand nombre de caractéristiques à traiter pour être supporté par la RAM que nous avions à disposition.

3.2.3 Résultats GradCam

Les explications résultantes de GradCam sur le ResNet50 nuancent les résultats SHAP. Les figures 5.p et 5.q, présentant ces explications, montrent que la conserve aurait un impact négatif sur la prédiction car elle est de couleur claire sur la figure 5.p (qui met en avant les zones de chaleur pour la classe chat) et est mise en exergue avec des couleurs chaudes sur la figure 5.q (représentant les zones de chaleur pour la classe chien). Une zone orange sur le nez et les yeux du chat et jaune pour le contour du visage, constitue cependant un point commun avec les autres méthodes d'explicabilité.
La figure 5.s présentant le résultat de GradCam sur VGG16 montre une explication très similaire au ResNet50, ce qui rejoint les explications données par LIME et présentées dans la figure 5.e. La mauvaise prédiction donnée par InceptionV3 continue à se confirmer avec la carte de chaleur générée par GradCam. Une zone de chaleur qui pour la classe "chat" (figure 5.u) se répand sur la droite de l'image sans cibler d'objet et qui pour la classe "chien" (figure 5.v) prend en considération une grande partie de l'image. Soulignons également que dans les deux cas la zone de chaleur se répand sur toute l'image avec très peu de couleurs froides.
De son côté, GradCam appliqué au modèle ViT (figure 5.o) a mis en avant une zone bien plus importante de l'image, prenant en compte le chat dans son ensemble et de manière assez précise, ainsi que quelques éléments du fond de l'image. L'explication ici peut être interprétée comme étant plus juste car plus "complète" que celles des autres algorithmes d'XAI.

3.3 Discussion

3.3.1 Résumé des résultats

Nous avons mené une étude comparative des performances et des explications locales générées à partir de 4 algorithmes de classification d'images : ResNet50, VGG16, InceptionV3 et ViT. Pour notre étude, nous avons mesuré le temps nécessaire aux algorithmes pour le fine-tuning, le temps d'inférence et *accuracy*. Nous avons ainsi mis en évidence que parmi ces 4 algorithmes, le ViT était le plus performant, bien que le plus long à fine-tuner, et qu'il existait des différences notables en termes de performance entre les 3 architectures CNN différentes. InceptionV3 s'est avéré être, dans le cadre de notre étude, le moins performant. Nous avons ensuite implémenté 3 algorithmes d'explicabilité locale générant des explications visuelles, LIME, SHAP et GradCam, sur chacun des algorithmes de classification d'images dans le cas des classes "chat" et "chiens". Nous avons ainsi pu étudier les explications visuelles lorsque la classification est correcte pour ViT, ResNet50, et VGG16, et lorsque la classification est erronée pour InceptionV3. Pour chaque explication, nous avons analysé les superpixels contribuant positivement et négativement à la classification émise.

3.3.2 Discussion des explications par modèle

Plus en détails, concernant ViT, l'analyse des explications générées par LIME laisse supposer que ViT prend plus de temps pour s'ajuster aux données et pour effectuer une inférence car il traite les images par blocs plutôt que par pixels individuels, ce qui nécessiterait plus de calculs pour apprendre les relations spatiales entre les différentes parties de l'image. Cette approche semble permettre au modèle d'atteindre un seuil de précision important tout en lui permettant de prendre en compte l'intégralité de l'objet d'intérêt (ici le chat) pour la classification, comme le montrent les résultats avec la méthode GradCam (figure 5.o).

Les modèles ResNet50 et VGG16 présentent beaucoup plus de points communs. Ils semblent se concentrer sur des pixels similaires de l'image pour prédire la classe correcte, ce qui peut expliquer pourquoi, malgré leurs architectures différentes, ils ont des performances largement comparables. L'analyse de l'explication de la classification erronée d'InceptionV3, montre que l'algorithme ne semble pas avoir réussi à distinguer les formes présentes sur l'image et à se concentrer sur une partie en particulier. De plus, les zones mises en évidence par les 3 algorithmes d'XAI ne semblent pas cohérentes : les zones mises en lumière sont différentes d'un algorithme d'XAI à un autre. Ces résultats peuvent être mis en lien avec *accuracy* du modèle de 0,56 bien inférieure aux autres modèles, ce qui suggère que le modèle n'a pas réussi à converger en 10 epochs. Nous pensons donc que des simulations supplémentaires seraient nécessaires dans le cas de InceptionV3 pour explorer le lien entre l'évolution de *accuracy* du modèle et la précision des explications générées. Nous pourrions alors comparer nos résultats avec les explications d'une mauvaise classification sur un algorithme de classification d'image avec une forte *accuracy*. Un tel questionnement permettrait de mieux explorer la question de la représentation latente artificielle des algorithmes.

3.3.3 Discussion des méthodes d'explicabilité

Les différentes méthodes d'explicabilité que nous avons éprouvées lors de cette étude présentent à la fois des similarités et des différences. L'importance accordée au visage du chat est notablement présente dans l'ensemble des méthodes ayant bien classé l'image de chat. Mais des différences sont aussi ressorties, comme par exemple pour ResNet50 où SHAP considère un objet (conserve) comme important à la prédiction, là où GradCam ne le fait pas.

Ces différences peuvent s'expliquer par les approches utilisées par ces méthodes pour déterminer l'importance des caractéristiques. Chaque méthode se concentre sur différents aspects des modèles. SHAP prend en compte les interactions entre les pixels, tandis que GradCam se concentre sur les gradients des sorties du modèle par rapport aux entrées. LIME et SHAP sont des méthodes d'approximation locale : elles expliquent les prédictions du modèle pour des instances individuelles, mais ont plus de mal à capturer le comportement global du modèle. Alors que GradCam utilise les gradients de la sortie du modèle par rapport à ses entrées, ce qui peut fournir une vue plus globale de l'importance des caractéristiques.

Les résultats des méthodes d'XAI utilisées dans cette étude montrent que même si les performances des modèles de classification d'images peuvent être similaires, les caractéristiques sur lesquelles ils basent leur classification (autrement dit leur façons de classifier des images) peuvent être différentes. C'est pourquoi, le choix du modèle de classification, en plus de tenir compte de ces performances, doit prendre en compte les explications qu'on peut lui fournir. Nos résultats soulignent ainsi l'importance de considérer plusieurs méthodes d'explicabilité pour comprendre le comportement des modèles boîtes noires et d'éprouver leur robustesse.

Soulignons que nous avons relancé les simulations de LIME plusieurs fois et force est de constater que l'explicabilité locale fournie, peut varier et cela sans avoir modifié les paramètres de l'algorithme d'XAI ou la donnée d'entrée. Cela soulève la question de la robustesse des explications et de son impact sur la confiance des utilisateurs dans les explications générées. Nous souhaitons préciser ici que la robustesse des explications correspond à leur résilience face à des perturbations de l'algorithme d'XAI. C'est une notion distincte de la stabilité de l'algorithme d'XAI, qui correspond à la capacité du modèle à absorber des perturbations en dessous d'un seuil critique tout en maintenant un comportement stable et cohérent. Ces deux notions sont essentielles et complémentaires, et contribuent à la fiabilité d'un algorithme d'XAI [2, 27]). Des simulations supplémentaires seraient intéressantes à mener afin d'explorer plus en détails cette question de lien entre robustesse des explications et de stabilité des algorithmes d'XAI.

Enfin, nous pensons que l'alliance de méthodes d'ablation - dédiées à l'explication du comportement de modèles IA [28] - appliquées aux algorithmes de computer vision, et d'algorithmes d'explicabilité locale serait intéressant à explorer afin d'identifier les composants des modèles jouant un rôle dans la fiabilité des explications.

4 Conclusion

Notre étude est une introduction à la question de la confiance dans les algorithmes d'IA explicable. À travers ces travaux, nous avons questionné les limites des algorithmes d'explicabilité locale appliqués à la classification d'image. Nous avons montré que l'explicabilité peut être un outil pour questionner la représentation artificielle d'un algorithme et son comportement pour une classification lorsqu'elle est correcte ou non. Cependant, pour un même algorithme de classification d'images, il est important de multiplier les outils d'explicabilité afin de vérifier la fiabilité des explications et questionner les informations extraites.

Dans nos travaux futurs, nous souhaiterions explorer plus en détail les forces et faiblesses des algorithmes d'explicabilité en élargissant : (i) notre panel d'algorithmes de classification d'images, (ii) notre étude à un domaine d'application plus critique tel que le diagnostic médical et (iii) les dataset utilisés lors pour le fine-tuning pour en évaluer l'impact sur la représentation latente des algorithmes et l'impact sur les explications générées pour corroborer - ou non - les écarts de performance observés pour nos algorithmes. Par ailleurs, nous souhaiterions également explorer la question du choix de l'algorithme d'explicabilité locale selon la criticité du métier et des données notamment via des questionnaires d'évaluations des explications soumis auprès de différents métiers et niveaux d'expertise. Cela pourrait aider à la conception d'une cartographie des méthodologies d'explicabilité locale visuelles selon le besoin métier et permettre également l'élaboration de nouveaux algorithmes d'XAI qui tireraient profit des forces des différentes méthodologies existantes. Plus globalement, cela contribuera à la

question de la confiance dans les explications et par extension, à celle de l'appropriation de ces outils par la société civile.

Références

[1] IA GEHN. Les lignes directrices en matière d'éthique pour une ia digne de confiance, 2019.

[2] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable artificial intelligence (xai) : What we know and what is left to attain trustworthy artificial intelligence. *Information fusion*, 99 :101805, 2023.

[3] Daniel S. Weld and Gagan Bansal. The challenge of crafting intelligible intelligence. *Commun. ACM*, 62(6) :70–79, 2019.

[4] David Gunning. Darpa's explainable artificial intelligence (XAI) program. In *IUI*. ACM, 2019.

[5] Alejandro Barredo Arrieta, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI) : Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*, 58 :82–115, 2020.

[6] Vinay Chamola, Vikas Hassija, A. Razia Sulthana, Debshishu Ghosh, Divyansh Dhingra, and Biplab Sikdar. A review of trustworthy and explainable artificial intelligence (XAI). *IEEE Access*, 11 :78994–79015, 2023.

[7] Hans de Bruijn, Martijn Warnier, and Marijn Janssen. The perils and pitfalls of explainable AI : strategies for explaining algorithmic decision-making. *Gov. Inf. Q.*, 39(2) :101666, 2022.

[8] Tim Miller. Explanation in artificial intelligence : Insights from the social sciences. *Artif. Intell.*, 267 :1–38, 2019.

[9] Christian Meske and Enrico Bunde. Transparency and trust in human-ai-interaction : The role of model-agnostic explanations in computer vision-based decision support. In Helmut Degen and Lauren Reinerman-Jones, editors, *Artificial Intelligence in HCI - First International Conference, AI-HCI 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19-24, 2020, Proceedings*, volume 12217 of *Lecture Notes in Computer Science*, pages 54–69. Springer, 2020.

[10] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5) :93 :1–93 :42, 2019.

[11] Jeremy Elson, John R. Douceur, Jon Howell, and Jared Saul. Asirra : a CAPTCHA that exploits interest-aligned manual image categorization. In Peng Ning, Sabrina De Capitani di Vimercati, and Paul F. Syverson, editors, *Proceedings of the 2007 ACM Conference on Computer and Communications Security, CCS 2007, Alexandria, Virginia, USA, October 28-31, 2007*, pages 366–374. ACM, 2007.

[12] Jacquemont Mikaël. Cats and dogs sample, August 2021.

[13] Jeff Heaton. Ian goodfellow, yoshua bengio, and aaron courville : Deep learning - the MIT press, 2016, 800 pp, ISBN : 0262035618. *Genet. Program. Evolvable Mach.*, 19(1-2) :305–307, 2018.

[14] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks : an overview and application in radiology. *Insights into imaging*, 9 :611–629, 2018.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words : Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[18] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society, 2016.

[19] Aravinda S Rao, Tuan Nguyen, Marimuthu Palaniswami, and Tuan Ngo. Vision-based automated crack detection using convolutional neural networks for condition assessment of infrastructure. *Structural Health Monitoring*, 20(4) :2124–2142, 2021.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30 : Annual*

Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008, 2017.

[21] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you ?" : Explaining the predictions of any classifier. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM, 2016.

[22] Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022.

[23] Tobias Huber, Benedikt Limmer, and Elisabeth André. Benchmarking perturbation-based saliency maps for explaining atari agents. *Frontiers Artif. Intell.*, 5, 2022.

[24] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam : Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.*, 128(2) :336–359, 2020.

[25] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 782–791. Computer Vision Foundation / IEEE, 2021.

[26] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30 : Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774, 2017.

[27] Ann-Kathrin Dombrowski, Christopher J. Anders, Klaus-Robert Müller, and Pan Kessel. Towards robust explanations for deep neural networks. *Pattern Recognit.*, 121 :108194, 2022.

[28] Richard Meyes, Melanie Lu, Constantin Waubert de Puiseau, and Tobias Meisen. Ablation studies in artificial neural networks. *CoRR*, abs/1901.08644, 2019.

Effacement des croyances en logique propositionnelle

N. Creignou[1], R. Ktari[1,2], 0. Papini[1]

[1] Aix Marseille Univ, CNRS, LIS, Marseille, France
[2] Université de Sfax, ISIMS, CES-Lab LR11ES49, Sfax, Tunisie

Correspondante : nadia.creignou@lis-lab.fr

Résumé

L'un des thèmes importants de la représentation des connaissances et du raisonnement en intelligence artificielle est le changement de croyances. Dans le cadre logique l'approche AGM s'est imposée comme un standard et différentes opérations de changement de croyances ont été envisagées. Alors que la révision, la contraction, et la mise à jour ont donné lieu à de nombreux travaux, l'effacement a suscité jusqu'à présent moins d'intérêt. L'effacement est à la contraction ce que la mise à jour est à la révision. L'article porte sur l'étude de l'effacement dans le cadre de la logique propositionnelle. Il prolonge l'approche de Katsuno et Mendelzon par des postulats supplémentaires capturant la minimalité du changement et propose deux théorèmes de représentation pour les opérateurs d'effacement, l'un en termes de préordres totaux sur les interprétations, l'autre en termes de préordres partiels sur les interprétations. Enfin, il complète les travaux de Caridroit, Konieczny et Marquis pour la contraction en proposant un nouveau théorème de représentation pour les opérateurs de contraction en termes de préordres partiels sur les interprétations.

Mots-clés

Changement de croyances, représentation des connaissances, logique propositionnelle, contraction, effacement

Abstract

Belief change is an important topic of knowledge representation and reasoning in artificial intelligence. Within the logical framework, the AGM approach has become a standard and various belief change operations have been considered. While revision, contraction and updating have given rise to a great deal of work, erasure has so far attracted less interest. Erasure is to contraction what update is to revision. This article deals with the study of erasure within the framework of propositional logic. It extends Katsuno and Mendelzon's approach with additional postulates capturing the minimality of change and proposes two representation theorems for erasure operators, one in terms of total preorders on interpretations, the other in terms of partial preorders on interpretations. Finally, it completes the work of Caridroit, Konieczny and Marquis for contraction by proposing a new representation theorem for contraction operators in terms of partial preorders on interpretations.

Keywords

Belief change, knowledge representation, propositional logic, contraction, erasure

1 Introduction

Le changement de croyances est une thématique importante dans le domaine de la représentation des connaissances et du raisonnement en intelligence artificielle. Les approches logiques ont donné lieux à de nombreux travaux depuis l'approche AGM qui s'est imposée comme un standard [1, 10]. Différentes opérations de changement de croyances ont été étudiées, révision [1, 10, 6, 20], mise à jour [9, 21, 15, 11], contraction [2, 10, 4] [1].

Afin de caractériser différentes approches sémantiques de la révision dans un même cadre, Katsuno et Mendelzon [14] ont restreint l'approche AGM à la logique propositionnelle. Ils ont reformulés les postulats AGM, et proposé deux théorèmes de représentation, l'un en termes de préordres totaux sur les interprétations et l'autre en termes de préordres partiels sur les interprétations. Ce cadre sémantique a permis de clairement distinguer les opérations de révision et de mise à jour [15]. D'un point de vue sémantique, les opérateurs de révision reposent sur une minimisation globale, ils sélectionnent les modèles de la nouvelle information les plus "proches" des modéles de l'ensemble de croyances intial, alors que les opérateurs de mise à jour reposent sur une minimisation locale, ils sélectionnent, pour chaque modèle de l'ensemble de croyances intial, les modèles de la nouvelle information les plus "proches" de celui-ci.

Plus récemment, s'inspirant de Katsuno et Mendelzon, Caridroit, Konieczny et Marquis ont reformulé les postulats AGM de la contraction dans le cadre propositionnel et ont proposé un théorème de représentation pour les opérateurs de contraction en termes de préordres totaux sur les interprétations [4].

Dans [15], Katsuno et Mendelzon définissent une nouvelle opération de changement de croyances, qu'ils nomment *effacement*, qui est à la contraction ce que la mise à jour est à la révision. Effacer une information repose sur une minimisation locale, c'est à dire que les opérateurs d'effacement sélectionnent les modèles de l'ensemble de croyances initiales ainsi que les modèles qui falsifient l'information

1. Pour plus de détails sur les différentes approches proposées pour ces opérations consulter par exemple [18] ou [19].

à effacer, les plus "proches" de chaque modèle de l'ensemble de croyances initiales. Katsuno et Mendelzon proposent une définition sémantique de l'effacement ainsi que des postulats basiques que cette opération doit satisfaire.

L'effacement a été étudié comme beaucoup d'opérations de changement de croyances dans le cadre de fragments propositionnels [5], cependant les études réalisées sur l'effacement restaient, jusqu'à présent, incomplètes car elles ne proposaient qu'un ensemble de postulats basiques pour cette opération insuffisant pour conduire à un théorème de représentation.

L'objet de cet article est de compléter l'étude de l'effacement en proposant :

— des postulats supplémentaires qui capturent le changement minimal de l'opération d'effacement,
— un théorème de représentation pour les opérateurs d'effacement en termes de préordres totaux,
— un théorème de représentation pour les opérateurs d'effacement en termes de préordres partiels.

Par ailleurs, l'article complète également les travaux de Caridroit, Konieczny et Marquis sur la contraction [4] en présentant un théorème de représentation pour les opérateurs de contraction en termes de préordres partiels qui, à ce jour, manquait.

Ainsi, ce travail met une touche finale au panorama de l'étude sémantique des opérateurs de changement de croyances pour la contraction et l'effacement.

L'article se décline comme suit. Après quelques préliminaires en Section 2 qui rappellent quelques notions utiles à la suite de l'article, la Section 3 présente un rapide état de l'art sur les opérations de changement de croyances, révision, contraction et mise à jour. La Section 4 présente la contribution de l'article sur l'effacement (postulats supplémentaires, théorèmes de représentation) ainsi que le théorème de représentation pour les opérateurs de contraction en termes de préordres partiels avant de conclure en Section 6.

2 Préliminaires

Nous considérons $\mathcal{L}$ le langage de la logique propositionnelle défini sur un ensemble fini de variables propositionnelles (ou atomes), noté $\mathcal{U}$, muni des connecteurs logiques usuels $\vee$, $\wedge$, $\neg$ et des constantes $\top$, $\bot$. Les lettres minuscules romaines a, b, c, ... désignent les atomes, les lettres grecques α, β, φ, ... désignent les formules bien formées et les lettres majuscules désignent les ensembles d'atomes ou les ensembles de formules bien formées.

Une *interprétation* sur $\mathcal{U}$ est une fonction $w\colon \mathcal{U} \to \{0, 1\}$ qui associe à chaque atome une valeur de vérité, 0 pour faux ou 1 pour vrai. ($w(\bot) = 0$ et $w(\top) = 1$). Une telle interprétation est représentée par un ensemble $w \subseteq \mathcal{U}$ d'atomes évalués à vrai ou par son vecteur caractéristique correspondant de longueur $|\mathcal{U}|$. Une interprétation qui satisfait une formule φ est appelée *modèle* de φ et nous notons $\mathrm{Mod}(\varphi)$ l'ensemble des modèles de φ. De plus, $\varphi \models \psi$ si $\mathrm{Mod}(\varphi) \subseteq \mathrm{Mod}(\psi)$ et $\psi \equiv \varphi$ (ψ et φ sont équivalentes) si $\mathrm{Mod}(\psi) = \mathrm{Mod}(\varphi)$.

Une formule ψ est dite *complète* si pour toute formule $\mu \in \mathcal{L}$ soit $\psi \models \mu$ soit $\psi \models \neg\mu$. D'une manière équivalente, une formule satisfaisable est complète si et seulement si elle admet exactement un modèle.

Soit $\mathcal{I}$ un ensemble, un préordre est une relation binaire de $\mathcal{I} \times \mathcal{I}$, noté $\leq$, qui est réflexive et transitive. Le préordre strict associé à $\leq$ est défini par $w < w'$ si $w \leq w'$ et $w' \not\leq w$. Une relation d'équivalence induite par $\leq$, notée $\approx$, est définie par $w \approx w'$ si $w \leq w'$ et $w' \leq w$. Un préordre est dit total si pour tout w, w' de $\mathcal{I}$, soit $w \leq w'$ ou $w' \leq w$. L'ensemble des éléments minimaux de $\mathcal{I}$ selon le préordre $\leq$, noté $\min_{\leq}(\mathcal{I})$ est défini par $\min_{\leq}(\mathcal{I}) = \{w \in \mathcal{I} \mid$ il n'existe pas $w' \in \mathcal{I}$ tel que $w' < w\}$.

Nous rappelons deux pseudo-distances entre un modèle w et une formule. L'une basée sur la cardinalité, notée $|\Delta|_w^{min}(\mu)$, comme étant le nombre minimal de variables propositionnelles pour lesquelles w et les modèles de μ diffèrent. Formellement, $|\Delta|_w^{min}(\mu) = min\{|w\Delta w'| : w' \in \mathrm{Mod}(\mu)\}$. L'autre basée sur l'inclusion ensembliste, notée $\Delta_w^{min}(\mu)$, comme étant les ensembles minimaux, au sens de l'inclusion, de variables propositionnelles pour lesquelles w et les modèles de μ diffèrent. Formellement, $\Delta_w^{min}(\mu) = min_{\subseteq}(\{w\Delta w' : w' \in \mathrm{Mod}(\mu)\})$.

3 Etat de l'art

3.1 Révision des croyances

La révision de croyances est une opération qui permet d'accepter une nouvelle information, en préservant la cohérence tout en modifiant le moins possible les croyances initiales. Plus formellement, une opération de révision, est une fonction, notée $\circ$, de $\mathcal{L} \times \mathcal{L}$ vers $\mathcal{L}$, qui à partir d'une formule ψ (les croyances initiales d'un agent), et d'une formule μ (la nouvelle information), fournit une nouvelle formule, $\psi \circ \mu$ (les croyances de l'agent révisées).

Alchourròn, Gärdenfors et Makinson [1] ont étudié la révision lorsque les croyances d'un agent sont représentées par une théorie (ensemble de formules déductivement clos) et ont proposé un ensemble de postulats, appelés *postulats AGM*, que devrait satisfaire tout opérateur de révision "rationnel". Dans le cadre propositionnel, Katsuno et Mendelzon [13] ont reformulé les postulats AGM lorsqu'une théorie est représentée par les modèles d'une formule. Dans ce cadre, la révision de ψ par μ revient à rechercher les modèles de μ les plus proches de ceux de ψ. Ces postulats connus sous le nom de postulats KM sont les suivants :

Soit $\psi, \psi_1, \psi_2, \mu, \mu_1, \mu_2 \in \mathcal{L}$.

(R1) $\psi \circ \mu \models \mu$.
(R2) Si $(\psi \wedge \mu)$ est satisfaisable, alors $\psi \circ \mu \equiv \psi \wedge \mu$.
(R3) Si μ est satisfaisable, alors $\psi \circ \mu$ est satisfaisable.
(R4) Si $\psi_1 \equiv \psi_2$ et $\mu_1 \equiv \mu_2$, alors $\psi_1 \circ \mu_1 \equiv \psi_2 \circ \mu_2$.
(R5) $(\psi \circ \mu_1) \wedge \mu_2 \models \psi \circ (\mu_1 \wedge \mu_2)$.
(R6) Si $(\psi \circ \mu_1) \wedge \mu_2$ est satisfaisable alors $\psi \circ (\mu_1 \wedge \mu_2) \models (\psi \circ \mu_1) \wedge \mu_2$.
(R7) Si $(\psi \circ \mu_1) \models \mu_2$ et $(\psi \circ \mu_2) \models \mu_1$, alors $\psi \circ \mu_1 \equiv \psi \circ \mu_2$.
(R8) $(\psi \circ \mu_1) \wedge (\psi \circ \mu_2) \models \psi \circ (\mu_1 \vee \mu_2)$.

Une description détaillée de ces postulats est dans [14].

Katsuno et Mendelzon [13, 14] ont montré que l'opération de révision $\circ$, selon l'ensemble de postulats KM qu'elle satisfait, peut se traduire par un préordre total sur les interprétations ou par un préordre partiel sur les interprétations. Plus formellement, une *affectation fidèle* est une fonction qui associe à chaque formule ψ un préordre $\leq_\psi$ sur les interprétations tel que :

1. Si $w, w' \in \mathrm{Mod}(\psi)$ alors $w \approx_\psi w'$
2. Si $w \in \mathrm{Mod}(\psi)$ et $w' \notin \mathrm{Mod}(\psi)$ alors $w <_\psi w'$
3. Si $\psi \equiv \psi'$, alors $\leq_\psi = \leq_{\psi'}$

Ils ont fourni le théorème de représentation suivant :

Théorème 1. [14]
— *Un opérateur de révision $\circ$ satisfait tous les postulats* (R1)–(R6) *si et seulement si il existe une affectation fidèle qui associe à chaque formule ψ un préordre total $\leq_\psi$ tel que* $\mathrm{Mod}(\psi \circ \mu) = \min_{\leq_\psi}(\mathrm{Mod}(\mu))$.
— *Un opérateur de révision $\circ$ satisfait les postulats $(R1)$–$(R5)$, $(R7)$ and $(R8)$ si et seulement si il existe une affectation fidèle qui associe à chaque formule ψ un préordre partiel $\leq_\psi$ tel que* $\mathrm{Mod}(\psi \circ \mu) = \min_{\leq_\psi}(\mathrm{Mod}(\mu))$.

Il existe de nombreux opérateurs de révision dans la littérature, nous nous limitons ici à présenter l'opérateur de révision de Dalal, noté ($\circ_D$), [6], et l'opérateur de révision de Satoh, noté ($\circ_S$), [20]. La proximité pour l'opérateur Dalal se mesure en terme de cardinalité de la différence symétrique entre modèles,

$$\mathrm{Mod}(\psi \circ_D \mu) = \{m \in \mathrm{Mod}(\mu) : \exists m' \in \mathrm{Mod}(\psi) \text{ tel que } |m \Delta m'| = |\Delta|_m^{min}(\mu)\}.$$

Tandis que la proximité pour l'opérateur Satoh se mesure en terme d'inclusion ensembliste de la différence symétrique entre modèles.

$$\mathrm{Mod}(\psi \circ_S \mu) = \{m \in \mathrm{Mod}(\mu) : \exists m' \in \mathrm{Mod}(\psi) \text{ tel que } m \Delta m' \in \Delta_m^{min}(\mu)\}.$$

Notons que l'opérateur de révision de Dalal satisfait (R1)-(R6) [8, 14] tandis que l'opérateur de révision de Satoh satisfait (R1)-(R5), (R7) et (R8) [14].

3.2 Contraction des croyances

La contraction de croyances est une opération qui permet de retirer une croyance de l'ensemble des croyances intitiales en respectant le principe de changement minimal tout en préservant la cohérence. Formellelement, une opération de contraction est une fonction, notée $-$, de $\mathcal{L} \times \mathcal{L}$ vers $\mathcal{L}$ qui à partir d'une formule ψ (les croyances initiales d'un agent) et d'une formule μ (la croyance à retirer), fournit $\psi - \mu$ (les croyances de l'agent contractées).

Alchourrón, Gärdenfors et Makinson [1] ont étudié la contraction lorsque les croyances d'un agent sont représentées par une théorie et ont proposé des postulats que tout opérateur de contraction devrait satisfaire. Katsuno et Mendelzon [15] ont proposé une reformulation de certains postulats AGM pour la contraction lorsqu'une théorie est représentée par les modèles d'une formule. Plus récemment, Caridroit, Konieczny et Marquis [4] ont revisité les postulats de la contraction, en proposant, en particulier, des postulats supplémentaires caractérisant le principe de changement minimal.
Soit $\psi, \psi_1, \psi_2, \mu, \mu_1, \mu_2 \in \mathcal{L}$.

(C1) $\psi \models \psi - \mu$.
(C2) Si $\psi \not\models \mu$, alors $\psi - \mu \models \psi$.
(C3) Si $\psi - \mu \models \mu$, alors $\models \mu$.
(C4) Si $\psi_1 \equiv \psi_2$ et $\mu_1 \equiv \mu_2$, alors $\psi_1 - \mu_1 \equiv \psi_2 - \mu_2$.
(C5) Si $\psi \models \mu$, alors $(\psi - \mu) \wedge \mu \models \psi$.
(C6) $\psi - (\mu_1 \wedge \mu_2) \models (\psi - \mu_1) \vee (\psi - \mu_2)$.
(C7) Si $\psi - (\mu_1 \wedge \mu_2) \not\models \mu_1$, alors $\psi - \mu_1 \models \psi - (\mu_1 \wedge \mu_2)$.

Une description détaillée de ces postulats est dans [15] et [4]. [2]

Caridroit, Konieczny et Marquis [4] ont montré qu'une opération de contraction $-$ satisfaisant l'ensemble des postulats qu'ils ont proposés peut se traduire par un préordre total sur les interprétations.

Théorème 2. [3] *Un opérateur de contraction $-$ satisfait les postulats* (C1)–(C7) *si et seulement si il existe une affectation fidèle qui associe à chaque formule ψ un préordre total $\leq_\psi$ tel que* $\mathrm{Mod}(\psi - \mu) = \mathrm{Mod}(\psi) \cup \min_{\leq_\psi}(\mathrm{Mod}(\neg\mu))$.

Les opérations de révision et de contraction sont extrêmement liées comme le reflètent les identités de Levi et Harper :

$$\psi - \mu \equiv \psi \vee (\psi \circ \neg\mu) \text{ (identité de Harper)}$$

$$\psi \circ \mu \equiv (\psi - \neg\mu) \wedge \mu \text{ (identité de Lévi)}$$

Ces identités ont permis à Caridroit, Konieczny et Marquis de montrer les correspondances entre les postulats de la révision et ceux de la contraction [4].

2. Dans un souci de cohérence nous avons adopté la numérotation des postulats initialement proposée par Katsuno et Mendelzon, et qui diffère de celle proposée par Caridroit, Konieczny et Marquis, les postulats (C4) et (C5) étant inversés.

Théorème 3. [4]

- *Si l'opérateur de contraction $-$ satisfait* (C1)–(C5) *alors son analogue pour la révision $\circ$ défini par l'identité de Lévi satisfait* (R1)–(R4). *De plus si* (C6) *est satisfait, alors* (R5) *est satisfait pour la révision. Enfin, si le postulat* (C7) *est satisfait, alors le postulat* (R6) *l'est aussi.*
- *Si l'opérateur de révision $\circ$ satisfait* (R1)–(R4) *alors son analogue pour la contraction $-$ défini par l'identité de Harper satisfait* (C1)–(C5). *De plus si* (R5) *est satisfait, alors* (C6) *est satisfait pour la contraction. Enfin, si* (R6) *est satisfait, alors* (C7) *l'est aussi.*

Cela conduit naturellement à définir des opérateurs de contraction à partir d'opérateurs de révision. L'opérateur de contraction $-_D$, analogue à l'opérateur de révision de Dalal $\circ_D$, et l'opérateur de contraction $-_S$, analogue à l'opérateur de révision de Satoh $\circ_S$, sont respectivement définis comme suit :

$$\mathrm{Mod}(\psi -_D \mu) = \mathrm{Mod}(\psi) \cup \mathrm{Mod}(\psi \circ_D \neg\mu).$$

$$\mathrm{Mod}(\psi -_S \mu) = \mathrm{Mod}(\psi) \cup \mathrm{Mod}(\psi \circ_S \neg\mu).$$

Notons que l'opérateur $-_D$ satisfait (C1)–(C7) tandis que l'opérateur $-_S$ satisfait (C1)–(C6), mais viole (C7) [4].

3.3 Mise à jour des croyances

Keller et Winslett [16], puis Katsuno et Mendelzon [15] ont mis en évidence les differences entre révision et mise à jour. La mise à jour de croyances est une opération qui incorpore une nouvelle information reflétant un changement de l'environnement, en préservant la cohérence tout en modifiant le moins possible les croyances initiales. Formellement, une opération de la mise à jour est une fonction, notée $\diamond$, de $\mathcal{L} \times \mathcal{L}$ vers $\mathcal{L}$ qui à partir d'une formule ψ (les croyances initiales d'un agent) et d'une formule μ (la nouvelle information reflétant le changement de l'environnement), fournit $\psi \diamond \mu$ (les croyances de l'agent mises à jour). Les postulats KM [15] de la mise à jour sont les suivants :
Soit $\psi, \psi_1, \psi_2, \mu, \mu_1, \mu_2 \in \mathcal{L}$.

(U1) $\psi \diamond \mu \models \mu$.
(U2) Si $\psi \models \mu$, alors $\psi \diamond \mu \equiv \psi$.
(U3) Si ψ et μ sont satisfaisables, alors $\psi \diamond \mu$ l'est aussi.
(U4) Si $\psi_1 \equiv \psi_2$ et $\mu_1 \equiv \mu_2$, alors $\psi_1 \diamond \mu_1 \equiv \psi_2 \diamond \mu_2$.
(U5) $(\psi \diamond \mu) \wedge \phi \models \psi \diamond (\mu \wedge \phi)$.
(U6) Si $(\psi \diamond \mu_1) \models \mu_2$ et $(\psi \diamond \mu_2) \models \mu_1$, alors $\psi \diamond \mu_1 \equiv \psi \diamond \mu_2$.
(U7) Si ψ est complète, alors $(\psi \diamond \mu_1) \wedge (\psi \diamond \mu_2) \models \psi \diamond (\mu_1 \vee \mu_2)$.
(U8) $(\psi_1 \vee \psi_2) \diamond \mu \equiv (\psi_1 \diamond \mu) \vee (\psi_2 \diamond \mu)$.
(U9) Si ψ est complète et $(\psi \diamond \mu) \wedge \phi$ est satisfaisable, alors $\psi \diamond (\mu \wedge \phi) \models (\psi \diamond \mu) \wedge \phi$.

Une description détaillée de ces postulats est dans [15]. Les quatre postulats (U1), (U4), (U5) et (U6) correspondent respectivement aux postulats de révision (R1), (R4), (R5) et (R7). Le postulat (U8) est spécifique à la mise à jour et exprime le fait qu'un opérateur de mise à jour doit accorder la même attention à chacun des modèles des croyances initiales. Les postulats (U7) et (U9) correspondent respectivement à (R8) et (R9), mais sont limités aux formules complètes (ce qui est logique en présence de (U8)). Les postulats (U2) et (U3) diffèrent de (R2) et (R3), ce sont des versions plus faibles des postulats de révision. Une conséquence pour la mise à jour est qu'une fois qu'une incohérence est introduite dans les croyances initiales, il n'y a aucun moyen de l'éliminer. Noter que ce n'est pas le cas pour la révision.

Katsuno et Mendelzon [15] ont montré que l'opération de mise à jour $\diamond$, selon l'ensemble de postulats KM qu'elle satisfait, peut se traduire par un préordre total sur les interprétations ou par un préordre partiel sur les interprétations [15]. Plus formellement, une *affectation fidèle ponctuelle* est une fonction qui associe à chaque interprétation w un préordre $\leq_w$ sur les interprétations, tel que pour toute interpretation w', si $w' \neq w$ alors $w <_w w'$. Ils ont fourni le théorème de représentation suivant :

Théorème 4. [15]

- *Un opérateur de mise à jour $\diamond$ satisfait les postulats* (U1)–(U5) *et* (U8)–(U9) *si et seulement si il existe une affectation fidèle ponctuelle qui associe à chaque interpretation w un préordre total $\leq_w$ tel que* $\mathrm{Mod}(\psi \diamond \mu) = \bigcap_{w \in \mathrm{Mod}(\psi)} min(\mathrm{Mod}(\mu), \leq_w)$.
- *Un opérateur de mise à jour $\diamond$ satisfait les postulats* (U1)–(U8) *si et seulement si il existe une affectation fidèle ponctuelle qui associe à chaque interpretation w un préordre partiel $\leq_w$ tel que* $\mathrm{Mod}(\psi \diamond \mu) = \bigcap_{w \in \mathrm{Mod}(\psi)} min(\mathrm{Mod}(\mu), \leq_w)$.

Ce théorème de représentation permet de mettre en évidence la différence entre révision et mise-à-jour. La mise à jour repose sur une minimisation ponctuelle, modèle par modèle de ψ alors que la révision repose sur une minimisation globale. Il est à noter que lorsque ψ est une formule complète révision et mise à jour coïncident.
Nous nous limitons à présenteur deux opérateurs de mise à jour, l'opérateur de Forbus [9], noté $\diamond_F$, et l'opérateur de de Winslett est aussi appelé *PMA (Possible Models Approach)*[21], noté $\diamond_W$, définis respectivement comme suit :

$$\mathrm{Mod}(\psi \diamond_F \mu) = \bigcap_{w \in \mathrm{Mod}(\psi)} \{w' \in \mathrm{Mod}(\mu) : |w \Delta w'| = |\Delta|_w^{min}(\mu)\}.$$

$$\mathrm{Mod}(\psi \diamond_W \mu) = \bigcap_{m \in \mathrm{Mod}(\psi)} \{m' \in \mathrm{Mod}(\mu) : m \Delta m' \in \Delta_m^{min}(\mu)\}.$$

Notons que l'opérateur $\diamond_F$ satisfait (U1)–(U9) [15, 11] tandis que l'opérateur $\diamond_W$ satisfait (U1)–(U8) [15], mais viole (U9)[17].

4 Effacement des croyances

L'effacement, introduit par Katsuno and Mendelzon [15], est à la contraction ce que la mise à jour est à la révision.

Intuitivement, effacer une croyance signifie que l'environnement de l'agent a changé de telle sorte que cette croyance peut ne plus être valide. D'un point de vue logique lorsque les croyances d'un agent sont représentées par une formule ψ, effacer une croyance μ de ψ signifie ajouter localement les modèles de $\neg\mu$ aux modèles de ψ. Les opérateurs d'effacement que nous étudions sont des fonctions, notées $\triangleleft$ de $\mathcal{L} \times \mathcal{L}$ vers $\mathcal{L}$ qui à partir d'une formule ψ qui représente les croyances initiales d'un agent et d'une formule μ à effacer, renvoie une formule $\psi \triangleleft \mu$.

Exemple 1. *Considérons l'exemple inspiré de celui utilisé dans [15] où les croyances décrivent deux objets A et B qui se trouvent dans une pièce. Il y a une table dans la pièce et les objets peuvent être sur la table ou pas. La formule a signifie que* "l'objet A est sur la table" *et la formule b signifie que* "l'objet B est sur la table". *Supposons que les croyances d'un agent sont représentées par la formule $\psi = (a \wedge \neg b) \vee (\neg a \wedge b)$, qui exprime que soit l'objet A, soit l'objet B est sur la table mais pas les deux. La contraction de ψ par b, c'est à dire retirer l'information que b est sur la table, ne change pas les croyances de l'agent et $\psi - b \equiv \psi$. En revanche, supposons que l'environnement de l'agent a changé, un robot a été envoyé avec pour instruction de faire en sorte que B ne soit pas sur la table, ce changement de l'environnement se traduit par l'effacement de ψ par b et dans ce cas $\psi \triangleleft b \equiv (\neg a \wedge b) \vee \neg b$.*
Plus formellement, $\mathrm{Mod}(\psi) = \{a, b\}$, $\mathrm{Mod}(b) = \{b, ab\}$ et $\mathrm{Mod}(\neg b) = \{a, \emptyset\}$, les modèles des croyances de l'agent après la contraction par b sont $\mathrm{Mod}(\psi - b) = \{a, b\}$ tandis que les modèles des croyances de l'agent après l'effacement de b sont $\mathrm{Mod}(\psi \triangleleft b) = \{a, b, \emptyset\}$.
La différence intuitive entre la contraction et l'effacement peut s'expliquer comme suit. La contraction par b signifie que rien n'a changé dans la pièce, comme l'agent croit que l'objet A ou l'objet B est sur la table mais pas les deux, la contraction n'a aucun effet sur les croyances de l'agent. En revanche, l'effacement par b signifie que l'état de la pièce a changé, si l'objet B était sur la table avant le changement, il a été déplacé mais on ne peut rien déduire sur la position de l'objet A du fait que l'objet B n'est pas sur la table.

4.1 Postulats de base pour l'effacement

Des postulats que tout opérateur d'effacement devrait satisfaire ont été proposés par Katsuno et Mendelzon [15] dans le même esprit que ceux proposés pour la contraction et la mise à jour.
Soit $\psi, \psi_1, \psi_2, \mu, \mu_1, \mu_2 \in \mathcal{L}$.

- (E1) $\psi \models \psi \triangleleft \mu$.
- (E2) Si $\psi \not\models \mu$, alors $\psi \triangleleft \mu \equiv \psi$.
- (E3) Si ψ est satisfaisable et $\not\models \mu$, alors $\psi \triangleleft \mu \not\models \mu$.
- (E4) Si $\psi_1 \equiv \psi_2$ et $\mu_1 \equiv \mu_2$, alors $\psi_1 \triangleleft \mu_1 \equiv \psi_2 \triangleleft \mu_2$.
- (E5) $(\psi \triangleleft \mu) \wedge \mu \models \psi$.
- (E8) $(\psi_1 \vee \psi_2) \triangleleft \mu \equiv (\psi_1 \triangleleft \mu) \vee (\psi_2 \triangleleft \mu)$.

De même que pour la révision et la mise à jour, le postulat (E8) est spécifique à l'effacement, tandis que les autres postulats pour la contraction et l'effacement ne diffèrent que sur les deuxième et troisième postulats. Les postulats (E2) et (E3) sont plus faibles que les postulats (C2) et (C3). Ceci est illustré par l'exemple donné ci-dessus. Comme pour la mise à jour, une conséquence pour l'effacement est qu'une fois qu'une incohérence est introduite dans les croyances initiales, il n'y a aucun moyen de l'éliminer.

De façon similaire aux identités de Levi et Harper liant contraction et révision Katsuno and Mendelzon [15] ont proposé des identités reliant effacement et mise à jour :

(Id_1) $\psi \triangleleft \mu \equiv \psi \vee (\psi \diamond \neg\mu)$
(Id_2) $\psi \diamond \mu \equiv (\psi \triangleleft \neg\mu) \wedge \mu$

De plus, ils ont proposé le résultat suivant concernant la satisfaction des postulats.

Proposition 1. [15]

1. *Si un opérateur de mise à jour $\diamond$ satisfait (*U1)-(U4) et (U8), alors l'opérateur d'effacement $\triangleleft$ defini par l'identité (Id_1) satisfait (E1)-(E5) et (E8).
2. Si un opérateur d'effacement $\triangleleft$ satisfait (E1)-(E4) et (E8), alors l'opérateur de mise à jour $\diamond$ defini par l'identité (Id_2) satisfait (U1)-(U4) et (U8).
3. Si un opérateur de mise à jour $\diamond$ satisfait (U1)-(U4) et (U8), alors il est possible de définir un opérateur d'effacement grâce à (Id_1). L'opérateur de mise à jour obtenu à partir de cet opérateur d'effacement via (Id_2) est égal à l'opérateur de mise à jour initial $\diamond$.
4. Si un opérateur d'effacement $\triangleleft$ satisfait (E1)-(E5) et (E8), alors il est possible de définir un opérateur de mise à jour grâce à (Id_2). L'opérateur d'effacement obtenu à partir de cet opérateur de mise à jour via (Id_1) est égal à l'opérateur d'effacement initial $\triangleleft$.

L'identité (Id_1) nous a permis de définir des opérateurs d'effacement à partir des opérateurs de mise à jour connus [5]. Un opérateur d'effacement, noté $\triangleleft_F$, est défini par

$$\mathrm{Mod}(\psi \triangleleft_F \mu) = \mathrm{Mod}(\psi) \cup \mathrm{Mod}(\psi \diamond_F \neg\mu)$$

où $\diamond_F$ est l'opérateur de mise à jour de Forbus. Un opérateur d'effacement, noté $\triangleleft_W$, est défini par

$$\mathrm{Mod}(\psi \triangleleft_W \mu) = \mathrm{Mod}(\psi) \cup \mathrm{Mod}(\psi \diamond_W \neg\mu)$$

où $\diamond_W$ est l'opérateur de mise à jour de Winslett. Selon la Proposition 1, les opérateurs d'effacement $\triangleleft_F$ et $\triangleleft_W$ satisfont $(\mathrm{E1}) - (\mathrm{E5})$ et $(\mathrm{E8})$.
L'exemple suivant illustre les opérateurs $\triangleleft_F$ et $\triangleleft_W$.

Exemple 2. *Soit ψ, μ deux formules propositionnelles telles que $Mod(\psi) = \{abcd, a\}$ et $Mod(\mu) = \{a, c, d, ab, ac, ad, bc, cd, abc, abd, bcd, abcd, \emptyset\}$.*
Nous avons $Mod(\neg\mu) = \{acd, bd, b\}$ et d'après la Table 2, $Mod(\psi \diamond_F \neg\mu) = \{acd, b\}$ et $Mod(\psi \diamond_W \neg\mu) = \{acd, bd, b\}$. Les résultats de l'effacement avec les opérateurs $\triangleleft_F$ et $\triangleleft_W$ sont respectivement $Mod(\psi \triangleleft_F \mu) = \{abcd, a, acd, b\}$ et $Mod(\psi \triangleleft_W \mu) = \{abcd, a, acd, bd, b\}$.

	$Mod(\psi)$		
	Δ	$abcd$	a
$Mod(\neg\mu)$	acd	$\boldsymbol{b^*}$	$\boldsymbol{cd^*}$
	bd	ac^*	abd
	b	acd	$\boldsymbol{ab^*}$

TABLE 1 – Différences symétriques ; par colonne, les minimaux selon la cardinalité sont notés en gras et les minimaux selon l'inclusion sont notés avec une asterisque.

4.2 De nouveaux postulats capturant la minimalité du changement

Nous pouvons ajouter deux postulats supplémentaires pour capturer la minimalité du changement. Ils sont équivalents à (C6) et (C7), à la seule différence qu'en raison de la règle de disjonction (E8), le postulat (E7) peut être restreint aux formules complètes.

(E6) $\psi \lhd (\mu_1 \wedge \mu_2) \models (\psi \lhd \mu_1) \vee (\psi \lhd \mu_2)$.
(E7) Si ψ est complète et $\psi \lhd (\mu_1 \wedge \mu_2) \not\models \mu_1$, alors $\psi \lhd \mu_1 \models \psi \lhd (\mu_1 \wedge \mu_2)$.

On peut observer que l'opérateur d'effacement de Forbus $\lhd_F$ satisfait (E6) et (E7), tandis que l'opérateur de Winslett $\lhd_W$ satisfait (E6) mais pas (E7).

Il n'est guère surprenant que nous puissions énoncer un théorème de représentation pour les opérateurs d'effacement définis par des préordres totaux qui est la contrepartie du théorème de représentation obtenu par Caridroit *et al.* [4] pour la contraction. La preuve nécessite les deux lemmes suivants.
Le premier lemme explicite le résultat de l'effacement par une formule qui n'a qu'un seul contre-modèle et n'utilise que les postulats de base.

Lemma 1. *Soit $\lhd$ un opérateur d'effacement qui satisfait les postulats* (E1)–(E5) *et* (E8)*, ψ une formule satisfaisable, w une interpretation et α_w une formule ayant w comme unique modèle, alors $\psi \lhd \neg\alpha_w \equiv \psi \vee \alpha_w$.*

Démonstration. Puisque l'opérateur satisfait (E8) il est suffisant de prouver le lemme quand ψ est une formule complète ayant, disons w_0 comme unique modèle.
Si $w_0 = w$, alors $\psi \models \alpha_w$ et selon (E2), $\psi \lhd \neg\alpha_w \equiv \psi \equiv \psi \vee \alpha_w$.
Si $w_0 \neq w$, d'après (E5) $(\psi \lhd \neg\alpha_w) \wedge \neg\alpha_w \models \psi$, et donc $(\psi \lhd \neg\alpha_w) \models \psi \vee \alpha_w$. Par (E1), $\psi \models (\psi \lhd \neg\alpha_w)$. De plus, selon (E3), $(\psi \lhd \neg\alpha_w) \not\models \neg\alpha_w$. Ainsi nous obtenons $\alpha_w \models (\psi \lhd \neg\alpha_w)$ et donc $\psi \lhd \neg\alpha_w \equiv \psi \vee \alpha_w$. □

Le deuxième lemme met en jeu les deux postulats additionnels qui traitent de la minimalité du changement.

Lemma 2. *Soit $\lhd$ un opérateur d'effacement qui satisfait les postulats* (E1)–(E8)*, ψ une formule complète, et α et β deux formules qui ne sont pas des tautologies, alors $\psi \lhd (\alpha \wedge \beta) \equiv \psi \lhd \alpha$ ou $\psi \lhd \beta$ ou $(\psi \lhd \alpha) \vee (\psi \lhd \beta)$.*

Démonstration. Sous les hypothèses faites sur les formules la preuve est similaire à celle donnée dans [4, preuve de la Proposition 9]. □

Dans tout ce qui suit, étant donné une interprétation w, ψ_w (resp. α_w) représente une formule complète dont l'unique modèle est w. Egalement, étant donné une interprétation w_i, on note α_i une formule complète dont l'unique modèle est w_i.

4.3 L'effacement en termes de préordres totaux

Nous sommes maintenant en mesure de prouver le théorème de représentation suivant, qui montre que les postulats capturent tous les opérateurs d'effacement définis par un préordre total.

Théorème 5. *Un opérateur d'effacement $\lhd$ satisfait les postulats* (E1)–(E8) *si et seulement si il existe une affectation fidèle ponctuelle qui associe à chaque interprétation w un préordre total $\leq_w$ tel que* $\mathrm{Mod}(\psi \lhd \mu) = \mathrm{Mod}(\psi) \cup \bigcup_{w \in \mathrm{Mod}(\psi)} min_{\leq_w}(\mathrm{Mod}(\neg\mu))$.

Démonstration. $\Leftarrow$) Suposons que nous avons une affectation fidèle ponctuelle qui associe à chaque interprétation w un préordre total $\leq_w$.
Considérons l'opérateur d'effacement $\lhd$ défini par

$$\mathrm{Mod}(\psi \lhd \mu) = \mathrm{Mod}(\psi) \cup \bigcup_{w \in \mathrm{Mod}(\psi)} min_{\leq_w}(\mathrm{Mod}(\neg\mu)).$$

Prouvons dans un premier temps que $\lhd$ satisfait les postulats (E1)–(E8). Il est évident que $\lhd$ satisfait (E1), (E3), (E4), (E5) et (E8). Si ψ est insatisfaisable alors (E2), (E6) et (E7) sont trivialement vérifiés. Nous supposons donc dans la suite que ψ est satisfaisable.

Il découle de la définition d'une affectation fidèle ponctuelle que si w est un modèle de $\neg\mu$, alors $\psi_w \lhd \mu$ est équivalent à ψ_w. On obtient donc (E2) en utilisant (E8).

Puisque $min_{\leq_w}\mathrm{Mod}(\neg\mu_1 \vee \neg\mu_2) \subseteq min_{\leq_w}\mathrm{Mod}(\neg\mu_1) \cup min_{\leq_w}\mathrm{Mod}(\neg\mu_2)$, (E6) est vérifié.

Soit ψ une formule complète telle que $\mathrm{Mod}(\psi) = \{w_0\}$. Supposons que $\psi \lhd (\mu_1 \wedge \mu_2) \not\models \mu_1$. Cela signifie qu'il existe $w \in \mathrm{Mod}(\psi) \cup min_{\leq_{w_0}}(\mathrm{Mod}(\neg(\mu_1 \wedge \mu_2)))$ tel que $w \in \mathrm{Mod}(\neg\mu_1)$. Si $w \in \mathrm{Mod}(\psi \wedge \neg\mu_1)$, alors $\mathrm{Mod}(\psi) = \{w\}$ et dans ce cas puisque l'affectation est fidèle ponctuelle $\mathrm{Mod}(\psi \lhd \mu_1) = \mathrm{Mod}(\psi)$. Le fait que (E7) est valide découle alors de (E1). Si $w \notin \mathrm{Mod}(\psi)$, pour vérifier que (E7) est valide il suffit de montrer que $min_{\leq_{w_0}}(\mathrm{Mod}(\neg\mu_1)) \subseteq min_{\leq_{w_0}}(\mathrm{Mod}(\neg(\mu_1 \wedge \mu_2))$. Soit $w' \in min_{\leq_{w_0}}(\mathrm{Mod}(\neg\mu_1))$. Puisque $\leq_{w_0}$ est un préordre total et que $w \in \mathrm{Mod}(\neg\mu_1)$, nous avons $w' \leq_{w_0} w$. Supposons qu'il existe $w" \in \mathrm{Mod}(\neg(\mu_1 \wedge \mu_2))$ tel que $w" <_{w_0} w'$, alors on a également $w" <_{w_0} w$, ce qui contredit le fait que $w \in min_{\leq_{w_0}}(\mathrm{Mod}(\neg(\mu_1 \wedge \mu_2)))$. Donc (E7) est bien vérifié.

$\Rightarrow$) Soit $\lhd$ un opérateur d'effacement qui satisfait les postulats(E1)–(E8). Définissons une relation binaire $\leq_w$ sur les interprétations comme suit :

$$w_1 \leq_w w_2 \text{ si ou bien } w_1 = w \text{ ou } w_1 \in \mathrm{Mod}(\psi_w \lhd \neg(\alpha_1 \vee \alpha_2)).$$

Montrons tout d'abord que $\leq_w$ est un préordre total. Il découle du postulat (E3) que ou bien w_1 ou w_2 appartient à $\text{Mod}(\psi_w \lhd \neg(\alpha_1 \vee \alpha_2))$, prouvant ainsi que la relation binaire est totale. Le fait que $\leq_w$ est reflexive découle du Lemme 1.
La preuve que $\leq_w$ est transitive est complètement similaire à la preuve donnée par Caridroit *et al.* dans le cas de la contraction (voir [4, Preuve du Theorem 14]). Cette dernière s'appuie sur un lemme analogue au Lemme 2 pour la contraction et utilise alors seulement les postulats (C1), (C6) et (C7), qui sont analogues à (E1), (E6) et (E7) (quand le dernier est restreint à une formule complète).

Il découle de (E2) que l'application $w \mapsto \leq_w$ est une affectation fidèle ponctuelle.

Il reste à montrer que

$$\text{Mod}(\psi \lhd \mu) = \text{Mod}(\psi) \cup \bigcup_{w \in \text{Mod}(\psi)} min_{\leq_w}(\text{Mod}(\neg\mu)).$$

Si ψ est insatisfaisable, alors les deux côtés de l'équation sont vides et l'égalité est vérifiée. Si ψ est satisfaisable, alors selon (E8) étant donné une interprétation w il est suffisant de prouver que $\text{Mod}(\psi_w \lhd \mu) = \{w\} \cup min_{\leq_w}(\text{Mod}(\neg\mu))$. Si $w \in \text{Mod}(\neg\mu)$ alors il découle de (E2) que $\psi_w \lhd \mu \equiv \psi$ et l'égalité est valide puisque nous utilisons une affectation fidèle. Nous supposons donc dans la suite que $w \notin \text{Mod}(\neg\mu)$. Si μ est une tautologie, selon (E1) et (E5) $\psi_w \lhd \mu \equiv \psi$ et l'égalité est valide. Supposons maintenant qu'il existe $w_1 \in \text{Mod}(\psi_w \lhd \mu)$ qui est dans $\text{Mod}(\neg\mu)$ mais qui n'appartient pas à $min_{\leq_w}(\text{Mod}(\neg\mu))$. Alors il existe $w_2 \in \text{Mod}(\neg\mu)$ tel que $w_2 <_w w_1$. Nous avons alors $w_1 \notin \text{Mod}(\psi_w \lhd \neg(\alpha_1 \vee \alpha_2))$. Considérons maintenant la formule $\beta = \neg\mu \wedge \neg\alpha_1 \wedge \neg\alpha_2$. Clairement nous avons $\neg\mu \equiv \beta \vee (\alpha_1 \vee \alpha_2)$. Puisque (E4) est satisfait, $\psi_w \lhd \mu \equiv \psi_w \lhd (\neg\beta \wedge \neg\alpha_1 \wedge \neg\alpha_2)$. Et donc selon (E6), $\psi_w \lhd \mu \models (\psi_w \lhd \neg\beta) \vee (\psi_w \lhd \neg(\alpha_1 \vee \alpha_2))$. Nous avons supposé que $w_1 \in \text{Mod}(\psi_w \lhd \mu)$ et $w_1 \notin \text{Mod}(\psi_w \lhd \neg(\alpha_1 \vee \alpha_2))$, donc $w_1 \in \text{Mod}(\psi_w \lhd \neg\beta)$. Selon (E5) nous avons $(\psi_w \lhd \neg\beta) \wedge \neg\beta \models \psi_w$. Puisque nous avons également $w_1 \in \text{Mod}(\neg\beta)$, nous obtenons $w_1 \in \text{Mod}(\psi_w)$, c'est-à-dire, $w_1 = w$, contradiction.
Cela prouve que $\text{Mod}(\psi_w \lhd \mu) \subseteq \{w\} \cup min_{\leq_w}(\text{Mod}(\neg\mu))$.

Montrons maintenant l'inclusion inverse. Selon (E1), $w \in \text{Mod}(\psi_w \lhd \mu)$. Considérons w_1 appartenant à $min_{\leq_w}(\text{Mod}(\neg\mu))$ et dans le but d'obtenir une contradiction supposons $w_1 \notin \text{Mod}(\psi_w \lhd \mu)$. Dans ce cas μ n'est pas une tautologie et selon (E3) $\psi_w \lhd \mu \not\models \mu$, ainsi il existe w_2 un modèle de $\neg\mu$ qui est dans $\text{Mod}(\psi_w \lhd \mu)$. Si $w_2 = w$ alors $w_2 <_w w_1$, ce qui contredit la minimalité de w_1.
Si $w_2 \neq w$. Puisque $w_2 \in \text{Mod}(\psi_w \lhd \mu)$ nous avons $\psi_w \lhd \mu \not\models \neg(\alpha_1 \vee \alpha_2)$. Puisque w_1 et w_2 sont tous les deux des modèles de $\neg\mu$ observons que $\mu \wedge \neg(\alpha_1 \vee \alpha_2) \equiv \mu$, et donc selon (E7), $\psi_w \lhd \neg(\alpha_1 \vee \alpha_2) \models \psi_w \lhd \mu$. Puisque par hypothèse $w_1 \notin \text{Mod}(\psi_w \lhd \mu)$, nous avons $w_1 \notin \text{Mod}(\psi_w \lhd \neg(\alpha_1 \vee \alpha_2)$. Donc, d'après le Lemme 2 et le Lemme 1, $\text{Mod}(\psi_w \lhd \neg(\alpha_1 \vee \alpha_2)) = \{w, w_2\}$, ce qui contredit la minimalité de w_1. □

4.4 L'effacement en termes de préordres partiels

Rappelons que l'opérateur d'effacement de Winslett ne satisfait pas (E7), donc cet opérateur n'est pas pris en compte par le théorème précédent. Cet opérateur est défini par un préordre partiel et non total. Nous pouvons modifier les postulats de l'effacement de façon à ce qu'ils s'adaptent aux préordres partiels.
Dans la preuve du Théorème 5 étant donné un opérateur d'effacement associé à un préordre total, seul le postulat (E7) exige que le préordre soit total. Par conséquent, comme l'ont fait Katsuno et Mendelzon pour la révision, afin d'obtenir un théorème de représentation par le biais de préordres partiels, nous supprimons le postulat (E7) et le remplaçons par deux postulats plus faibles.

(E9) Si $\psi \models \mu_1 \wedge \mu_2$, $(\psi \lhd \mu_1) \models \psi \vee \neg\mu_2$ et $(\psi \lhd \mu_2) \models \psi \vee \neg\mu_1$, alors $(\psi \lhd \mu_1) \equiv (\psi \lhd \mu_2)$.

(E10) Si ψ est complète, alors $(\psi \lhd \mu_1) \wedge (\psi \lhd \mu_2) \models \psi \lhd (\mu_1 \vee \mu_2)$.

Rappelons qu'effacer une information μ d'une formule ψ revient à ajouter des modèles à ψ, pour chaque modèle w de ψ il s'agit d'ajouter les modèles qui falsifient μ les plus proches de w. Le postulat (E9) est l'analogue du postulat (U6) pour la mise à jour. Supposons que ψ satisfait à la fois μ_1 et μ_2, (E9) dit que si effacer μ_1 de ψ revient à ajouter seulement des modèles qui falsifient μ_2 et si effacer μ_2 de ψ revient à ajouter seulement des modèles qui falsifient μ_1, alors les effacements des deux informations ont le même effet. Le postulat (E10) est l'analogue du postulat (U7) pour la mise à jour, il s'applique aux formules qui ont exactement un modèle. Si un modèle résulte à la fois de l'effacement de μ_1 et de l'effacement de μ_2 d'une formule ψ, alors ce modèle doit également résulter de l'effacement de $\mu_1 \vee \mu_2$ de ψ.
Les opérateurs d'effacement $\lhd_F$ et $\lhd_W$ satisfont tous les deux postulats (E9) et (E10).
La définition de ces postulats permet de concevoir une classe d'opérateurs d'effacement basés sur des préordres partiels. Le théorème suivant montre que les opérateurs d'effacement basés sur des préordres partiels sont complètement caractérisés par les postulats (E1)–(E6) et (E8)–(E10).

Théorème 6. *Un opérateur d'effacement $\lhd$ satisfait les postulats* (E1)–(E6) *et* (E8)–(E10) *si et seulement si il existe une affectation fidèle ponctuelle qui associe à chaque interprétation w un préordre partiel $\leq_w$ tel que* $\text{Mod}(\psi \lhd \mu) = \text{Mod}(\psi) \cup \bigcup_{w \in \text{Mod}(\psi)} min_{\leq_w}(\text{Mod}(\neg\mu))$.

Démonstration. $\Leftarrow$) Supposons que nous avons affectation fidèle ponctuelle qui associe à chaque interprétation

w un préordre partiel $\leq_w$. Considérons l'opérateur d'effacement $\triangleleft$ défini par $\text{Mod}(\psi \triangleleft \mu) = \text{Mod}(\psi) \cup \bigcup_{w \in \text{Mod}(\psi)} min_{\leq_\psi}(\text{Mod}(\neg\mu))$.
Prouvons dans un premier temps que l'opérateur $\triangleleft$ satisfait les postulats (E1)–(E6) et (E8)–(E10). La preuve que $\triangleleft$ satisfait les postulats (E1)–(E6) et (E8) est similaire à la preuve du Théorème 5 ci-dessus, le fait que le préordre est partiel (et non nécessairement total) n'a pas d'impact sur la preuve de ces sept postulats.

Supposons que $\psi \models \mu_1 \wedge \mu_2$, $(\psi \triangleleft \mu_1) \models \psi \vee \neg\mu_2$ et $(\psi \triangleleft \mu_2) \models \psi \vee \neg\mu_1$. Dans le but d'obtenir une contradiction supposons qu'il existe une interprétation w telle $w \in \text{Mod}(\psi \triangleleft \mu_1)$ et $w \notin \text{Mod}(\psi \triangleleft \mu_2)$. Observons que $w \notin \text{Mod}(\psi)$. Puisque $(\psi \triangleleft \mu_1) \models \psi \vee \neg\mu_2$, nous avons $w \in \text{Mod}(\neg\mu_2)$. Par définition de l'opérateur, puisque $w \notin \text{Mod}(\psi \triangleleft \mu_2)$, pour chaque modèle w_i de ψ il existe $w'_i \in min_{\leq_{w_i}}(\text{Mod}(\neg\mu_2))$ tel que $w'_i <_{w_i} w$ et w'_i est minimal dans $\text{Mod}(\neg\mu_2)$. Alors chaque w'_i est un modèle de $(\psi \triangleleft \mu_2)$. Or $(\psi \triangleleft \mu_2) \models \psi \vee \neg\mu_1$, et donc ou bien $w'_i \in \text{Mod}(\psi)$ ou $w'_i \in \text{Mod}(\neg\mu_1)$. Si $w'_i \in \text{Mod}(\psi)$, alors $w'_i \in \text{Mod}(\psi) \cap \text{Mod}(\neg\mu_2)$, ce qui contredit le fait que $\psi \models \mu_2$. Si $w'_i \in \text{Mod}(\neg\mu_1)$, alors $w'_i <_{w_i} w$ ce qui contredit le fait que $w \in min_{\leq_{w_i}}(\text{Mod}(\neg\mu_1))$. Ainsi dans les deux cas nous obtenons une contradiction et nous avons prouvé que $\text{Mod}(\psi \triangleleft \mu_1) \subseteq \text{Mod}(\psi \triangleleft \mu_2)$. L'inclusion inverse est prouvée de la même façon ce qui montre que (E9) est satisfait.

Notons que $\text{Mod}((\psi_w \triangleleft \mu_1) \wedge (\psi_w \triangleleft \mu_2)) = \{w\} \cup min_{\leq_w}(\text{Mod}(\neg\mu_1)) \cap min_{\leq_w}(\text{Mod}(\neg\mu_2)) \subseteq \{w\} \cup min_{\leq_w}\text{Mod}(\neg\mu_1 \wedge \neg\mu_2)$. Donc $(\psi_w \triangleleft \mu_1) \wedge (\psi_w \triangleleft \mu_2) \models \psi_w \triangleleft (\mu_1 \vee \mu_2)$, ce qui prouve que (E10) est satisfait.

$\Rightarrow$) Soit $\triangleleft$ un opérateur d'effacement qui satisfait les postulats (E1)–(E6) et (E8)–(E10). Rappelons que dans le suite étant donné une interprétation w_i, nous notons α_i une formule dont le seul modèle est w_i.
Pour chaque interprétation w nous définissons une relation binaire $\leq_w$ sur les interprétations par :

$$w_1 \leq_w w_2 \text{ si } \text{Mod}(\psi_w \triangleleft \neg(\alpha_1 \vee \alpha_2)) = \{w\} \cup \{w_1\}.$$

Montrons dans un premier temps que $\leq_w$ est un préordre. D'après le Lemme 1, $\text{Mod}(\psi_w \triangleleft \neg\alpha_1) = \text{Mod}(\psi_w) \cup \{w_1\}$, et donc $\leq_w$ est réflexive.
Prouvons maintenant que la relation est transitive. Considérons trois interprétations deux à deux distinctes w_1, w_2 et w_3 telles que $w_1 \leq_w w_2$ et $w_2 \leq_w w_3$. Nous avons donc $\text{Mod}(\psi_w \triangleleft \neg(\alpha_1 \vee \alpha_2)) = \text{Mod}(\psi_w) \cup \{w_1\}$, c'est-à-dire $\psi \triangleleft \neg(\alpha_1 \vee \alpha_2) \equiv \psi_w \vee \alpha_1$ et $\text{Mod}(\psi_w \triangleleft \neg(\alpha_2 \vee \alpha_3)) = \text{Mod}(\psi_w) \cup \{w_2\}$, c'est-à-dire $\psi_w \triangleleft \neg(\alpha_2 \vee \alpha_3) \equiv \psi_w \vee \alpha_2$. Supposons tout d'abord qu'une de ces interprétations est égale à w. Si $w_1 = w$, alors par (E2), $\text{Mod}(\psi_w \triangleleft \neg(\alpha_1 \vee \alpha_3)) = \text{Mod}(\psi_w)$ et $w_1 \leq_w w_3$. Si $w_2 = w$, alors selon les hypothèses et par (E2), on a également $w_1 = w$ et donc $w_1 = w_2$. Si $w_3 = w$ alors selon les hypothèses et par (E2) nous avons $w_1 = w_2 = w_3 = w$. Supposons maintenant qu'aucune des interprétations w_1, w_2 et w_3 n'est égale à w, et donc que $\psi_w \models (\neg\alpha_1 \wedge \neg\alpha_2 \wedge \neg\alpha_3)$. D'une part $\psi_w \triangleleft (\neg\alpha_1 \wedge \neg\alpha_2) \equiv \psi_w \vee \alpha_1$, d'où $\psi_w \triangleleft (\neg\alpha_1 \wedge \neg\alpha_2) \models \psi_w \vee (\alpha_1 \vee \alpha_2 \vee \alpha_3)$. D'autre part d'après (E6), $\psi_w \triangleleft (\neg\alpha_1 \wedge \neg\alpha_2 \wedge \neg\alpha_3) \models (\psi_w \triangleleft \neg\alpha_1) \vee (\psi_w \triangleleft (\neg\alpha_2 \wedge \neg\alpha_3))$. Donc sous nos hypothèses $\psi_w \triangleleft (\neg\alpha_1 \wedge \neg\alpha_2 \wedge \alpha_3) \models (\psi_w \triangleleft \neg\alpha_1) \vee \psi_w \vee \alpha_2$. Selon le Lemme 1, $\psi_w \triangleleft \neg\alpha_1 \equiv \psi_w \vee \alpha_1$. D'où $\psi_w \triangleleft (\neg\alpha_1 \wedge \neg\alpha_2 \wedge \neg\alpha_3) \models \psi_w \vee \alpha_1 \vee \alpha_2$. Donc par (E9) $\psi_w \triangleleft (\neg\alpha_1 \wedge \neg\alpha_2) \equiv \psi_w \triangleleft (\neg\alpha_1 \wedge \neg\alpha_2 \wedge \neg\alpha_3)$. En conséquence, d'une part, $\text{Mod}(\psi_w \triangleleft (\neg\alpha_1 \wedge \neg\alpha_2 \wedge \neg\alpha_3)) = \{w, w_1\}$, et donc $\psi_w \triangleleft (\neg\alpha_1 \wedge \neg\alpha_2 \wedge \neg\alpha_3) \models \psi_w \vee \alpha_1 \vee \alpha_3$. D'autre part par (E5), $(\psi_w \triangleleft (\neg\alpha_1 \wedge \neg\alpha_3)) \wedge (\neg\alpha_1 \wedge \neg\alpha_3)) \models \psi_w$, et donc $(\psi_w \triangleleft (\neg\alpha_1 \wedge \neg\alpha_3)) \models \psi_w \vee (\alpha_1 \vee \alpha_3) \models \psi_w \vee (\alpha_1 \vee \alpha_2 \vee \alpha_3)$. Par (E9) nous obtenons que $\psi_w \triangleleft (\neg\alpha_1 \wedge \neg\alpha_3) \equiv \psi_w \triangleleft (\neg\alpha_1 \wedge \neg\alpha_2 \wedge \neg\alpha_3)$. On en déduit que $\psi_w \triangleleft (\neg\alpha_1 \wedge \neg\alpha_3) \equiv \psi_w \triangleleft (\neg\alpha_1 \wedge \neg\alpha_2)$. Donc $\text{Mod}(\psi_w \triangleleft \neg(\alpha_1 \vee \alpha_3)) = \text{Mod}(\psi_w) \cup \{w_1\}$, c'est-à-dire $w_1 \leq_w w_3$, ce qui prouve la transitivité de la relation $\leq_w$.

Il découle du postulat (E2) que l'application $w \mapsto \leq_w$ est une assignation fidèle ponctuelle.

Il reste à prouver que $\text{Mod}(\psi \triangleleft \mu) = \text{Mod}(\psi) \cup \bigcup_{w \in \text{Mod}(\psi)} min_{\leq_w}(\text{Mod}(\neg\mu))$. Si ψ is insatisfaisable, alors les deux membres de l'équation sont vides et l'égalité est trivialement vérifiée. Si ψ est satisfaisable, alors puisque le postulat (E8) est vérifié, étant donné une interprétation w il est suffisant de prouver que $\text{Mod}(\psi_w \triangleleft \mu) = \{w\} \cup min_{\leq_w}(\text{Mod}(\neg\mu))$. Si $w \in \text{Mod}(\neg\mu)$ alors il découle de (E2) que $\psi_w \triangleleft \mu \equiv \psi_w$ et l'égalité est vérifiée puisque l'assignation est fidèle. Si μ est une tautologie alors selon (E5) $\psi_w \triangleleft \mu \equiv \psi_w$ et l'égalité est vérifiée. Nous supposons désormais que $w \notin \text{Mod}(\neg\mu)$ et que μ n'est pas une tautologie.
Prouvons tout d'abord que $\text{Mod}(\psi_w) \cup min_{\leq_w}(\text{Mod}(\neg\mu)) \subseteq \text{Mod}(\psi_w \triangleleft \mu)$. On sait par (E1) que $w \in \text{Mod}(\psi_w \triangleleft \mu)$. Considérons maintenant $w_0 \neq w$ et $w_0 \in min_{\leq_w}(\text{Mod}(\neg\mu))$.
Supposons que $\text{Mod}(\neg\mu) = \{w_0, w_1, \ldots, w_n\}$. Pour toute interprétation $w_i \in \text{Mod}(\neg\mu)$, puisque ni $\neg\alpha_0$ ni $\neg\alpha_i$ ne sont des tautologies il découle du Lemme 2 que $\psi_w \triangleleft (\neg\alpha_0 \wedge \neg\alpha_i) \equiv \psi_w \vee \alpha_0$ ou $\psi_w \vee \alpha_i$ ou $\psi_w \vee \alpha_0 \vee \alpha_i$. Puisque $w_0 \in min_{\leq_w}(\text{Mod}(\neg\mu))$ il n'existe aucun $w_i \in \text{Mod}(\neg\mu)$ tel que $\psi_w \triangleleft (\neg\alpha_0 \wedge \neg\alpha_i) \equiv \psi_w \vee \alpha_i$, et donc $w_0 \in \psi_w \triangleleft (\neg\alpha_0 \wedge \neg\alpha_i)$.
Observons que $\mu \equiv \bigvee_{i=1}^{n}(\alpha_0 \vee \alpha_i)$. Donc selon (E4) et en appliquant de façon répétée (E10) nous obtenons que $w_0 \in \text{Mod}(\psi_w \triangleleft \mu)$, prouvant ainsi que
$\text{Mod}(\psi_w) \cup min_{\leq_\psi}(\text{Mod}(\neg\mu)) \subseteq \text{Mod}(\psi_w \triangleleft \mu)$.
Montrons maintenant l'inclusion inverse, $\text{Mod}(\psi_w \triangleleft \mu) \subseteq \text{Mod}(\psi_w) \cup min_{\leq_w}(\text{Mod}(\neg\mu))$. Considérons $w_0 \in \text{Mod}(\psi_w \triangleleft \mu)$ tel que $w_0 \neq w$. Par (E5) nous avons $w_0 \in \text{Mod}(\neg\mu)$. Afin d'obtenir une contradiction supposons que $w_0 \notin min_{\leq_w}(\text{Mod}(\neg\mu))$. Cela signifie qu'il existe $w_1 \in \text{Mod}(\neg\mu)$ tel que $w_1 <_w w_0$, c'est-à-dire, $\psi_w \triangleleft (\neg\alpha_0 \wedge \neg\alpha_1)) \equiv \psi_w \vee \alpha_1$. Considèrons maintenant la formule $\beta = \neg\mu \wedge \neg\alpha_0 \wedge \neg\alpha_1$. Clairement nous avons $\neg\mu \equiv \beta \vee \alpha_0 \vee \alpha_1$. Selon (E4), $\psi_w \triangleleft \mu \equiv \psi_w \triangleleft (\neg\beta \wedge \neg\alpha_0 \wedge \neg\alpha_1)$. Par (E6), $\psi_w \triangleleft (\neg\beta \wedge \neg\alpha_0 \wedge \neg\alpha_1) \models$

$(\psi_w \lhd \neg\beta) \vee (\psi_w \lhd (\neg\alpha_0 \wedge \neg\alpha_1))$. Puisque $w_1 <_w w_0$, $w_0 \notin \mathrm{Mod}(\psi_w \lhd (\neg\alpha_0 \wedge \neg\alpha_1)$, et donc $w_0 \in \mathrm{Mod}(\psi_w \lhd \neg\beta)$. On a également $w_0 \in \mathrm{Mod}(\neg\mu)$ et *a fortiori* $w_0 \in \mathrm{Mod}(\neg\beta)$. Alors selon (E5), $(\psi_w \lhd \neg\beta) \wedge (\neg\beta) \models \psi_w$, et nous obtenons $w_0 = w$, ce qui fournit une contradiction. □

4.5 La contraction en termes de préordres partiels

Comme nous l'avons observé dans la Section 3.2, il manque toujours un théorème de représentation pour les opérateurs de contraction par le biais de préordres partiels. Notre objectif est de combler cette lacune. De la même manière que pour l'effacement, nous sommes en mesure de donner une version des postulats de contraction qui prend en compte les préordres partiels et nous pouvons ainsi concevoir une classe d'opérateurs de contraction basés sur les préordres partiels.
Nous supprimons le postulat (C7) et le remplaçons par deux postulats plus faibles, (C8) et (C9). Ils sont similaires aux postulats (E9) et (E10), à l'exception du dernier qui n'est plus restreint aux formules complètes.

(C8) Si $\psi \models \mu_1 \wedge \mu_2$, $(\psi - \mu_1) \models \psi \vee \neg\mu_2$ et $(\psi - \mu_2) \models \psi \vee \neg\mu_1$, alors $(\psi - \mu_1) \equiv (\psi - \mu_2)$.
(C9) $(\psi - \mu_1) \wedge (\psi - \mu_2) \models \psi - (\mu_1 \vee \mu_2)$.

Les opérateurs de contraction $-_D$ et $-_S$ satisfont (C8) et (C9).

Le théorème suivant montre que les opérateurs de contraction basés sur des préordres partiels sont parfaitement caractérisés par les postulats (C1)–(C6), (C8) et (C9).

Théorème 7. *Un opérateur de contraction $-$ satisfait les postulats* (C1)–(C6) *et* (C8)–(C9) *si et seulement si il existe une affectation fidèle qui associe à chaque formule ψ un préordre partiel $\leq_\psi$ tel que* $\mathrm{Mod}(\psi - \mu) = \mathrm{Mod}(\psi) \cup min_{\leq_\psi}(\mathrm{Mod}(\neg\mu))$.

La preuve de ce thèorème suit exactement les mêmes lignes que la preuve du théorème 6.

5 Panorama des opérations de changement de croyances

Nous avons proposé de nouveaux postulats capturant le principe de changement minimal pour l'effacement, à savoir (E6) et (E7). Ceci nous a permis d'établir un premier théorème de représentation montrant qu'un opérateur d'effacement satisfaisant l'ensemble des postulats (E1)–(E8) correspond à un préordre total sur les interprétations. De plus, le remplacement du postulat (E7) par deux postulats plus faibles (E9) et (E10) nous a permis d'établir un second théorème de représentation montrant qu'un opérateur d'effacement satisfaisant l'ensemble des postulats (E1)–(E6) et (E8)–(E10) correspond à un préordre partiel sur les interprétations.
De plus, pour l'opération de contraction, nous avons montré qu'en remplaçant le postulat (C7) par deux postulats plus faibles (C8) et (C9), nous pouvons établir un théorème de représentation, qui manquait jusqu'à présent, montrant qu'un opérateur de contraction satisfaisant l'ensemble (C1)–(C6) et (C8)–(C9) correspond à un préordre partiel sur les interprétations.
Notre contribution permet ainsi de dresser un panorama intéressant de quatre opérations de changement de croyance bien connues, résumé dans le tableau suivant.

	Post. basiques.	Post. min.	Post. min.-
Rev.	(R1)-(R4)	(R5), **(R6)**	(R7), (R8)
Cont.	(C1)-(C5)	(C6), **(C7)**	(C8), (C9)
Màj	(U1)-(U4), (U8)	(U5), **(U9)**	(U6), (U7)
Eff.	(E1)-(E5), (E8)	(E6), **(E7)**	(E9), (E10)

TABLE 2 – Panorama des opérations de changement de croyances

Chaque ligne correspond à une opération. La première colonne représente les postulats de base pour chaque opération. Noter que la mise à jour et l'effacement ont un postulat spécifique, à savoir (U8) et (E8). Ces postulats indiquent que la mise à jour et l'effacement prennent en compte de manière égale chacun des modèles des croyances initiales. Les postulats de la deuxième colonne reflètent le principe de minimalité des changements. Pour chaque opération de changement de croyance, les postulats apparaissant dans ces deux premières colonnes sont ceux requis pour énoncer un théorème de représentation en termes de préordres totaux sur les interprétations. Pour chaque opération, le postulat en gras doit être supprimé et remplacé par deux postulats plus faibles, indiqués dans la troisième colonne, pour tenir compte des préordres partiels. Tous les théorèmes de représentation donnés dans ce document peuvent être facilement lus à partir de ce tableau. Les postulats soulignés sont ceux que nous avons introduits dans ce document.
Il est intéressant de noter que la numérotation des postulats pour la mise à jour dans ce tableau semble étrange (le dernier postulat (U9) apparaît tôt en tant que postulat capturant la minimalité du changement). Cela est dû au fait qu'à l'origine, Katsuno et Mendelzon [15] ont obtenu un théorème de représentation pour la mise à jour par le biais de préordres partiels. Contrairement aux autres opérations, le théorème de représentation en termes de préordres totaux a été obtenu dans un deuxième temps. En particulier, le postulat (U9) implique (U6) et (U7).

6 Conclusion

Dans cet article consacré à l'effacement des croyances en logique propositionnelle nous poursuivons et complètons les travaux initiés par Katsuno et Mendelzon [15]. Ils ont formellement défini l'effacement des croyances dans un cadre sémantique et ont proposé un ensemble de postulats basiques. Pour la révision et la mise à jour, ils ont proposé des postulats supplémentaires qui reflètent le principe de minimalité des changements. Ils ont ensuite montré qu'un

opérateur de révision (ou de mise à jour) satisfait ces postulats si et seulement s'il est induit par un préordre total ou partiel sur les interprétations. En 2017, Caridroit, Konieczny et Marquis [4] ont poursuivi cette étude en s'intéressant à la contraction. Ils ont obtenu un théorème de représentation pour la contraction en termes de préordres totaux.

Dans cet article, nous avons considéré l'effacement et avons d'abord adapté leur travail en définissant des postulats similaires aux leurs pour capturer la minimalité du changement. Dans un deuxième temps, nous avons affaibli l'un de ces postulats pour tenir compte des préordres partiels, et nous avons ainsi obtenu un théorème de représentation pour les opérateurs d'effacement en termes de préordres partiels. Comme sous-produit, nous avons obtenu un résultat similaire pour la contraction. Enfin, nous avons dressé un tableau complet des quatre opérations fondamentales de modification des croyances, à savoir la révision, la mise à jour, la contraction et l'effacement, dans le cadre sémantique, comme le montre la table 2.

Une suite naturelle à ce travail serait l'étude de l'opération, appelée *Forget*, proposée par Winslett [22], qu'elle compare à la contraction. Soit ψ et μ deux formules propositionnelles et soit $\diamond$ un opérateur de mise à jour, l'opération de Forget est équivalente à $(\psi \diamond \mu) \vee (\psi \diamond \neg\mu)$.

Une autre perspective serait l'étude de la contraction et de l'effacement itérés. Alors que de nombreux travaux ont été développés sur la révision itérée suite aux travaux de Darwiche et Pearl [7], la contraction itérée n'a suscité que peu d'intérêt jusqu'à présent si ce n'est le travail de Hild et Spohn [12].

Par ailleurs, une étude plus ambitieuse serait l'étude de la complexité de problèmes de décision comme la vérification de modèles pour les opérateurs de contraction et d'effacement.

Références

[1] C.E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change : Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50 :510–530, 1985.

[2] C.E. Alchourrón and D. Makinson. On the logic of theory change : Safe contraction. *Studia Logica*, 44(4) :14–37, 1985.

[3] T. Caridroit, S. Konieczny, and P. Marquis. Contraction in propositional logic. In *Proceedings of ECSQARU'15*, pages 186–196, 2015.

[4] T. Caridroit, S. Konieczny, and P. Marquis. Contraction in propositional logic. *Int. J. Approx. Reason.*, 80 :428–442, 2017.

[5] N. Creignou, R. Ktari, and O. Papini. Belief contraction and erasure in fragments of propositional logic. *J. Log. Comput.*, 32(7) :1436–1468, 2022.

[6] M. Dalal. Investigations into a theory of knowledge base revision : preliminary report. In *Proceedings of AAAI'88*, pages 475–479, 1988.

[7] A. Darwiche and J. Pearl. On the logic of iterated belief revision. *Artif. Intell.*, 89(1-2) :1–29, 1997.

[8] T. Eiter and G. Gottlob. On the complexity of propositional knowledge base revision, updates, and counterfactuals. *Artificial Intelligence*, 57(2–3) :227–270, 1992.

[9] K. D. Forbus. Introducing actions into qualitative simulation. In *Proceedings International Joint Conferences on Artificial Intelligence Organization (IJCAI)*, pages 1273–1278, 1989.

[10] P. Gärdenfors. Knowledge in flux. In *Cambridge University Press, Cambridge UK*, 1988.

[11] A. Herzig and O. Rifi. Propositional belief base update and minimal change. *Artificial Intelligence*, 115(1) :107–138, 1999.

[12] Matthias Hild and Wolfgang Spohn. The measurement of ranks and the laws of iterated contraction. *Artif. Intell.*, 172(10) :1195–1218, 2008.

[13] H. Katsuno and A. O. Mendelzon. A unified view of propositional knowledge base updates. In *Proceedings of IJCAI'89*, pages 1413–1419, 1989.

[14] H. Katsuno and A.O. Mendelzon. Propositional knowledge base revision and minimal change. *Artificial Intelligence*, 52(3) :263–294, 1991.

[15] H. Katsuno and A.O. Mendelzon. On the difference between updating a knowledge base and revising it. In P. Gärdenfors, editor, *Belief revision*, pages 183–203. Cambridge University Press, 1992.

[16] A.M. Keller and M. Winslett. On the use of an extended relational model to handle changing incomplete information. *IEEE Trans. Software Eng.*, 11(7) :620–633, 1985.

[17] R. Ktari. *Changement de croyances dans des fragments de la logique propositionnelle*. PhD thesis, Aix-Marseille Université, 5 2016.

[18] P. Marquis, H. Prade, and O. Papini, editors. *Panorama de l'intelligence artificielle : Volume 1 : Représentation des connaissances et formalisation des raisonnements*. Cepaduès, 2014.

[19] P. Marquis, H. Prade, and O. Papini, editors. *A Guided Tour of Artificial Intelligence Research : Volume I : Knowledge Representation, Reasoning and Learning*. Springer, 2020.

[20] K. Satoh. Nonmonotonic reasoning by minimal belief revision. In *Proceedings of FGCS'88*, pages 455–462, Tokyo, 1988.

[21] M. Winslett. Reasoning about action using a possible models approach. In *Proc. AAAI*, pages 89–93, 1988.

[22] M. Winslett. Sometimes updates are circumscription. In *Proceedings International Joint Conferences on Artificial Intelligence Organization (IJCAI)*, pages 859–863, 1989.

Multi-objective reinforcement learning: an ethical perspective

T. Deschamps[1], R. Chaput[1], L. Matignon[1]

[1] Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205, F-69622 Villeurbanne, France

timon.deschamps@liris.cnrs.fr

Abstract

Reinforcement learning (RL) is becoming more prevalent in practical domains with human implications, raising ethical questions. Specifically, multi-objective RL has been argued to be an ideal framework for modeling real-world problems and developing human-aligned artificial intelligence. However, the ethical dimension remains underexplored in the field, and no survey covers this aspect. Hence, we propose a review of multi-objective RL from an ethical perspective, highlighting existing works, gaps in the literature, important considerations, and potential areas for future research.

Keywords

Reinforcement learning, multi-objective decision making, machine ethics.

Résumé

L'apprentissage par renforcement est de plus en plus employé pour des applications pratiques impactant l'humain, soulevant ainsi des questions éthiques. Spécifiquement, l'apprentissage par renforcement multi-objectif est considéré comme un cadre idéal pour la modélisation de problèmes concrets et le développement de systèmes d'intelligence artificielle alignés sur l'humain. Peu de travaux du domaine adoptent une perspective éthique, et les études existantes ne couvrent pas cet aspect. Ainsi, nous proposons une revue de l'apprentissage par renforcement multi-objectif d'un point de vue éthique, en détaillant les travaux existants, les lacunes de la littérature, les considérations importantes, et les potentielles pistes de recherche futures.

Mots-clés

Apprentissage par renforcement, prise de décision multi-objectifs, éthique computationnelle.

1 Introduction

The field of *reinforcement learning* (RL) has recently seen numerous breakthroughs, notably featuring artificial intelligence (AI) agents beating humans at a wide variety of games [50, 8]. RL has also been applied to multiple real-world problems, with a potentially large impact on societies, e.g., nuclear fusion control [16], healthcare [69]. This calls for the study of the ethical issues that may arise from such uses, and the development of techniques to ensure that the agents have a behavior deemed *ethically-aligned* with human principles; so as to guarantee this technology will be beneficial to humanity. This is a complex endeavor, and a few works have started paving the way [66, 52].

In this paper, we focus on *multi-objective reinforcement learning* (MORL), a sub-field of RL in which multiple potentially conflicting goals are considered rather than a single one. Following the RL trend, MORL is being increasingly used in real world applications such as public bicycle dispatching [14] or energy management [19]. It has been argued that aligning AI with human goals is a multi-objective problem [58], making the study of MORL interesting in this regard. A few multi-objective decision making surveys have been published [25, 48], focusing on the theory and applications of multi-objective decision making algorithms. The goal of this work is to highlight the need for ethically-aligned multi-objective methods and to conduct an analysis of MORL from a moral standpoint. To do so, we start by discussing and categorizing existing MORL methods, before introducing important ethical considerations, which we use to emphasize important gaps in the literature.

2 A motivating example

To illustrate the ethical concerns that can arise when AI agents are deployed in the real-world, we propose to study the case of self-driving vehicles. This sector has been increasingly interested in RL [28], which is viewed as a suitable paradigm: vehicles can be represented by agents taking actions such as steering and accelerating within an environment (road network).

RL agents typically optimize for a single objective, e.g., *speed*. However, when dealing with complex use-cases or when humans can be impacted, more flexibility is desirable to account for additional goals like *cost saving* and *comfort*.

MORL is ideal in such contexts, as it allows for representing and compromising between multiple objectives. This multi-objective aspect is essential when autonomous vehicles are deployed on real roads, as human error, technical malfunctions or unexpected situations will inevitably occur, leading the machine to have to handle complex ethical dilemmas which require weighting between conflicting moral values, e.g., ensuring safety for both passengers and surrounding pedestrians in an inevitable accident scenario.

This example motivates the study of MORL agents with an ethically-aligned behavior, and we will extend it throughout this paper to illustrate some of the notions discussed.

3 Background

3.1 Reinforcement learning

Reinforcement learning is a general framework to solve problems in which an agent alternatively takes *actions* and receives *observations* and *rewards* from an environment, and aims at maximizing the cumulative reward obtained. RL is usually modeled as a *Markov decision process* (MDP), defined as a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma \rangle$, where:

- $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, respectively;
- $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition function, i.e., the probability of transitioning to a state s_{t+1} given that the action a_t was taken at time step t in state s_t;
- $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the reward function, which outputs a scalar reward for a given (s_t, a_t, s_{t+1}) tuple;
- $\gamma \in [0, 1)$ is a discount factor modulating the importance of long term rewards.

The agent acts according to a *stochastic policy* $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$, which gives the probability of taking any action $a \in \mathcal{A}$ given the current state $s \in \mathcal{S}$. If in every state one of the actions is selected with probability 1, the policy becomes *deterministic*, denoted $\pi : \mathcal{S} \to \mathcal{A}$[1].

At any time step t, we can compute the sum of future rewards, or *return*, defined as:

$$G_t = R_{t+1} + \gamma R_{t+2} + \ldots = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k. \quad (1)$$

The *value* of a state $V^\pi(s) = \mathbb{E}_\pi [G_t \mid S_t = s]$ is the expected return for an agent located in this state at time step t and following policy π. In turn, our goal is to find the optimal policy π^* which, when followed, maximizes the value for all states in $\mathcal{S}$.

To this day, RL remains a highly active discipline, with many emerging sub-fields such as multi-agent RL [23, 11], model-based RL [33] and multi-objective RL, the latter of which we discuss in the following section.

3.2 Multi-objective reinforcement learning

The field of multi-objective reinforcement learning (MORL) deals with *multi-objective Markov decision processes* (MOMDPs). MOMDPs differ from regular MDPs only in that the reward (and by extension the value) is vector-valued: $\mathbf{r} \in \mathbb{R}^m$ with m objectives[2]. This implies that finding a single optimal policy via a simple maximization process becomes impossible, as maximizing one of the component of the reward vector (called objective) could lead to a decrease in another one.

Utility functions, also referred to as scalarization functions, map the value vector $\mathbf{V}^\pi$ of a given policy π to a single scalar ($u : \mathbb{R}^m \to \mathbb{R}$). They provide a convenient way to formalize a decision maker's preferences and trade-offs over the objectives.

A common and simple class of utility functions are linear utilities, denoted as $u(\mathbf{V}^\pi) = \mathbf{w}^\top \mathbf{V}^\pi$, which combines a weight vector $\mathbf{w}$ in the $(m-1)$-simplex[3] and the value vector using a linear combination. Intuitively, each weight $w_o \in \mathbf{w}$ represents the importance of the associated objective $\mathbf{V}^\pi_o$.

If we have access to a linear utility function for the user, we can use it to simplify the problem back into the single-objective RL setting and solve it with classical methods. However, this is not an option when the utility function is not fully known in advance or is non-linear, which represents a large portion of real-life scenarios (see the motivating scenarios presented in [25]).

In these settings, we focus instead on a set of optimal policies: the *Pareto front* (PF). A policy $\pi \in \Pi$ belongs to the Pareto front PF(Π) if it is not Pareto-dominated by any other policy. The Pareto-dominance of a policy π over a policy π' is defined as:

$$\pi \succ_P \pi' := (\forall o : \mathbf{V}^\pi_o \geq \mathbf{V}^{\pi'}_o) \wedge (\exists o : \mathbf{V}^\pi_o > \mathbf{V}^{\pi'}_o). \quad (2)$$

In plain words, π's associated value vector is greater or equal to the one associated with π' for all objectives o, and strictly greater for at least one.

As the PF can have multiple policies with the same induced value function, we often refer to a *Pareto coverage set* (PCS), which simply retains a single policy for each non Pareto-dominated value function. Computing a PCS guarantees that we have access to all policies that are optimal under some monotonically increasing utility function. This allows to adapt to changes in the user's preferences while making minimal assumptions about u. In practice, however, PF and PCS can be prohibitively large to compute. Recent works [48, 25, 41] have argued for a utility-based approach, in which we use information we have about the utility function to guide our search in the space of policies. For example, when u is known to be linear, we can restrict our focus to subsets of the PF referred as *convex coverage sets* (CCS), which contain all maximal policies under this assumption.

To illustrate these concepts, let's take our example from section 2. Keeping only 2 objectives (speed and comfort) for ease of representation, we can visualize the PF and a CCS in figure 1. Each point represents a policy and its associated value vector, compromising between the two objectives. We can see that increasing speed usually leads to a decrease in comfort, but it is not always the case (for instance, faster speeds on very uneven roads could smooth out the cruise). Notice that points belonging to the represented CCS are also part of the Pareto front (in fact CCS(Π) $\subseteq$ PF(Π)). Here, point b is not Pareto-dominated by any other point. Furthermore, there is no $\mathbf{w}$ for which a linear scalarization would lead to b being maximal. Thus, we can conclude that b belongs to the PF but not to a CCS.

When using a scalarization function, two optimization criteria naturally arise: *scalarized expected returns* (SER) and

[1] Some works also use $\mu(s) = a$ specifically for deterministic policies.

[2] Note that we use the standard notation of boldface for vector variables.

[3] The m-simplex, denoted Δ^m, is the set of all nonnegative vectors of $m + 1$ dimensions whose components sum to 1.

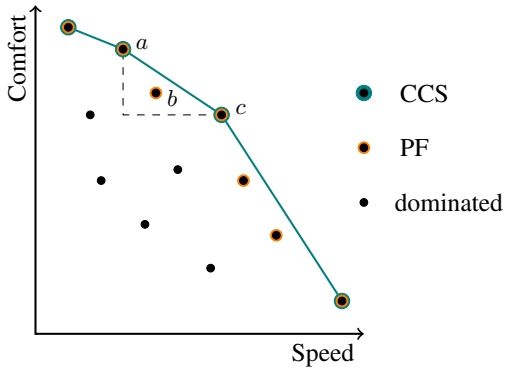


Figure 1: Visualization of the Pareto front and a convex coverage set for a 2-objective self-driving car example.

expected scalarized returns (ESR). To optimize for SER, we scalarize an expectation over multiple runs of the vector-valued returns of a policy, whereas optimizing for ESR requires having a scalarized return for each run, and then computing an expectation over them. These two criteria have different properties and should be used in different scenarios. SER, the most studied one, is particularly suited when we aim to optimize over many policy executions, whereas using the ESR criterion is better to ensure that each execution is maximal over our utility function.

See [48, 25] for a detailed overview of the theory and methods of multi-objective reinforcement learning.

3.3 Machine ethics

As autonomous machines are increasingly integrated into domains with significant human implications, their impact, whether it be positive or negative, requires investigation. *Machine ethics* is concerned with ensuring that AI agents demonstrate ethically-aligned behaviors, i.e., behaviors whose outcomes are acceptable according to some human-chosen ethical framework [6]. In turn, we aim for them to be *explicit ethical agents* [35], i.e., agents who are not simply constrained to avoid unethical behaviors but who integrate algorithmic capabilities [18] allowing them to perform ethics-related computations and to consider ethical considerations in their decision-making process. To evaluate the ethical alignment of these behaviors, we leverage insights from *normative ethics*. As it is concerned with the morality of actions, this field provides a suitable framework for such an analysis.

Normative ethics encompasses three main schools of thought: *consequentialism*, *virtue ethics* and *deontology*. According to consequentialism, only the outcomes of actions are necessary to judge whether these actions are ethical or not. Consequentialist ethics are most known for utilitarianism, which argues that in every situation, the ethical action is the one that maximizes happiness and well-being for all. Virtue ethics shift the focus from the action to its motivation. In this view, an agent is ethical if it acts according to set values (e.g., confidence, honour, freedom). Deontology takes a rule-based approach, in which actions can either be right or wrong according to a list of principles. Kantian ethics is a prime example of deontological ethical theories. We refer the interested reader to [54] for an extensive review of western moral philosophy.

As discussed in section 3.1, a defining feature of reinforcement learning agents is their ability to take actions in an environment, making normative ethics a natural framework for studying the ethical alignment of their behavior.

In fact, reinforcement learning has been characterized as an ideal framework to develop ethical agents [1], and recent work has surveyed RL-based moral learning agents [52]. Furthermore, we argue that the formulation of the reinforcement learning objective as the maximization of a future reward signal naturally aligns with a number of branches of consequentialism. Although some methods allow for the application of deontological ethics into RL [24, 5], none to our knowledge directly takes a moral perspective and is adapted to the multi-objective setting. Finally, it has been argued that MORL, on top of being ideal to model a number of real-world problems [25], is a particularly fitting framework to develop human-aligned artificial intelligence [58]. Moreover, we suggest that it is also suited for modeling virtue ethics, as each component of the vector-valued reward can encode a virtue to be followed. For a comprehensive overview of machine ethics implementations, refer to the survey of Tolmeijer et al. [55].

4 Classical MORL methods

The most commonly used taxonomy for multi-objective sequential decision making [48, 25] classifies methods depending on the type of policy and utility function they consider, resulting in a number of criteria:

- *single vs. multiple policies*: As mentioned in sec. 3.2, algorithms can either output a single solution (if the utility is fixed and known in advance) or a set of optimal policies. Multi-policy methods are **more costly**, but allow for **greater flexibility**: since fewer assumptions are made on the utility function, the user can adapt in the face of new data or changing contexts.

- *deterministic vs. stochastic policies*: While it was shown that **stochastic policies can outperform deterministic ones** in some environments [63, 56], their use can become ethically questionable or impossible in domains requiring strong guarantees (e.g., medical treatments).

- *linear vs. monotonically increasing* u: Using linear utility functions **simplifies the learning process**, allowing the MORL problem to be reduced to a single-objective one (for single-policy algorithms) or to restrict the policy search to a CCS (for multi-policy algorithms). Using monotonically increasing utility functions enables the expression of a much **richer relationship between the objectives**, at the cost of a **more complex learning process**, as the entire PF has to be considered.

	single policy (known u)		multiple policies (unknown u)	
	deterministic	stochastic	deterministic	stochastic
linear scalarization	one policy in Π_{DS}: DQN [32], REINFORCE [51]		CCS of policies in Π_{DS}: Envelope [68], PG-MORL [67], PD-MORL [9], CN [2]	
monotonically increasing scalarization	one policy in Π_{D}: EUPG [46], MOCAC [43], Q-steering [60]	mixture of policies in Π_{DS}: π-mix [56], S-rand [63]	PCS of policies in Π_{D}: PQL [34], PCN [42]	mixture of policies in Π_{DS}: CAPQL [29], π-mix [56], S-rand [63]

Table 1: Non-exhaustive classification of MORL algorithms, following the common utility-based taxonomy from [48, 25]. Here, Π_{D} and Π_{DS} denote the policy space restricted to deterministic and deterministic stationary policies, respectively.

For each combination of criteria, this taxonomy allows us to define a *solution set*, i.e., the type of policies that will constitute the solution to our given problem. In table 1, we categorize a non-exhaustive list of popular MORL methods according to said taxonomy. In this section, we present each class of solution set alongside its corresponding methods.

4.1 Linear scalarization

When the utility function is linear, Roijers et al. [48] show that **deterministic stationary[4] policies are optimal**. Furthermore, adding non-stationarity and stochasticity greatly increases the size of the policy space. Thus, MORL methods developped for linear utility functions tend to limit their search to deterministic stationary policies. In scenarios where u is known, only a single optimal policy is required. Conversely, when the utility is unknown or may change, we seek to retrieve a convex coverage set.

Note that by definition, the SER and ESR optimization criteria are equivalent under linear utility, and as such no distinction is made between them in this section.

4.1.1 One deterministic stationary policy

When a linear utility function is used, any single policy MORL problem can be cast into single-objective RL by scalarizing the reward vector. This setting can be solved with most of the existing RL methods (e.g., value-based methods, policy gradients).

For example, take the autonomous driving example discussed in section 2. Let's assume our user is budget-conscious, not in a hurry, and has recurrent back pain. They might then decide on a preference (weight) vector of $[0.1, 0.5, 0.4]$, meaning that they assign an importance factor of 0.1 to speed, 0.5 to cost saving, and 0.4 to comfort. When driving towards a speed bump, the car can either brake or accelerate. The brake option yields a reward of $[-0.4, 0.4, 2.1]$ which gets scalarized to $0.1 \cdot -0.4 + 0.5 \cdot 0.4 + 0.4 \cdot 2.1 = 1$. Accelerating gives $[5, -0.2, -1]$, resulting in a scalarized reward of $u([5, -0.2, -1]) = 0$. This indicates that braking is to be favored in this context. When the agent receives a reward vector from the environment, single-objective RL methods like REINFORCE [51] or DQN [32] can scalarize it as such before using the resulting value as their reward input.

[4] A policy π is stationary if the distribution of actions is constant in all states, i.e., it is not conditional on time step-dependent information.

4.1.2 CCS of deterministic stationary policies

As mentioned in section 3.2, using a linear utility function implies that all optimal policies lie on a convex coverage set. This means that a multi-policy algorithm able to recover a CCS has access to an optimal policy for any possible weight vector $\mathbf{w}$.

Most algorithms use some form of neural network conditioned on a weight vector in their architecture and train it with random values, allowing the model to produce robust outputs over any input $\mathbf{w}$. Conditioned Networks (CN) [2] popularized this approach by showing the potential of conditioned deep Q-networks to generalize across the weight space. Following work kept the same general structure, while focusing on efficient exploration and alignment of weight vectors. The authors of Envelope [68] propose to use multiple schemes such as homotopy optimization and Hindsight Experience Replay [7] and show that it allows them to consistently outperform CN. PG-MORL [67] was one of the first methods to tackle environments with large continuous action spaces. It features an evolutionary stage that allows it to efficiently search the space of policies and weights to best improve the CCS. PD-MORL [9] was able to beat Envelope and PG-MORL (on discrete and continuous action tasks respectively) by adding a preference guidance term to a double deep Q-network loss [62]. Note that some of these works use the terms Pareto coverage sets and convex coverage sets interchangeably, but their nature in fact strictly limit them to the retrieval of CCS.

4.2 Monotically increasing scalarization

When the utility function is non-linear, **deterministic stationary policies are not guaranteed to be optimal**. To retrieve policies from the Pareto front that do not lie on convex coverage sets, we need to introduce either non-stationarity or stochasticity.

Note that in this context of non-linear scalarization functions, the ESR and SER optimization criteria are distinct. Although not explicitly mentioned here, each method presented in this section optimizes for one of them.

4.2.1 Deterministic non-stationary policies

When the solution policies must be deterministic and the utility function is non-linear, White shows that non-stationary policies can dominate stationary ones [65]. Consequently, it is necessary to consider non-stationary policies to retrieve a PCS in this context.

Imagine an autonomous delivery company working for two large clients A and B. Its goal is to distribute as many items as possible, while avoiding to neglect either A or B as not to lose an important partnership. An autonomous truck receives a reward of $[1, 0]$ when customer A gets a successful delivery, and $[0, 1]$ for customer B. The utility function to use could then be $u(\mathbf{V}^\pi) = \min(V_A^\pi, V_B^\pi)$, effectively maximizing the total number of deliveries while ensuring no client is left out. Here, a deterministic non-stationary policy would be able to yield a satisfying utility while a stationary one would not. Indeed, instead of always acting the same in each state — which would be equivalent to always picking the same client and thus yielding a utility of 0 — the non-stationary policy could condition on the time-dependent past rewards. This allows the agent to make informed decisions about actions to take depending on whether A or B was most chosen until now.

The first and third cells in the second row of table 1 respectively represent the single and multi-policy (PCS) solution sets for deterministic non-stationary policies. Constructing such policies is often done by conditioning them on the current timestep t (EUPG [46], PCN[42][5]), or by splitting $\mathbf{G}$ (see eq. 1) into past (also known as accrued) and future returns (PQL [34], EUPG [46], MO-CAC [43]). For example, the EUPG algorithm employs a modified policy gradient loss including both accrued rewards and a t-conditioned policy. Q-steering [60] takes another approach, forming non-stationary combinations of deterministic stationary base policies. Q-steering is based on Q-learning, and as such is limited to discrete state and action spaces.

4.2.2 Deterministic stationary mixture policies

As previously mentioned, there are contexts in which having a predictable, deterministic policy is essential. Conversely, other applications can tolerate some degree of stochasticity. For example, when designing a fleet of autonomous cars, we might want to add randomness to the path-finding algorithm, such that not all agents converge to the same road, thus avoiding congested traffic and globally sub-optimal behaviors.

When allowed, stochastic policies should be considered as part of the solution, as they can dominate deterministic policies under non-linear utility function [48]. It was shown that in some cases, we can construct a Pareto front from a mixture (i.e., a stochastic combination) of deterministic stationary policies [56, 63]. This is ideal, as it means that recovering a CCS is sufficient to construct the entire PF, greatly reducing the amount of computation needed to find optimal policies.

For example, Vamplew et al. [56] introduce a new algorithm, which we refer to as π-mix, that randomly selects a deterministic policy at the start of each episode and for its entire duration. Although this method works as expected under SER, using one deterministic policy per episode is not suitable for learning under ESR. Following our autonomous delivery example from section 4.2.1, π-mix could learn to alternate between two policies, each favoring only client A or B. In expectation over multiple episodes, this would indeed result in a fair delivery between them. However, on a per-episode basis, one customer would not be supplied, and thus could end the contract.

The ESR case is more complex, as the choice of policy needs to happen at each state (instead of each episode), being effectively equivalent to a stochastic policy. Wakuta [63] introduces a such method in a simplified setting, which we designate as S-rand, where the probability of picking one of k policy is the same at each state.

However, Lu et al. [29] show that finding the correct weights of a stochastic policy to retrieve a specific value vector is in practice infeasible. They propose CAPQL which uses reward augmentation to recover otherwise unreachable value functions from the Pareto front, although the resulting policies are not stochastic.

4.3 Challenges and way forward

As seen throughout this section, the field of multi-objective reinforcement learning, despite its growing popularity, remains sparse and fragmented. The recent work of Hayes et al. [25] identifies a few understudied areas of MORL that require further exploration: *complex multi-objective benchmarks*, dedicated *many-objectives methods*, specificities of *multi-agent settings* and the *dynamical identification and evolution of objectives*.

In particular, the study of many-objectives methods seems like an important future research area for MORL. Indeed, most MORL algorithms suffer from the *curse of dimensionality*, i.e., the exponential growth of the search space in the number of objectives makes retrieving satisfying policies highly complex. Note that the lack of MORL benchmarks has been partly addressed since the survey. Notably, the widely-used RL library *Gymnasium* was extended to the multi-objective case with *MO-Gymnasium* [21].

5 MORL and ethics

While it is important to take into account the normative ethics considerations mentioned in section 3.3, deploying MORL agents in society introduces additional concerns. Drawing from the machine ethics literature and considering potential issues caused by the use of naive MORL algorithms in real life scenarios, we identify four desirable features associated with ethical MORL agents.

They should have the ability to: (a) **prioritize user experience**, (b) **adapt to an evolving society**, (c) **adhere to a set of norms**, and (d) **account for other agents**. Interestingly, the evolution of objectives and the multi-agent aspect are part of the list of open challenges for MORL research mentioned in section 4.3. Note that these properties are pointers for researchers wanting to consider the impact of their algorithms, and not an exhaustive list of required attributes to develop agents with ethically-aligned behaviors. These features can even be contradictory in some cases, e.g., when a user's preferences are incompatible with the set of norms the agent ought to follow.

[5] Pareto Conditioned Networks can be seen as a sort of deterministic non-stationary policy method, as the agent follows a policy trained using supervised learning that conditions on the "desired horizon".

In this section, we define each of the aforementioned properties, review their place in the MORL literature, highlight potential future work, and conclude by discussing ways of benchmarking ethics in a MORL settings. A summarizing classification of existing methods according to our four principles is presented in table 2.

5.1 The user-centric approach

User-centric methods bring an **explicit consideration of the user** alongside the traditional performance goals. These approaches aim to empower users with agency, helping them to make informed decisions while minimizing their cognitive load. Algorithms mentioned in section 4 are capable of producing one policy (or a set of policies) that efficiently solves the input problem. However, most of them do not tackle how to find what utility function to use or which policy to pick from the Pareto front. Consequently, the end-user is tasked with making these decisions which can be non-trivial, for when the Pareto Front is not easily visualizable ($m > 3$). Etzioni and Etzioni [20] advocate for the *ethics bot*, an AI program that "extracts specific ethical preferences from a user and subsequently applies these preferences to the operations of the user's machine". This resonates with the example discussed in sec. 4.1.1 in which we want the agent to learn the passenger's preferences (e.g., prioritize speed if they are in a hurry or low costs if they want to save up) and adapt its driving profile accordingly. Zintgraf et al. [71] noticed this gap in the literature and made a first step to address it by proposing and evaluating several preference elicitation strategies. Following this work, a number of papers have focused on making the human decision maker a bigger part of the MORL process.

With GUTS [47], Roijers et al. introduce an interactive approach for multi-armed bandits, where the agent learns simultaneously about the environment and the user's preferences. Contrary to previous methods, GUTS is able to learn non-linear utility functions, while querying the user a provably limited number of times.

MORAL [40] proposes a two-step method for aligning an agent's behavior with the preferences of a user. First, a set of reward functions is learned from expert demonstrations using adversarial inverse reinforcement learning [22]. The user is then faced with multiple queries, allowing the agent to find a preference vector between expert reward functions, while simultaneously optimizing a policy on this combination. Empirically, the authors show that an adversarial user would not be able to teach the agent behaviors actively avoided by the expert demonstrations, although no formal proof is given. DWPI [30] learns the user's preference vector from demonstrations of their behavior in the environment (in a way reminiscent of inverse RL [70]). Chaput et al. [13] argue for a more contextual and intelligible approach, and propose QSOM-MORL, which learns to identify and solve ethical dilemmas using contextual human preferences.

Although not discussed in this work, it is important to consider potential biases in the construction of the utility function when developing single-policy user-centric algorithms. For example, some work (notably in the economics literature) show that there can be a gap between observed and ground truth preferences [10]. As MORL algorithms get better, this discrepancy may become a bottleneck in user satisfaction, further emphasizing the need to take these factors into account.

5.2 Evolving values and preferences

The methods for learning a user's preferences or utility function introduced in the previous section assume that this target is fixed and not subject to change. However, the owner of a self-driving vehicle, who usually favors comfort and savings over speed, may radically change their preferences in the case of an emergency. Similarly, the vehicle could be part of an autonomous taxi fleet, having to adapt to each customer profile. Therefore, it can be desirable for autonomous agents to have the ability to **detect and adapt to user preference changes**.

A few MORL methods have been developed to tackle this problem. CN [2] and DMCRL [37] take similar approaches, using prior information from learned policies to adapt to changing preferences. Q-steering [60] includes an interactive mode, allowing the user to update the target during or after the learning phase.

As society evolves, the three values proposed in our example of section 2 could fail to address emerging considerations such as environmental impact. Pavaloiu and Koose [39] emphasize that morality is subjective, varies across cultures, and continuously evolves. Thus, we may want our agent to **adapt to newly introduced objectives** while retaining previously learned knowledge. One naive way to approach this aspect could be to use a linear scalarization function, and take advantages of methods which support non-stationary reward functions (e.g., continual RL [27], Q(D)SOM [12]). Hayes et al. [25] identify the challenge of dynamic identification and addition of objectives as one of the main areas for future work in MORL, and to our knowledge the formulation of a variable sized vector-valued reward function has not been studied yet.

5.3 Lawful agents

Approaches for the ethical alignment of agents behavior can be categorized into 3 classes [4]:

- *Bottom-up* approaches do not enforce any obligatory or prohibited actions. Instead, the ethical behavior is learned through experience, and emerges from the definition of the agent and environment.
- *Top-down* approaches are rule-based, and incorporate a priori knowledge (such as deontological duties).
- Some works [52, 17] argue for *hybrid* methods which combine the top-down and bottom-up approaches.

When discussing their ethics bots, Etzioni and Etzioni [20] mention that they only address moral preferences, and disregard normative aspects (e.g., a legal framework). Thus, a MORL-based implementation of an ethics bot would only learn in a bottom-up fashion. Although some works

MORL methods	user-centered	adaptable	normative	multi-agent
CN [2], DMCRL [37], Q-steering [60]	✓	✓		
MAEE [44]			✓	✓
GUTS [47], MORAL [40], DWPI [30], QSOM-MORL [13]	✓			
EE [45], TLO [59]			✓	
MO-MIX [26], PRBS/D [31], moral rewards [53]				✓

Table 2: Qualification of MORL methods with regards to ethical properties.

[64, 53] argues that top-down approaches are challenging and pose some risks, having a set of guarantees (via top-down or hybrid agents) can be crucial in some applications. Typically, we want to ensure that self-driving vehicles deployed on real roads act according to the locally enforced traffic regulations, so that their behavior is safe and predictable for human drivers. In fact, Pagallo [38] argues that values alone are not enough for the coordination of AI agents and that rules are needed. Thus, it is desirable for our agents to **be able to follow a set of norms**.

In MORL, Rodriguez-Soto et al. [45] take the perspective of the environment designer, allowing them to derive theoretical guarantees for the alignment of agents w.r.t. chosen ethical values. To do so, they start from a MOMDP whose reward functions are built upon a value system. Their proposed Multi-Valued Ethical Embedding (EE) algorithm then proceeds to compute a solution weight vector, resulting in a linearly scalarized MDP with the desired properties.

Using potential-based rewards, TLO [59] focuses on impact-minimizing agents, i.e., agents performing a primary task while aiming at disrupting the environment as little as possible. This approach is bottom-up by design, yet the authors demonstrate strong empirical results showing the ethical alignment of trained agents. These results are for now limited to discrete states and actions, although the algorithms proposed are theoretically extensible to the continuous cases.

For single-objective RL, a few works propose top-down or hybrid approaches. Shielding [5] uses temporal logic to enforce a set of properties on the resulting policy. AJAR [3] uses argumentation-based judges to compute the rewards based on a set of moral values. Extending such methods to the multi-objective case presents promising possibilities for future research.

5.4 Ethics as a multi-agent problem

Murukannaiah et al. [36] argue that the study of ethics intrinsically needs to be done in a multi-agent context, highlighting that research in AI ethics is to this day largely constituted of single-agent works and ignores the societal context. As trained MORL algorithms are deployed a real-life situations, they are likely to encounter other actors, both artificial and human. Therefore, we argue that our agents should **be able to account for and interact with other actors**. The field of multi-objective multi-agent reinforcement learning (MOMARL) accounts by design for the interactions that can emerge in these cases. Being at the intersection of two sub-fields, MOMARL remains relatively understudied. Rădulescu et al. [41] have surveyed the field of multi-objective multi-agent decision making and concluded that many gaps still exist in the literature, particularly for RL-based methods. Although some MOMARL approaches have been proposed [26, 31], and there has been work on ethics in the multi-agent setting [15], very few MOMARL papers specifically take an ethical perspective. Rodriguez-Soto et al. [44] propose a method (MAEE) to construct environments in which agents are guaranteed to have an ethically-aligned behavior, while pursuing their individual goals. However, the multi-objective reward function they use is very simple, with only two component: an individual objective and an ethical objective (itself split between a normative and evaluative part). QSOM and QDSOM [12] are multi-agent algorithms based on self-organizing maps. Although not multi-objective, they were tested with various reward functions combining ethical stakes, analogously to ESR-optimized MORL. Tennant et al. [53] analyze the behavior of intrinsically-motivated RL agents rewarded according to moral theories when faced with moral dilemmas.

5.5 Benchmarking ethics

While some papers tackle the evaluation of MORL algorithms and the available benchmarks [57], few environments have become standard, and most of them are too simple for modern methods [25].

When trying to ensure the ethical alignment of an AI agent's behavior, the metric of success may be more complex than a simple sum of reward signals. Few MORL environments with an ethics-first approach have been proposed. The ethical gathering game by Rodriguez-Soto et al. [44] extends the regular gathering game, with the addition of beneficence as a moral value. Scheirlinck et al. [49] introduce the ethical smart grid, a complex environment with continuous actions and observations. They propose to use a number of (sometimes conflicting) moral values from the literature to evaluate the behavior of agents.

Additionally, there is a number of environments which are not created with ethics in mind but allow for the inclusion of one or more of the constraints previously mentioned. As such, any MORL environment (e.g., DST [61]) can be viewed through a user-centric lens by changing the setting or adding queries to a user to learn their preferences. Similarly, we can modify multi-agent multi-objective environ-

ments (e.g., MOBDP [31]) to shift the focus towards the alignment of agents with some specified ethical values.

6 Conclusion

As artificial intelligence agents are being increasingly deployed in society, there is a growing need to study ways of ensuring the ethical alignment of their behaviors. In this paper, we have focused on multi-objective reinforcement learning, a framework that has been deemed ideal for modeling the complexities of both ethics and real-world problems. First, we proposed a classification of existing multi-objective RL methods according to the prevalent taxonomy. Then, we explored the considerations required when one wishes to work in MORL while adopting an ethics-centered perspective. The literature at the intersection of MORL and ethics is still very limited, and a lot of work remains to be done, notably on methods explicitly implementing one or more of the four desirable properties for ethical agents highlighted in section 5: adherence to user preferences, adaptability to societal changes, compliance with norms and regulations, and considerations of other agents. We hope that this work can serve researchers at the intersection of MORL and ethics to visualize the state of current research and the still lacking areas deserving of further investigations.

Acknowledgements

This work was funded by ANR project ACCELER-AI (ANR-22-CE23-0028-01).

References

[1] David Abel, James MacGlashan, and Michael L Littman. Reinforcement learning as a framework for ethical decision making. In *AAAI Workshop: AI, Ethics, and Society*, 2016.

[2] Axel Abels, Diederik Roijers, Tom Lenaerts, Ann Nowé, and Denis Steckelmacher. Dynamic Weights in Multi-Objective Deep Reinforcement Learning. In *ICML*, 2019.

[3] Benoît Alcaraz, Olivier Boissier, Rémy Chaput, and Christopher Leturc. Ajar: An argumentation-based judging agents framework for ethical reinforcement learning. In *AAMAS*, 2023.

[4] Colin Allen, Iva Smit, and Wendell Wallach. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and information technology*, 2005.

[5] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *AAAI*, 2018.

[6] Michael Anderson and Susan Leigh Anderson. Machine ethics: Creating an ethical intelligent agent. *AI magazine*, 2007.

[7] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight Experience Replay. In *NeurIPS*, 2018.

[8] Adrià Puigdomènech Badia, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Zhaohan Daniel Guo, and Charles Blundell. Agent57: Outperforming the atari human benchmark. In *ICML*, 2020.

[9] Toygun Basaklar, Suat Gumussoy, and Umit Y. Ogras. PD-MORL: Preference-Driven Multi-Objective Reinforcement Learning Algorithm. In *ICLR*, 2023.

[10] John Beshears, James J Choi, David Laibson, and Brigitte C Madrian. How are preferences revealed? *Journal of public economics*, 2008.

[11] Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on SMC*, 2008.

[12] Rémy Chaput, Olivier Boissier, and Mathieu Guillermin. Adaptive reinforcement learning of multi-agent ethically-aligned behaviours: the QSOM and QDSOM algorithms. *arXiv e-prints*, 2023.

[13] Rémy Chaput, Laetitia Matignon, and Mathieu Guillermin. Learning to identify and settle dilemmas through contextual user preferences. In *ICTAI*, 2023.

[14] Jianguo Chen, Kenli Li, Keqin Li, Philip S Yu, and Zeng Zeng. Dynamic bicycle dispatching of dockless public bicycle-sharing systems using multi-objective reinforcement learning. *ACM TCPS*, 2021.

[15] Nicolas Cointe, Grégory Bonnet, and Olivier Boissier. Ethical Judgment of Agents' Behaviors in Multi-Agent Systems. In *AAMAS*, 2016.

[16] Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 2022.

[17] Virginia Dignum. *Responsible artificial intelligence: how to develop and use AI in a responsible way*. 2019.

[18] Virginia Dignum, Matteo Baldoni, Cristina Baroglio, Maurizio Caon, Raja Chatila, Louise Dennis, Gonzalo Génova, Galit Haim, Malte S Kließ, Maite Lopez-Sanchez, et al. Ethics by design: Necessity or curse? In *AAAI/ACM Conference on AI, Ethics, and Society*, 2018.

[19] Muhammad Diyan, Bhagya Nathali Silva, and Kijun Han. A multi-objective approach for optimal energy management in smart home using the reinforcement learning. *Sensors*, 2020.

[20] Amitai Etzioni and Oren Etzioni. Incorporating ethics into artificial intelligence. *The Journal of Ethics*, 2017.

[21] Florian Felten, Lucas Nunes Alegre, Ann Nowe, Ana L. C. Bazzan, El Ghazali Talbi, Grégoire Danoy, and Bruno Castro da Silva. A Toolkit for Reliable Benchmarking and Research in Multi-Objective Reinforcement Learning. In *NeurIPS Datasets and Benchmarks Track*, 2023.

[22] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *ICLR*, 2018.

[23] Sven Gronauer and Klaus Diepold. Multi-agent deep reinforcement learning: A survey. *Artificial Intelligence Review*, 2022.

[24] Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, Yaodong Yang, and Alois Knoll. A review of safe reinforcement learning: Methods, theory and applications. *arXiv preprint arXiv:2205.10330*, 2022.

[25] Conor F. Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M. Zintgraf, Richard Dazeley, Fredrik Heintz, Enda Howley, Athirai A. Irissappane, Patrick Mannion, Ann Nowé, Gabriel Ramos, Marcello Restelli, Peter Vamplew, and Diederik M. Roijers. A practical guide to multi-objective reinforcement learning and planning. *AAMAS*, 2022.

[26] Tianmeng Hu, Biao Luo, Chunhua Yang, and Tingwen Huang. MO-MIX: Multi-Objective Multi-Agent Cooperative Decision-Making With Deep Reinforcement Learning. *IEEE PAMI*, 2023.

[27] Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *JAIR*, 2022.

[28] B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[29] Haoye Lu, Daniel Herman, and Yaoliang Yu. Multi-Objective Reinforcement Learning: Convexity, Stationarity and Pareto Optimality. In *ICLR*, 2022.

[30] Junlin Lu, Patrick Mannion, and Karl Mason. Inferring Preferences from Demonstrations in Multi-objective Reinforcement Learning: A Dynamic Weight-based Approach. In *ALA (AAMAS)*, 2023.

[31] Patrick Mannion, Sam Devlin, Jim Duggan, and Enda Howley. Reward shaping for knowledge-based multi-objective multi-agent reinforcement learning. *The Knowledge Engineering Review*, 2018.

[32] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *NIPS*, 2013.

[33] Thomas M Moerland, Joost Broekens, Aske Plaat, Catholijn M Jonker, et al. Model-based reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 2023.

[34] Kristof Van Moffaert and Ann Nowé. Multi-Objective Reinforcement Learning using Sets of Pareto Dominating Policies. *JMLR*, 2014.

[35] James H Moor. The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems*, 2006.

[36] Pradeep K Murukannaiah, Nirav Ajmeri, Catholijn M Jonker, and Munindar P Singh. New Foundations of Ethical Multiagent Systems. In *AAMAS*, 2020.

[37] Sriraam Natarajan and Prasad Tadepalli. Dynamic preferences in multi-criteria reinforcement learning. In *ICML*, 2005.

[38] Ugo Pagallo et al. Even angels need the rules: Ai, roboethics, and the law. In *ECAI*, 2016.

[39] Alice Pavaloiu and Utku Kose. Ethical artificial intelligence-an open question. *JOMUDE*, 2017.

[40] Markus Peschl, Arkady Zgonnikov, Frans A Oliehoek, and Luciano C Siebert. Moral: Aligning ai with human norms through multi-objective reinforced active learning. In *AAMAS*, 2022.

[41] Roxana Rădulescu, Patrick Mannion, Diederik M. Roijers, and Ann Nowé. Multi-objective multi-agent decision making: A utility-based analysis and survey. In *AAMAS*, 2020.

[42] Mathieu Reymond, Eugenio Bargiacchi, and Ann Nowé. Pareto Conditioned Networks. In *AAMAS*, 2022.

[43] Mathieu Reymond, Conor F. Hayes, Denis Steckelmacher, Diederik M. Roijers, and Ann Nowé. Actor-critic multi-objective reinforcement learning for non-linear utility functions. In *AAMAS*, 2023.

[44] Manel Rodriguez-Soto, Maite Lopez-Sanchez, and Juan A. Rodriguez-Aguilar. Multi-objective reinforcement learning for designing ethical multi-agent environments. *Neural Computing and Applications*, 2023.

[45] Manel Rodriguez-Soto, Roxana Rădulescu, Juan A Rodriguez-Aguilar, and Maite Lopez-Sanchez. Multi-objective reinforcement learning for guaranteeing alignment with multiple values. In *ALA (AAMAS)*, 2023.

[46] Diederik Roijers, Denis Steckelmacher, and Ann Nowe. Multi-objective Reinforcement Learning for the Expected Utility of the Return. In *ALA (AAMAS)*, 2018.

[47] Diederik M. Roijers, Luisa M. Zintgraf, Pieter Libin, and Ann Nowé. Interactive multi-objective reinforcement learning in multi-armed bandits for any utility function. In *ALA (AAMAS)*, 2020.

[48] Diederik Marijn Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A Survey of Multi-Objective Sequential Decision-Making. *JAIR*, 2013.

[49] Clément Scheirlinck, Rémy Chaput, and Salima Hassas. Ethical Smart Grid: A Gym environment for learning ethical behaviours. *JOSS*, 2023.

[50] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 2016.

[51] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, 1999.

[52] Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. Learning machine morality through experience and interaction. 2023.

[53] Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. Modeling moral choices in social dilemmas with multi-agent reinforcement learning. In *IJCAI*, 2023.

[54] Mark Timmons. *Moral theory: An introduction*. Rowman & Littlefield publishers, 2012.

[55] Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. Implementations in Machine Ethics: A Survey. *ACM Computing Surveys*, 2021.

[56] Peter Vamplew, Richard Dazeley, Ewan Barker, and Andrei Kelarev. Constructing Stochastic Mixture Policies for Episodic Multiobjective Reinforcement Learning Tasks. In *Advances in Artificial Intelligence*. 2009.

[57] Peter Vamplew, Richard Dazeley, Adam Berry, Rustam Issabekov, and Evan Dekker. Empirical evaluation methods for multiobjective reinforcement learning algorithms. *Machine Learning*, 2011.

[58] Peter Vamplew, Richard Dazeley, Cameron Foale, Sally Firmin, and Jane Mummery. Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology*, 2018.

[59] Peter Vamplew, Cameron Foale, Richard Dazeley, and Adam Bignold. Potential-based multiobjective reinforcement learning approaches to low-impact agents for ai safety. *Engineering Applications of Artificial Intelligence*, 2021.

[60] Peter Vamplew, Rustam Issabekov, Richard Dazeley, Cameron Foale, Adam Berry, Tim Moore, and Douglas Creighton. Steering approaches to pareto-optimal multiobjective reinforcement learning. *Neurocomputing*, 2017.

[61] Peter Vamplew, John Yearwood, Richard Dazeley, and Adam Berry. On the limitations of scalarisation for multi-objective reinforcement learning of pareto fronts. In *Advances in Artificial Intelligence*, 2008.

[62] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *AAAI*, 2016.

[63] Kazuyoshi Wakuta. A note on the structure of value spaces in vector-valued Markov decision processes. *Mathematical Methods of Operations Research*, 1999.

[64] Wendell Wallach and Colin Allen. *Moral machines: Teaching robots right from wrong*. Oxford University Press, 2008.

[65] D. J White. Multi-objective infinite-horizon discounted Markov decision processes. *Journal of Mathematical Analysis and Applications*, 1982.

[66] Jess Whittlestone, Kai Arulkumaran, and Matthew Crosby. The societal implications of deep reinforcement learning. *JAIR*, 2021.

[67] Jie Xu, Yunsheng Tian, Pingchuan Ma, Daniela Rus, Shinjiro Sueda, and Wojciech Matusik. Prediction-Guided Multi-Objective Reinforcement Learning for Continuous Robot Control. *ICML*, 2020.

[68] Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. A Generalized Algorithm for Multi-Objective Reinforcement Learning and Policy Adaptation. In *NeurIPS*, 2019.

[69] Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM CSUR*, 2021.

[70] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *AAAI*, 2008.

[71] Luisa M Zintgraf, Diederik M Roijers, Sjoerd Linders, Catholijn M Jonker, and Ann Nowé. Ordered preference elicitation strategies for supporting multi-objective decision making. *AAMAS*, 2018.

Grands modèles de langage (*large language models*) et règles logiques pour une automatisation décisionnelle avancée

Pierre Feillet

IBM France Lab

feillet@fr.ibm.com

Résumé

Les grands modèles de langage (LLM, pour large language models*) impactent tous les secteurs d'activité en promettant une nouvelle expansion de l'automatisation dans tous les métiers. L'enjeu pour les organisations est de bénéficier des gains de productivité annoncés par l'IA générative tout en sécurisant le raisonnement algorithmique dans la prise de leurs décisions critiques.*

Il est opportun d'évaluer les qualités des LLMs et des moteurs d'inférence de règles logiques ainsi que leur hybridation pour préparer une meilleure automatisation des décisions.

Mots-clés

IA hybride, règles logiques, grands modèles de langage, IA générative, extraction de connaissances.

Abstract

Large language models (LLMs) are impacting all sectors of activity by promising a new expansion for automation. The challenge for organizations is to benefit from the productivity gains announced by generative AI while securing algorithmic reasoning in critical decision-making.

It is timely to evaluate the qualities of LLMs and rule inference engines and prepare their hybridization for improved decision- making automation.

Keywords

Hybrid AI, rules, logic, large language models, generative AI, knowledge extraction.

1 Introduction

L'intelligence artificielle générative disrupte le monde de l'IA en ouvrant de nouvelles perspectives d'interactions Humain Machine et repousse les limites des tests de Turing.

Étant donnés les progrès récents une question se pose : les grands modèles de langage (LLM) seuls peuvent-ils suffire à automatiser les décisions critiques dans nos organisations ? Sont-ils compétents pour raisonner avec confiance afin de prendre une décision à fort impact pour les entreprises et citoyens ?

2 Les grands modèles de langage

2.1 Définition générale

En résumé, un grand modèle de langage compresse les informations lues à partir d'un corpus de textes d'entraînement pour générer de nouveaux textes à partir d'une requête donnée. Basé sur une architecture de réseau de neurones, son comportement est statistique. Il prend une séquence de *tokens* (groupes de caractères) exprimés dans la requête et produit une autre séquence de *tokens* qui sont les plus probables en regard de sa base documentaire d'entraînement. Ce n'est pas dans sa construction un algorithme de raisonnement ; il ne repose pas sur des mécanismes logiques et peut être vu comme un perroquet stochastique [10].

Les modèles de langage à grande échelle (LLM) se sont multipliés depuis 2022, disponible en source fermé ou ouvert et disponible en *Model as a Service* ou localement [7].

2.2 Apprentissage par renforcement à partir de rétroaction humaine

Au-delà de leur apprentissage auto-supervisé utilisant des corpus textuels, certains LLM bénéficient d'un apprentissage par renforcement par rétroaction humaine (*reinforcement learning from human feedback*). Cette approche d'apprentissage par renforcement repose sur les commentaires et évaluations humaines pour guider l'apprentissage du modèle sur la base d'une évaluation humaine de ses résultats.

2.3 Chaîne de pensée

Le *Chain-of-thought prompting* (en français, requête par chaîne de pensée) est une méthode qui vise à améliorer les capacités de raisonnement des grands modèles de langage. Cette technique fonctionne en décomposant la résolution d'un problème en modélisant les étapes successives d'un processus de pensée.

Cette approche popularisée par *LangChain* [6] permet au modèle non seulement d'arriver à des réponses plus précises, mais également de développer une méthode de résolution de problèmes plus structurée et explicative, rendant les réponses générées plus compréhensibles pour les utilisateurs. Cette décomposition garde néanmoins des limites en termes de raisonnement.

2.4 *Retrieval-Augmented Generation*

Le RAG pour *Retrieval-Augmented Generation*, est une technique conçue pour améliorer les réponses du modèle en intégrant des informations extraites de documents externes. Ce processus se décompose en deux étapes principales :

— récupération (*Retrieval*) : dans cette première étape, le modèle utilise une requête (généralement basée sur la question ou la tâche posée par l'utilisateur) pour chercher et récupérer des informations pertinentes à partir d'une base de données de documents. Cette base de données peut être constituée de textes provenant de diverses sources telles que des livres, des articles ou des sites internet.

— génération (*Augmented Generation*) : ensuite, le modèle utilise les informations récupérées comme contexte supplémentaire pour générer une réponse. Ce faisant, le modèle peut fournir des réponses plus précises, informatives et basées sur des données concrètes, plutôt que de se baser uniquement sur ce qu'il a appris durant son entraînement.

L'approche RAG est particulièrement utile pour les tâches qui nécessitent des réponses détaillées ou des informations spécifiques qui ne sont pas nécessairement contenues dans le corpus d'entrainement du modèle. Elle permet aux modèles de langage d'enrichir les requêtes afin de rendre la génération plus performante. Toutefois elle n'influe pas sur les qualités intrinsèques de raisonnement du LLM.

3 Les moteurs d'inférence de règles

Depuis plusieurs décennies, les entreprises s'appuient sur des solutions logicielles pour automatiser leurs décisions critiques dans le but de déterminer des éligibilités règlementaires, approuver des prêts et souscrire à des offres de services. Au- delà des applications encodant en Java, Javascript ou autre langage, la logique métier, les systèmes de gestion de règles ont été adoptés par les industriels, notamment par les services financiers. Ces systèmes permettent le développement, le test, la simulation, le déploiement et la maintenance de politiques métier en s'appuyant sur des règles logiques. Cette approche repose sur la capture de connaissances et leur formalisation. Les experts métier définissent une ontologie avec des prédicats logiques et des structures de données représentant leur domaine. Ces règles, communément formulées sous forme d'instructions `Si <conditions> Alors <Actions>`, d'arbres ou tables de décision, sont évalués par un moteur d'inférence causale capable d'instancier et de chaîner l'exécution dans un contexte de données structurées, automatisant ainsi un chemin de raisonnement.

Cette technologie a mûri au fil des ans, permettant de capturer des modèles de décision complexes avec des milliers de règles et de tables de décision. Des produits comme IBM *Operational Decision Manager* (ODM) [3] ont prouvé leur capacité à capturer des politiques complexes et exécuter plus d'un milliard de décisions quotidiennement tout en prenant quelques millisecondes pour réaliser chaque décision sur une CPU.

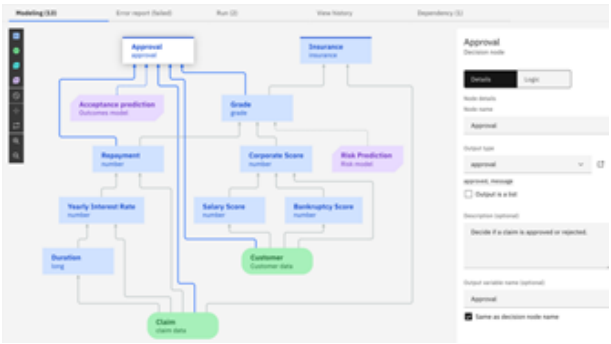


FIGURE 1 – Un modèle de décision défini comme un graphe acyclique dirigé dans IBM ADS.

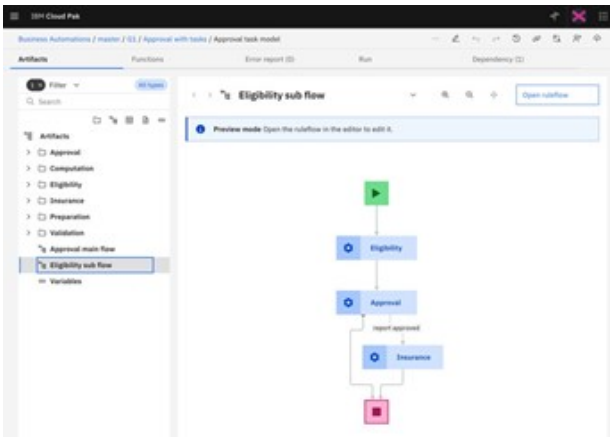


FIGURE 2 – Un micro-flow appelé *Ruleflow* pouvant contenir un cycle dans IBM ADS.

Les moteurs de règles comme ceux d'IBM ODM et ADS permettent de prescrire un micro-flot de raisonnement, afin par exemple de procéder à la validation d'un dossier d'emprunt, puis au calcul d'éligibilité, et enfin de calculer les modalités de remboursements et d'assurance en fonction du profil, de la politique en cours et des taux bancaires. Ces micro-flots peuvent être des DAG (*Direct Acyclic Graph*) ou bien autoriser des cycles dans des flots de regles (*ruleflows*). Ils organisent les étapes de raisonnements, et évaluent dans chacune des taches rencontrées un ensemble de règles.

4 Les exigences d'automatisation décisionnelle

4.1 Quelles sont les exigences clés pour la prise de décision en entreprise ?

L'automatisation des décisions en entreprise implique l'utilisation de la technologie et des logiciels pour systématiser les processus de prise de décision au sein d'une organisation. Les critères clés pour une automatisation réussie des décisions en entreprise sont les suivants :

— exactitude et fiabilité : le système doit être précis dans le traitement des données et la prise de déci-

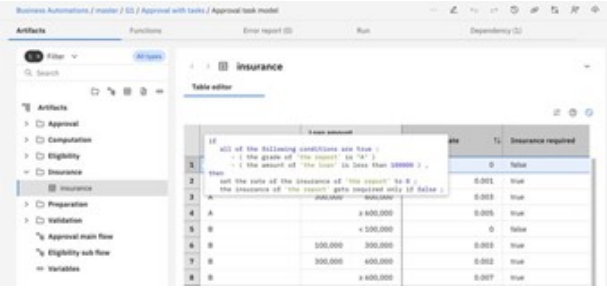


FIGURE 3 – Une table de décision regroupant des règles dans IBM ADS

sions pour assurer la confiance des utilisateurs et des parties prenantes ;
— évolutivité : la solution doit être capable de gérer de grands volumes de données et de demandes de prise de décision sans dégradation significative des performances ;
— flexibilité et adaptabilité : les environnements économiques et réglementaires sont dynamiques, et les exigences en matière de prise de décision peuvent changer. Le système doit être suffisamment flexible pour s'adapter à de nouvelles règles commerciales, politiques et réglementations sans nécessiter de coûts majeurs de développement ;
— prise de décision en temps réel : dans certains scénarios, des capacités de prise de décision en temps réel ou quasi-temps réel sont cruciales. Le système doit traiter les données et fournir des décisions dans des délais acceptables ;
— transparence et auditabilité : les décisions d'entreprise impactent souvent des processus critiques, et les parties prenantes ont besoin de comprendre comment les décisions sont prises. Le système doit fournir des explications claires pour les décisions, et il doit être auditable à des fins de conformité réglementaire ;
— sécurité et confidentialité des données : puisque l'automatisation des décisions traite de données sensibles, la sécurité et la confidentialité des données sont primordiales. Le système doit employer des mesures de sécurité robustes pour protéger les données contre des accès ou manipulations non autorisés ;
— observations des performances : le système doit disposer de capacités de mesures et de rapport pour suivre les performances des processus de prise de décision et identifier les domaines à améliorer ;
— maîtrise des coûts : la considération du coût du système par rapport à ses avantages est essentielle. La solution doit offrir un bon retour sur investissement et s'aligner sur les contraintes budgétaires de l'organisation.

Dans l'ensemble, une solution réussie d'automatisation des décisions en entreprise doit s'aligner sur les objectifs de l'organisation, rationaliser les processus de prise de décision et contribuer à une efficacité et une productivité accrue.

4.2 Les LLMs atteignent-ils seuls toutes ces exigences ?

Bien que certains LLMs montrent des résultats impressionnants et certaines capacités de raisonnement, ils échouent tout aussi facilement lors de la répétition de l'expérience, ou lors de légères modifications dans le requêtage. Vous pouvez expérimenter ce comportement à double face avec le chatbot de commande de pizza présenté dans le tutoriel OpenAI de DeepLearning.ai [1]. Selon les essais, le chatbot fournit le résultat attendu ou un résultat surprenant, même avec seulement un léger changement de requêtes.
Un autre défi avec les LLM est leur limite maximale de *tokens*, qui restreint la quantité de contexte qu'ils peuvent gérer. Une technique pour résoudre ce problème est d'appliquer une approche *Retrieve Augment Generate* [9] ou d'étendre l'apprentissage du LLM par *fine-tuning* sur un corpus complémentaire.
Par ailleurs, l'apprentissage par renforcement avec retour humain appliqué aux LLM pour en atténuer leurs réponses purement statistiques peut s'avérer incomplet voir contre-productif dans certaines tâches [8].

4.3 Pouvons-nous les utiliser en combinaison avec des moteurs de décision basés sur des règles ?

Les moteurs d'inférence de règles permettent l'automatisation de prises de décisions. Ils reposent sur la capture d'un savoir-faire métier dans une ontologie et la spécification non ambiguë de la logique formalisant l'expertise sur des données structurées. Cela garantit une prise de décision déterministe, robuste et transparente en échange d'un effort de capture et de formalisation de la connaissance.
Le défi réside dans la fusion de ces technologies pour capitaliser sur leurs points forts, à l'instar de la création de matériaux composites qui surpassent les propriétés individuelles de chaque élément.
Nous proposons différentes méthodes combinant les LLM avec des moteurs de règles.

5 Hybridations LLM et règles logiques

5.1 Compréhension du langage naturel suivie par un raisonnement basé sur des règles

Dans cette approche, nous utilisons d'abord un modèle de langage basé sur l'apprentissage automatique (LLM) pour comprendre le texte brut et en extraire des données structurées. Ensuite, nous passons ces données structurées extraites à un moteur de règles logiques pour raisonner de manière déterministe sur ces données, potentiellement combinées avec des informations supplémentaires provenant de référentiels d'entreprise.

Avantages. L'intégration séquentielle d'un LLM dédie à la compréhension du langage naturel afin d'en extraire

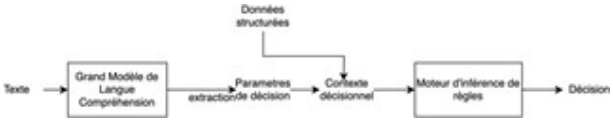


FIGURE 4 – Compréhension du langage naturel suivie par un raisonnement basé sur des règles logiques

des entités structurées suivie par l'exécution d'un moteur de règles est simple à mettre en place. Le pipeline LLM-Moteur de règles implique de transmettre les résultats du LLM comme paramètres d'entrée pour alimenter le contexte du moteur de règles.

Inconvénients. La performance de l'ensemble dépend de l'aptitude du LLM à extraire correctement les données structurées. Pour se protéger des cas ou celles-ci s'avèrent incomplètes ou mal formatées, il est crucial de mettre en place des garde-fous avec des vérifications, des valeurs par défaut ou des heuristiques de repli pour alimenter la phase de raisonnement. L'extraction du contexte décisionnel doit se révélée robuste via une ou plusieurs appels au LLM, et éventuellement par une chaîne de pensée. Elle peut s'appuyer sur des interactions utilisateur dans le cadre d'une expérience conversationnelle.

5.2 Raisonnement basé sur des règles suivi par la génération de langage naturel avec un LLM

Dans cette intégration, un moteur de règles prend d'abord une décision basée sur des données structurées. La décision structurée est ensuite transmise à un modèle de langage qui en fait déduit une transformation dans un texte de langue comme un résume ou un courriel.
Cette hybridation est particulièrement utile pour fournir des explications, justifications, résumer les décisions d'entreprise et les communiquer. Le contenu généré peut être adapté en fonction de du profil du destinataire et la communication visée.

FIGURE 5 – Raisonnement basé sur des règles suivi par une génération de texte en langage naturel avec un LLM

Avantages. Cette approche offre l'avantage de tirer parti des puissantes capacités de génération de langage naturel (NLG) fournies par le LLM pour créer des textes bien formulés pour communiquer des décisions automatisées.
Elle peut être facilement mise en œuvre facilement via une ingénierie de requêtage *prompt engineering* en passant la décision structurée au LLM.

Inconvénients. Alors que les règles se concentrent sur le raisonnement avec des données structurées, le LLM se concentre sur les tâches de traitement du langage naturel (NLP). Cette séparation nécessite une mise au point minutieuse pour gérer les variations de génération tant dans leurs formats que dans leurs contenus. Des tests automatisés du résultat de génération de langage naturel (NLG) sont nécessaire pour mesurer la performance de cette hybridation ainsi que la couverture des erreurs et variations.

5.3 Raisonnement basé sur des règles pilotant le traitement du langage naturel avec le LLM

Le moteur d'inférence de règles agit ici comme le moteur principal, invoquant le modèle de langage à la demande. Le moteur de règles pilote l'évaluation logique et appelle dynamiquement le LLM pour déléguer deux tâches :

- traitement du texte reçu dans son contexte de décision pour la compréhension (NLU) et l'extraction de données structurées ;
- générer du texte (NLG) pour produire, par exemple, un résumé de la décision automatisée en texte brut.

Cette intégration se poursuit dans la continuité de l'appel de tout modèle d'apprentissage automatique depuis IBM ODM [3] ou IBM ADS [2, 11], pour considérer les probabilités de risque ou d'opportunité lors d'une prise de décision. De même, le LLM peut être appelé à partir d'une règle, soit à distance, soit localement, en fonction de son facteur de forme.
Il est possible d'envisager cette hybridation comme une implémentation de chaînes de pensée par un moteur logique, en considérant que le LLM est appelé à chaque étape ou le traitement du langage intervient.

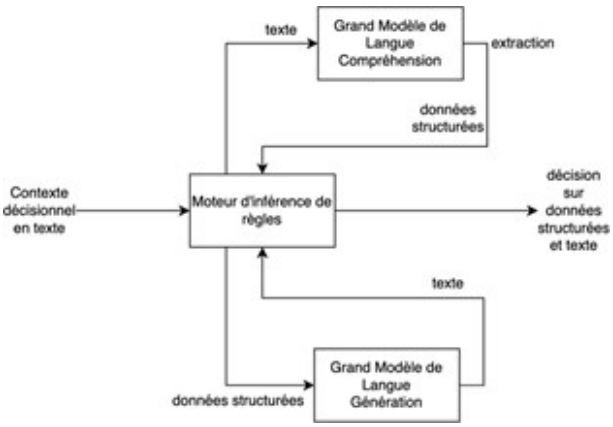


FIGURE 6 – Raisonnement basé sur des règles pilotant le traitement du langage naturel avec le LLM

Avantages. Le contrôle est assuré par le moteur de règles permet d'orchestrer les étapes de raisonnement, et procéder par déduction. Les LLM sont appelés à la demande en fonction du chemin de raisonnement pour réaliser les tâches de NLP.

Inconvénients. Les moteurs de raisonnement et LLM sont étroitement couplés, nécessitant une intégration fine. Les données structurées utilisées dans la prise de décision doivent s'aligner avec les tâches de NLP. Des garde-fous

appropriés sont nécessaires pour gérer la frontière entre données structurées et non structurées pour garantir la qualité et fiabilité de l'hybridation.

5.4 Extraire des règles métier depuis du texte avec un LLM pour les exécuter dans un moteur logique

Cette approche utilise un modèle de langage pour extraire des éléments d'automatisation, y compris des règles logiques, des modèles de données et des signatures fonctionnelles, à partir de politiques d'entrepris exprimées en texte brut. Ces éléments extraits sont ensuite utilisés pour générer un projet d'automatisation dans un moteur de règles comme ceux d'IBM ADS ou ODM. Cette approche a déjà été prototypée avec succès avec IBM ADS en passant par la génération d'un descripteur pivot en JSON.

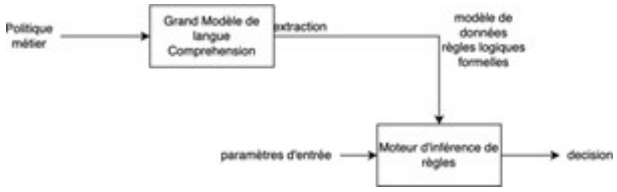


FIGURE 7 – Extraire des règles logiques du texte avec un LLM et exécuter ces règles dans un moteur d'inférence

Avantages. Tirer parti du LLM comme outil d'extraction de connaissances permet d'extraire des règles et une ontologie de données sous-jacente. Les règles extraites, une fois révisées par un humain, permettent d'automatiser des décisions avec traçabilité et déterminisme. C'est un chemin prometteur pour assister l'extraction de connaissances expertes et capitaliser sur des moteurs d'inférence pour l'exécution des décisions.

Inconvénients. Cette approche nécessite une chaîne de requêtes efficaces, éventuellement un modèle réglé finement pour réaliser des extractions pertinentes de la logique, pour tout domaine d'expertise et formulation de la politique métier. Elle nécessite également des compétences et outils pour extraire efficacement la connaissance à automatiser, et en assurer la maintenance lorsque les documents sources évoluent ou sur la base de retours opérationnels.

5.5 Règles pour apporter un raisonnement fiable dans un agent conversationnel

Un grand modèle de langage est ici utilisé pour piloter l'expérience conversationnelle, gérant les tâches de traitement du langage naturel. Le LLM délègue à un moteur de décision basé sur des règles pour appliquer des décisions logiques.
Cette intégration exige que le chatbot reconnaisse, pendant la conversation, quand déclencher un service de décision. Le chatbot guide le dialogue pour fournir le contexte et invoque le moteur de décision basé sur des règles lorsque tous les paramètres d'entrée sont définis. Le moteur de décision renvoie des paramètres de sortie, qui sont ensuite restitués dans la conversation par la génération de langage naturel (NLG).
IBM travaille activement à incorporer ce schéma pour apporter des capacités de prise de décision à Watson Orchestrate [5]. Par ailleurs les clients peuvent déjà développer des outils dans des solutions open-source comme LangChain pour invoquer des décisions basées sur des règles à partir d'un chatbot [4].

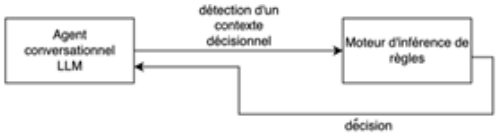


FIGURE 8 – Décision à base de règles logiques injectée pour raisonner dans un agent conversationnel

Avantages. Cette approche permet de bénéficier d'une expérience utilisateur conversationnelle tout en déléguant le raisonnement à des moteurs déterministes spécialisés.

Inconvénients. La mise en œuvre de cette approche nécessite une détection performante des décisions à déléguer, de synchroniser un contexte entre les 2 moteurs, et de gérer les cas d'erreur à la frontière entre les domaines de données structurées et non structurées. Les défis peuvent inclure le traitement de différents formats de données et des informations de contexte incomplètes.

6 Conclusion

Nous avons exploré le pouvoir de transformation des modèles de langage dans l'automatisation des décisions d'entreprise. Les LLM offrent des capacités impressionnantes dans le traitement du langage. En étant combinés à un apprentissage par renforcement à partir de rétroaction humaine, à des chaines de pensées ou du RAG ils améliorent leurs performances.

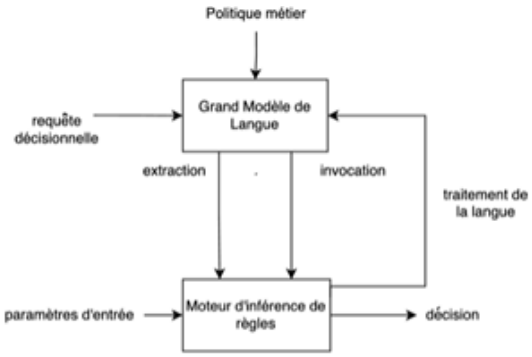


FIGURE 9 – Récapitulatif des combinaisons LLM et règles logiques

Ils manquent néanmoins de compétences pour un raisonnement fiable et répétable afin de répondre aux exigences

strictes associées à la prise de décision critique. Pour combler cette lacune, nous avons introduit cinq approches d'hybridation des LLM avec des moteurs de raisonnement basés sur des règles logiques.
Ces motifs d'intégration incluent l'invocation de LLM avant ou pendant l'exécution des règles, pour le traitement du texte brut en complément des données structurées, des règles pilotant le traitement du texte, l'extraction par LLM de règles logiques à partir de documents, et l'utilisation de services de décision basés sur des règles dans des agents conversationnels.
En combinant l'IA générative et symbolique, ces hybridations d'IA visent à promouvoir de nouveaux usages et démocratiser les solutions d'automatisation de prise de décision tout en satisfaisant les enjeux règlementaires et éthiques des entreprises.

Références

[1] Deep Learning OpenAI courses. https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/.

[2] IBM ADS. https://-www.ibm.com/products/automation- decision-services.

[3] IBM Operational Decision manager. https://www.ibm.com/products/operational-decision-manager.

[4] IBM ODM with LangChain. https://community.ibm.com/community/user/automation/blogs/laurent-grateau1/2023/06/09/integrating-o-dm-with-large- language-model.

[5] IBM watsonx Orchestrate. https://www.ibm.com/-products/watson-orchestrate.

[6] LangChain. https://arxiv.org/html/2402.06196v1.

[7] Large Language Models : A survey. https://arxiv.-org/html/2402.06196v1.

[8] Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. https://arxiv.org/abs/2307.15217.

[9] Retrieval-Augmented Generation for Knowledge- Intensive NLP Tasks. https://arxiv.org/abs/2005.11401.

[10] Stochastic parrot. https://en.wikipedia.org/wiki/Stochastic_parrot.

[11] Two ways of integrating Machine Learning in IBM ADS. https://www.youtube.com/watch?v=7ZJqt5b-exS8.

Enrichissement de fonctions de perte avec contraintes de domaine et co-domaine pour la prédiction de liens dans les graphes de connaissance

Nicolas Hubert[1,2], Pierre Monnin[3], Armelle Brun[2], Davy Monticolo[1]

[1] Université de Lorraine, ERPI, Nancy, France
[2] Université de Lorraine, CNRS, LORIA, Nancy, France
[3] Université Côte d'Azur, Inria, CNRS, I3S, Sophia-Antipolis, France

{nicolas.hubert, davy.monticolo}@univ-lorraine.fr, armelle.brun@loria.fr, pierre.monnin@inria.fr

Résumé

Les modèles à base d'embeddings pour la prédiction de liens dans les graphes de connaissance sont entrainés avec des fonctions de perte. Les approches traditionnelles considèrent que l'étiquette d'un triplet est soit vraie, soit fausse. Nous affirmons que les triplets négatifs qui sont sémantiquement valides au regard du profil de la relation devraient être traités différemment de ceux sémantiquement invalides. Nous proposons des fonctions de perte guidées par la sémantique. La généralité et la supériorité de notre approche sont clairement établies sur trois jeux de données publics.

Mots-clés

Graphe de connaissance, prédiction de liens, modèle à base d'embeddings, schémas.

Abstract

Knowledge graph embedding models used for link prediction are trained with loss functions w.r.t. a batch of labeled triples. Traditional approaches consider the label of a triple to be either true or false. We posit that negative triples that are semantically valid regarding relation's domain and range should be treated differently from semantically invalid ones. We then propose semantic-driven versions for the three main loss functions for link prediction. The generality and superiority of our approach is clearly demonstrated on three public benchmarks.

Keywords

Knowledge graph, link prediction, embedding models, schemas.

1 Introduction

Cet article a été accepté à ESWC 2024 [1].
Un graphe de connaissance (GC) est une collection de triplets (s, p, o) où le sujet s et l'objet o sont deux entités du graphe, et le prédicat p qualifie la nature de la relation entre ces deux entités. Les GCs sont intrinsèquement incomplets et une des principales tâches est de prédire les liens manquants [1].
La tâche de prédiction de liens (PL) est souvent abordée à l'aide de modèles à base d'embeddings qui représentent les entités et les relations sous forme de vecteurs.
Ces modèles sont entrainés avec des fonctions de perte visant à maximiser les scores assignés aux triplets positifs (présents dans le GC) et à minimiser ceux des triplets négatifs (non présents). Un pan de la littérature étudie l'influence des triplets négatifs [3]. En effet, leur génération (*negative sampling*) se fait en général en remplaçant le sujet ou l'objet d'un triplet positif par une autre entité du GC. Des travaux récents démontrent l'intérêt d'utiliser les informations extraites d'un schéma (ou ontologie) pour générer des triplets négatifs de meilleure qualité, notamment des triplets sémantiquement valides au regard du domaine et co-domaine d'un prédicat [4]. Cependant, ces travaux consistent à contraindre les triplets négatifs lors de l'échantillonnage. La possibilité d'échantillonner des triplets négatifs de tout type (sémantiquement valides ou non) et de les considérer différemment dans la fonction de perte n'a, à notre connaissance, jamais été étudiée.
Dans ce travail, nous proposons des fonctions de perte guidées par la sémantique. Ces fonctions tirent parti de la connaissance des domaines et co-domaines des prédicats extraits d'un schéma. La supériorité de notre approche par rapport aux approches traditionnelles est validée expérimentalement sur plusieurs modèles et de jeux de données.

2 Approche

Les fonctions de perte que nous proposons visent à distinguer les négatifs sémantiquement valides de ceux qui ne le sont pas. Les premiers sont définis comme respectant à la fois le domaine et co-domaine du prédicat, tandis que les seconds violent a minima le domaine ou co-domaine. Le domaine (resp. co-domaine) d'un prédicat est le type d'entités attendu comme sujet (resp. objet). Par exemple, le prédicat `président` attend une `Personne` comme sujet et un `Pays` comme objet.
Nous proposons des versions guidées par la sémantique pour les trois principales fonctions de perte utilisées dans la littérature [5] : la *pairwise hinge loss*, la *1-N binary cross-*

entropy loss et la *pointwise logistic loss*.
Dans ce qui suit, nous étayons le passage de la *pairwise hinge loss* (PHL) telle que définie traditionnellement, à sa version sémantique. La PHL est définie ci-dessous :

$$\mathcal{L}_{PHL} = \sum_{t \in \mathcal{T}^+} \sum_{t' \in \mathcal{T}^-} [\gamma + f(t') - f(t)]_+ \quad (1)$$

où $\mathcal{T}$, f, et $[x]_+$ représentent respectivement un ensemble de triplets, la fonction de score, et la partie positive de x. $\mathcal{T}$ est ensuite séparé en un ensemble de triplets positifs $\mathcal{T}^+$ et un ensemble de triplets négatifs $\mathcal{T}^-$. γ est un hyperparamètre de marge ajustable et qui spécifie à quel point les scores assignés aux triplets positifs doivent être distants des scores assignés aux triplets négatifs correspondants.
La version sémantique de la PHL est alors définie comme suit :

$$\mathcal{L}^S_{PHL} = \sum_{t \in \mathcal{T}^+} \sum_{t' \in \mathcal{T}^-} [\gamma \cdot \ell(t') + f(t') - f(t)]_+$$

$$\text{où } \ell(t') = \begin{cases} 1 \text{ si } t' \text{ est sémantiquement invalide} \\ \epsilon \text{ sinon} \end{cases} \quad (2)$$

La fonction de perte dans l'Équation (2) comporte désormais un exposant *S* pour clarifier qu'il s'agit de la version sémantique. Un choix de $\epsilon < 1$ conduit le modèle à appliquer une marge plus élevée entre les scores des triplets positifs et des triplets sémantiquement invalides qu'entre les triplets positifs et les triplets sémantiquement valides. Pour un triplet positif donné, cela permet de maintenir les scores de ses contreparties négatives sémantiquement valides relativement plus proches par rapport aux scores de ses contreparties sémantiquement invalides. Intuitivement, lorsque le modèle produit des prédictions erronées, un plus grand nombre d'entre elles sont néanmoins censées respecter les contraintes de domaine et de co-domaine imposées par les relations. Ainsi, les prédictions erronées sont supposés être sémantiquement plus proches du triplet positif.
Ce guidage sémantique est permis par l'introduction d'un terme ϵ, que nous appellons colloquialement le *facteur sémantique* et qui a pour but de rapprocher le score des négatifs sémantiquement valides de ceux des triplets positifs. Il est important de souligner le fait que ce facteur sémantique s'insère dans les trois fonctions de perte mentionnées ci-haut (et non seulement la PHL comme détaillé ci-dessus). Ceci démontre la généralité de notre approche, qui peut être étendue à d'autres fonctions de perte. Les définitions des fonctions de perte traditionnelles et de celles guidées par la sémantique sont détaillées dans l'article original [1].
Enfin, il convient de rappeler que notre apprche ne contraint aucunement le processus de génération de triplets négatifs. Au lieu de cela, notre approche distribue dynamiquement les triplets négatifs à différentes parties d'une même fonction de perte. Cela conduit à un traitement différencié selon la nature des triplets négatifs, pour un coût computationnel moindre que les approches les plus sophistiquées de *negative sampling* [1].

3 Résultats

Les jeux de données, expériences et résultats sont détaillés dans l'article original [1]. Le code source est également disponible [1]. Nous étudions spécifiquement notre approche sur 3 jeux de données et 8 modèles différents. A chaque fois, nous comparons les résultats obtenus en utilisant la fonction de perte originale du modèle et en utilisant notre version guidée par la sémantique, au regard de Sem@K [2] ainsi que des métriques traditionnelles basées sur le rang (MRR, Hits@K). Nous observons une nette amélioration des capacités sémantiques des modèles dans la quasi totalité des cas, ainsi qu'une amélioration satisfaisante des résultats en termes de MRR et Hits@K dans la majorité des cas. Ces résultats démontrent que notre approche n'améliore pas seulement la validité sémantique des prédictions, mais est également pertinente au regard des métriques traditionnelles basées sur le rang.

4 Conclusion

Ce travail se concentre sur les principales fonctions de perte utilisées pour la prédiction de liens dans les GCs. En nous appuyant sur l'hypothèse que tous les triplets négatifs ne sont pas égaux pour apprendre de meilleures représentations, nous proposons de les différencier en fonction de leur validité sémantique par rapport au domaine et co-domaine des prédicats lors de l'entrainement des modèles. Notre approche conduit les modèles à faire des prédictions sémantiquement plus plausibles et améliore également leur capacité à assigner un score plus élevé au triplet authentique.

Références

[1] Nicolas Hubert, Pierre Monnin, Armelle Brun, and Davy Monticolo. Treat different negatives differently : Enriching loss functions with domain and range constraints for link prediction. In *The Semantic Web - 21st International Conference, ESWC 2024, Proceedings*.

[2] Nicolas Hubert, Pierre Monnin, Armelle Brun, and Davy Monticolo. Sem@k : Is my knowledge graph embedding model semantic-aware ? volume 14, pages 1–37, 12 2023.

[3] Bhushan Kotnis and Vivi Nastase. Analysis of the impact of negative sampling on link prediction in knowledge graphs. *arXiv preprint 1708.06816*, 2017.

[4] Denis Krompaß, Stephan Baier, and Volker Tresp. Type-constrained representation learning in knowledge graphs. In *The Semantic Web - 14th International Semantic Web Conference (ISWC)*, volume 9366, pages 640–655, 2015.

[5] Andrea Rossi, Denilson Barbosa, Donatella Firmani, Antonio Matinata, and Paolo Merialdo. Knowledge graph embedding for link prediction : A comparative analysis. *ACM Transactions on Knowledge Discovery from Data*, 15(2) :14 :1–14 :49, 2021.

1. https://github.com/nicolas-hbt/semantic-lossfunc/

Techniques neurosymboliques probabilistes pour la classification supervisée informée par la logique

Arthur Ledaguenel[1,2], Celine Hudelot[2], Mostepha Khouadjia[1]

[1] IRT SystemX, Palaiseau
[2] CentraleSupélec, MICS

arthur.ledaguenel@irt-systemx.fr

Résumé

L'IA neurosymbolique est un champ de recherche émergent dont l'objectif est de combiner les capacités d'apprentissage des réseaux de neurones avec les aptitudes de raisonnement des systèmes symboliques. Ce papier présente un formalisme pour la classification supervisée informée par la logique, décrit succinctement les principales tâches et jeux de données traitées par la littérature, puis détaille un ensemble de techniques neurosymboliques basées sur le raisonnement probabiliste et analyse leur complexité asymptotique.

Mots-clés

Neurosymbolique, classification, logique.

Abstract

Neurosymbolic AI is a growing field of research aiming to combine neural networks learning capabilities with the reasoning abilities of symbolic systems. In this paper, we introduce a formalism for supervised multi-label classification informed by propositional background knowledge, desribe the main tasks and datasets tackled in the literature, then present a set of neurosymbolic techniques based on probabilistic logics and analyze their asymptotic complexity.

Keywords

Neurosymbolic, classification, logic.

1 Introduction

L'intelligence artificielle neurosymbolique est un champ de recherche émergent dont l'objectif est de combiner les capacités d'apprentissage des réseaux de neurones avec les aptitudes de raisonnement des systèmes symboliques. Cette hybridation peut prendre de nombreuses formes en fonction de la tâche traitée et des avantages ciblés [22, 44].

Un sous-domaine important de l'IA neurosymbolique est l'apprentissage machine informé (*informed machine learning*) [39], qui étudie comment exploiter de la connaissance *a priori* pour améliorer un système d'apprentissage. De nouveau, les techniques introduites dans la littérature peuvent être de natures très diverses selon le type de tâche (eg. régression, classification, détection, génération, etc.), le formalisme utilisé pour représenter la connaissance (eg. équations mathématiques, graphes de connaissances, logiques, etc.), l'étape à laquelle la connaissance est intégrée (eg. traitement des données, design de l'architecture du réseau de neurones, procédure d'apprentissage, procédure d'inférence, etc.) ou même les avantages attendus de l'hybridation (eg. explicabilité, performance, frugalité, etc.).

Dans ce papier, nous introduisons un formalisme pour la classification multi-classes supervisée informée par la logique.

Les contributions et le plan de l'article sont les suivants. Après quelques notions préliminaires sur la classification neuronale, la logique propositionnelle et le raisonnement probabiliste en Section 2, nous introduisons en Section 3 notre nouveau formalisme pour représenter une tâche de classification multi-classes supervisée informée par la logique. La connaissance *a priori* est exprimée par une formule propositionnelle qui décrit l'ensemble des combinaisons de classes sémantiquement *valides*. Nous illustrons ce formalisme par quelques exemples de tâches et jeux de données les plus fréquemment utilisés dans la littérature neurosymbolique. Puis nous nous appuyons sur ce formalisme dans la Section 4 pour reformuler les principales techniques neurosymboliques probabilistes existantes et analysons leurs principales propriétés dans la Section 5. Nous discutons dans la Section 6 des problèmes de complexité relatifs à ces techniques. Enfin, nous exposons les travaux connexes en Section 7 et concluons en Section 8 avec des pistes de recherche pour de futurs travaux.

2 Préliminaires

2.1 Classification neuronale

En apprentissage machine supervisé, l'objectif est d'apprendre une relation fonctionnelle $f : \mathcal{X} \mapsto \mathcal{Y}$ entre un **domaine d'entrée** $\mathcal{X}$ et un **domaine de sortie** $\mathcal{Y}$ à partir d'un jeu de données annoté $\mathsf{D} := (x^i, \mathbf{y}^i)_{1 \leq i \leq d} \in (\mathcal{X} \times \mathcal{Y})^d$. Les systèmes d'apprentissage profond sont habituellement décrits en deux modules : un réseau de neurones profond (*i.e.* un graphe computationnel paramétrique et différentiable) M est conçu pour émuler au mieux la fonction f et un module de coût différentiable L est utilisé pour mesurer la distance entre les prédictions et les annotations. Les poids du réseau sont alors optimisés en utilisant la descente de gra-

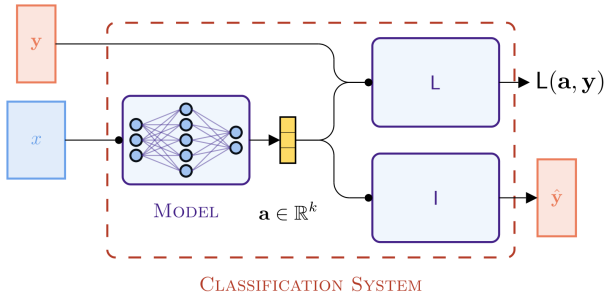


FIGURE 1 – Illustration d'un système neuronal de classification

dient afin de minimiser l'erreur empirique.
Cependant, cette description ne peut pas être appliquée telle quelle pour des tâches de classification : étant donné que l'espace de sortie est discret, un module de coût différentiable ne peut pas être défini sur $\mathcal{Y}^2$.
La classification multi-classes est un type de tâches d'apprentissage pour lesquelles les éléments de sortie sont des sous-ensemble d'un ensemble fini de classes $\mathbf{Y}$. On appelle un tel sous-ensemble un **état**, et le domaine de sortie (qui contient l'ensemble des états) est noté $\mathcal{Y} = \mathbb{B}^{\mathbf{Y}}$, avec $\mathbb{B} = \{0, 1\}$. Un état $\mathbf{y} \in \mathbb{B}^{\mathbf{Y}}$ peut également être vu comme une fonction qui associe à chaque variable une valeur dans $\mathbb{B}$ (*i.e.* une variable est associée à 1 par l'état si elle appartient au sous-ensemble décrit par l'état).
Ainsi, nous adoptons une description légèrement modifiée, dans laquelle un troisième module I, appelé le module d'inférence, doit être défini pour combler l'espace entre la nature continue du réseau de neurones (nécessaire à la descente de gradient) et la nature discrète du domaine de sortie. Ce troisième module, bien qu'essentiel, est rarement explicitement défini. Une illustration de cette description d'un système neuronal de classification peut être trouvée sur la Figure 1.

Définition 1. Un **système neuronal de classification** est constitué de :

- un module **paramétrique différentiable** (*i.e.* neuronal) M, appelé le **modèle**, qui prend en entrée une instance $x \in \mathcal{X}$, des paramètres $\theta \in \Theta$ et donne en sortie un vecteur de scores $\mathsf{M}(x, \theta) := \mathsf{M}_\theta(x) := \mathbf{a} \in \mathbb{R}^k$, appelé **scores pré-activation** ou **logits**.
- un module **non-paramétrique différentiable** L, appelé le module de **perte**, qui prend en entrée des logits $\mathbf{a} \in \mathbb{R}^k$ une annotation $\mathbf{y} \in \mathbb{B}^{\mathbf{Y}}$ et donne en sortie un scalaire positif $\mathsf{L}(\mathbf{a}, \mathbf{y})$.
- un module **non-paramétrique** I, appelé le module d'**inférence**, qui prend en entrée des logits $\mathbf{a} \in \mathbb{R}^k$ et donne en sortie une prédiction $\hat{\mathbf{y}} \in \mathbb{B}^{\mathbf{Y}}$.

Une approche classique pour définir un système neuronal de classification découle d'une interprétation probabiliste. Les *logits* produits par le réseau de neurones sont vus comme des paramètres d'une distribution conditionnelle de probabilité sur le domaine de sortie en fonction de l'entrée $\mathcal{P}(\cdot|\mathsf{M}_\theta(x))$, le module de perte calcule l'entropie croisée de cette distribution avec les annotations tandis que le module d'inférence donne la prédiction la plus probable étant donné la distribution apprise.
Quand aucune connaissance *a priori* n'est disponible (cas non-informé), il est classique de faire l'hypothèse que les variables de sorties sont indépendantes. On illustre ci-dessous comment cette hypothèse peut se traduire dans un système neuronal de classification.

Indépendante Pour la classification indépendante multi-classes (*cim*), on applique une couche sigmoïde par dessus le réseau pour transformer les *logits* en scores de probabilité. Le module de perte calcule l'entropie croisée entre ces scores et l'annotation, tandis que le module d'inférence prédit 1 pour toutes les variables dont la probabilité dépasse 0.5 et 0 pour les autres. Cela se traduit par les modules suivants :

$$\mathsf{L}_{cim}(\mathbf{a}, \mathbf{y}) := \mathtt{BCE}(\mathsf{s}(\mathbf{a}), \mathbf{y})$$
$$= -\sum_j y_j . \log(\mathsf{s}(a_j)) + (1 - y_j) . \log(1 - \mathsf{s}(a_j)) \quad (1)$$

$$\mathsf{I}_{cim}(\mathbf{a}) := \mathbb{1}[\mathbf{a} \geq 0] \quad (2)$$

où $\mathsf{s}(a_i) = \frac{e^{a_i}}{1+e^{a_i}}$ est la fonction sigmoïde, BCE est l'entropie croisée binaire et $\mathbb{1}[z] := \begin{cases} 1 & \text{si z vrai} \\ 0 & \text{sinon} \end{cases}$ est la fonction indicatrice.

2.2 Logique propositionnelle

Une signature propositionnelle est un ensemble $\mathbf{Y}$ de symboles appelés **variables** (*e.g.* $\mathbf{Y} = \{a, b\}$). Une **formule propositionnelle** est formée de manière inductive en combinant variables et autres formules par des connecteurs unaires ($\neg$, qui exprime la négation) et binaires ($\vee, \wedge$, qui expriment la disjonction et la conjonction respectivement). Par exemple, la formule $\kappa = a \wedge b$ exprime la conjonction des deux variables a et b. Un **état** de $\mathbf{Y}$ est une application $\mathbf{y} : \mathbf{Y} \mapsto \mathbb{B}$, où $\mathbb{B} := \{0, 1\}$. Un état $\mathbf{y}$ peut être prolongé en une **valuation** $\mathbf{y}^*$ sur l'ensemble des formules en suivant la sémantique usuelle (*e.g.* $\mathbf{y}^*(a \wedge b) = \mathbf{y}(a) \times \mathbf{y}(b)$). On dit qu'un état $\mathbf{y}$ **satisfait** une formule κ, noté $\mathbf{y} \models \kappa$, ssi $\mathbf{y}^*(\kappa) = 1$. Une formule est dite **satisfiable** s'il existe un état qui la satisfait. Deux formules sont dites équivalentes, noté $\kappa \equiv \gamma$, ssi elles sont satisfaites par les mêmes états. Dans la suite du papier et sauf mention contraire on supposera que l'ensemble de variables est fini et on notera $\mathbf{Y} = \{Y_j\}_{1 \leq j \leq k}$. On se réfère à [41] pour plus de détails sur la logique propositionnelle.

2.3 Raisonnement probabiliste

Un défi pour l'IA neurosymbolique est de combler l'écart entre la nature discrète de la logique et la nature continue des réseaux de neurones. Dans cette section, on définit les notion de distributions discrètes et de raisonnement probabiliste afin de fournir une interface entre ces deux univers.
Une **distribution de probabilité** sur un ensemble fini de variables $\mathbf{Y}$ est une application $\mathcal{P} : \mathbb{B}^{\mathbf{Y}} \mapsto \mathbb{R}^+$ qui associe un réel positif $\mathcal{P}(\mathbf{y})$ à chaque état $\mathbf{y} \in \mathbb{B}^{\mathbf{Y}}$, de sorte que la somme des valeurs donne $\sum_{\mathbf{y} \in \mathbb{B}^{\mathbf{Y}}} \mathcal{P}(\mathbf{y}) = 1$. Pour pouvoir

définir des opérations internes, comme la multiplication de deux distributions, nous étendons cette définition à des distributions non-normalisées $\mathcal{E} : \mathbb{B}^{\mathbf{Y}} \mapsto \mathbb{R}^+$. La distribution nulle associe tous les états à 0. La fonction de **partition** $\mathsf{Z} : \mathcal{E} \mapsto \sum_{\mathbf{y} \in \mathbb{B}^{\mathbf{Y}}} \mathcal{E}(\mathbf{y})$ donne la somme de la distribution sur tous ses états. On note $\overline{\mathcal{E}} := \frac{\mathcal{E}}{\mathsf{Z}(\mathcal{E})}$ la distribution **normalisée** (lorsque $\mathcal{E}$ est non nulle). Le **mode** d'une distribution $\mathcal{E}$ est son état le plus probable, *i.e.* $\underset{\mathbf{y} \in \mathbb{B}^{\mathbf{Y}}}{\operatorname{argmax}} \mathcal{E}(\mathbf{y})$.
Une distribution classique dans la littérature est la distribution exponentielle, qui est au coeur des modules de perte et d'inférence utilisés pour la classification indépendante multi-classes.

Définition 2. Étant donné un vecteur $\mathbf{a} \in \mathbb{R}^k$, on définit la **distribution exponentielle** comme :

$$\mathcal{E}(\cdot|\mathbf{a}) : \mathbf{y} \mapsto \prod_{1 \leq i \leq k} e^{a_i \cdot y_i}$$

On note également $\mathcal{P}(\cdot|\mathbf{a}) = \overline{\mathcal{E}(\cdot|\mathbf{a})}$ la distribution de probabilité correspondante.

Remarque 1. La distribution exponentielle est la distribution jointe de variables indépendantes de Bernouilli $\mathcal{B}(p_i)_{1 \leq i \leq k}$ avec $p_i = \mathsf{s}(a_i)$, où $\mathsf{s}(\mathbf{a}) = (\frac{e^{a_j}}{1+e^{a_j}})_{1 \leq j \leq k}$ est la fonction sigmoïde.

Exemple 1. *La Table 1 représente la distribution exponentielle sur deux variables Y_1 et Y_2 paramétrée par le vecteur $\mathbf{a} := (2, -1)$. La fonction de partition est $\mathsf{Z}(\mathcal{E}(\cdot|\mathbf{a})) = 11.5$. Le mode de la distribution $\mathcal{P}(\mathbf{y}|\mathbf{a})$ est $\hat{\mathbf{y}} = (1, 0)$ avec une probabilité de 0.64.*

y_1	y_2	$\mathcal{E}(\mathbf{y}\|\mathbf{a})$	$\mathcal{P}(\mathbf{y}\|\mathbf{a})$
0	0	e^0	0.09
0	1	e^{-1}	0.03
1	0	e^2	**0.64**
1	1	e^1	0.24
		11.5	1

TABLE 1 – Représentation tabulaire d'une distribution.

Traditionnellement, lorsqu'une croyance (*belief*) a propos de variables aléatoires est exprimée par une distribution de probabilité et que de nouvelles informations sont collectées sous la forme d'observations (*evidence*), deux choses nous intéressent : calculer la probabilité de ces observations et mettre à jour nos croyances en utilisant la règle de Bayes, en conditionnant la distribution sur les observations. Le raisonnement probabiliste nous permet d'effectuer les mêmes opérations avec de la connaissance logique à la place d'observations. Prenons une distribution de probabilité $\mathcal{P}$ sur un ensemble de variables $\mathbf{Y} := \{Y_j\}_{1 \leq j \leq k}$ et une formule propositionnelle **satisfiable** κ sur ce même ensemble de variables. Notons $\mathbb{1}_\kappa$ la fonction indicatrice de κ qui associe 1 aux états qui satisfont κ et 0 aux autres :

$$\mathbb{1}_\kappa(\mathbf{y}) = \begin{cases} 1 & \text{si } \mathbf{y} \models \kappa \\ 0 & \text{sinon} \end{cases}$$

Définition 3. La **probabilité** de κ sous $\mathcal{P}$ est la somme des probabilités des états qui satisfont κ, *i.e.* :

$$\mathcal{P}(\kappa) := \mathsf{Z}(\mathcal{P} \cdot \mathbb{1}_\kappa) = \sum_{\mathbf{y} \in \mathbb{B}^{\mathbf{Y}}} \mathcal{P}(\mathbf{y}) \cdot \mathbb{1}_\kappa(\mathbf{y}) \tag{3}$$

La distribution $\mathcal{P}$ **conditionée par** κ, notée $\mathcal{P}(\cdot|\kappa)$, est :

$$\mathcal{P}(\cdot|\kappa) := \overline{\mathcal{P} \cdot \mathbb{1}_\kappa} \tag{4}$$

Remarque 2. Les deux définitions données ci-dessus sont sémantiques et non syntaxiques : elles s'appuient uniquement sur l'ensemble des états qui satisfont la formule et pas sur la syntaxe de la formule, ce qui signifie que deux formules équivalentes auront la même probabilité et la même distribution conditionnée.

Lorsque la distribution utilisée est une distribution de probabilité exponentielle $\mathcal{P}(\cdot|\mathbf{a})$, on note :

$$\mathcal{P}(\kappa|\mathbf{a}) := \mathsf{Z}(\mathcal{P}(\cdot|\mathbf{a}).\mathbb{1}_\kappa) \tag{5}$$

$$\mathcal{P}(\cdot|\mathbf{a}, \kappa) := \frac{\mathcal{P}(\cdot|\mathbf{a}).\mathbb{1}_\kappa}{\mathcal{P}(\kappa|\mathbf{a})} \tag{6}$$

Étant donné que la distribution $\mathcal{P}(\cdot|\mathbf{a})$ est strictement positive (pour tous $\mathbf{a}$), si κ est satisfiable, alors $\mathcal{P}(\kappa|\mathbf{a}) > 0$. Calculer $\mathcal{P}(\kappa|\mathbf{a})$ est un problème de comptage appelé **Probabilistic Query Estimation** (PQE). Calculer le mode de $\mathcal{P}(\cdot|\mathbf{a}, \kappa)$ est un problème d'optimisation appelé **Most Probable Explanation** (MPE). Résoudre ces deux problèmes se révélera au coeur de plusieurs techniques neurosymboliques que l'on introduira (voir Section 4).

Exemple 2. *La Table 2 reprend la distribution de l'Exemple 1 et illustre le raisonnement probabiliste sur la formule $\kappa = \neg Y_1 \vee Y_2$. La probabilité de κ est $\mathcal{P}(\kappa|\mathbf{a}) = 0.36$. Le mode de la distribution $\mathcal{P}(\cdot|\mathbf{a}, \kappa)$ est $\hat{\mathbf{y}} = (1, 1)$ avec une probabilité de 0.67.*

y_1	y_2	$\mathcal{P}(\mathbf{y}\|\mathbf{a})$	$\mathcal{P}(\mathbf{y}\|\mathbf{a}) \cdot \mathbb{1}_\kappa(\mathbf{y})$	$\mathcal{P}(\mathbf{y}\|\mathbf{a}, \kappa)$
0	0	0.09	0.09	0.24
0	1	0.03	0.03	0.09
1	0	0.64	0	0
1	1	0.24	0.24	**0.67**
		1	**0.36**	1

TABLE 2 – Représentation tabulaire d'une distribution.

3 Classification supervisée informée par la logique propositionnelle

On dit qu'une tâche de classification multi-classes supervisée est **informée** lorsque lui est attachée de la connaissance *a priori*, exprimée par une formule propositionnelle κ satisfiable, qui spécifie quels états du domaine de sortie $\mathcal{Y}$ sont **sémantiquement valides**.
Un jeu de données supervisé D est **cohérent** avec la formule κ si toutes les annotations la satisfont (*i.e.* $\forall 1 \leq i \leq$

$n, \mathbf{y}^i \models \kappa$). Dans ce papier, nous faisons l'hypothèse que les jeux de données d'entraînement et de test sont cohérents à la connaissance *a priori*. Cependant, certaines techniques permettent d'assouplir cette hypothèse et de travailler avec des jeux de données contenant des incohérences.
Les techniques évoquées dans ce papier ne s'intéressent pas à l'architecture du réseau en elle-même (eg. perceptron multi-couches, réseau convolutif, réseau récurent, transformer, etc.), qui dépend principalement de la structure du domaine d'entrée (eg. images, textes, etc.), mais se focalisent sur les deux autres modules afin d'intégrer notre connaissance *a priori* sur le domaine de sortie. On donne ci-dessous quelques exemples du type de structures qui peuvent être exprimées en logique propositionnelle, ainsi que de la manière dont on peut intégrer cette connaissance dans les modules de perte et d'inférence.

Catégorique Dans une tâche de classification **catégorique**, une et une seule variable est classifiée comme *vraie* dans chaque état valide. Cette contrainte peut être facilement exprimée en logique propositionnelle :

$$\kappa_{\odot_k} := \left(\bigvee_{1 \leq j \leq k} Y_j \right) \wedge \left(\bigwedge_{1 \leq j < l \leq k} (\neg Y_j \vee \neg Y_l) \right) \quad (7)$$

où la première partie impose qu'une variable soit classifiée comme *vraie* et la seconde partie empêche deux variables d'être *vraies* en même temps.
Pour la classification catégorique, la couche sigmoïde est habituellement remplacée par une couche *softmax*, et la variable avec le score de probabilité maximum est prédite, ce qui donne les modules suivants :

$$\mathsf{L}_{\odot_k}(\mathbf{a}, \mathbf{y}) := \mathtt{CE}(\mathbf{s}(\mathbf{a}), \mathbf{y}) = -\log(\langle \sigma(\mathbf{a}), \odot_k(j) \rangle) \quad (8)$$

$$\mathsf{I}_{\odot_k}(\mathbf{a}) := \odot_k(\mathrm{argmax}(\mathbf{a})) \quad (9)$$

où CE est l'entropie croisée, $\sigma(\mathbf{a}) = (\frac{e^{a_j}}{\sum_l e^{a_l}})_{1 \leq j \leq k}$ et $\odot_k$ donne le *one-hot encoding* (en commençant à 1) de $j \in [\![1, k]\!]$, *e.g.* $\odot_4(2) = (0, 1, 0, 0)$.

Exemple 3 (MNIST)**.** *MNIST [28] est l'un des jeux de données les plus vieux de la vision par ordinateur et consiste en des petites images de chiffres manuscrits (e.g. 4 ou 9). La tâche de classification catégorique a pour domaine d'entrée l'espace des images* 28×28 *en niveaux de gris, et une classe pour chaque chiffre. Comme décrit plus haut, la connaissance a priori sur la tâche est simplement exprimée par la formule* $\kappa_{\odot_{0:9}}$.

Ce type de tâches peut être élargi pour inclure les tâches **multi-catégoriques** pour lesquelles l'ensemble de variables peut être partitionné en plusieurs groupes de variables catégoriques. Le formule propositionnelle serait alors une conjonction de formule catégoriques sur des ensembles de variables disjoints.

Exemple 4 (Leptograpsus)**.** *Le jeu de données Leptograpsus [9] décrit* 5 *mesures morphologiques effectuées sur* 200 *crabes de couleurs et de sexes différents. L'objectif de la tâche de classification multi-catégorique associée est de prédire la couleur et le sexe d'un crabe à partir de ces mesures. Le domaine d'entrée est l'espace des mesures morphologiques* $\mathcal{X} = \mathbb{R}^5$ *et le domaine de sortie est l'ensemble des états des variables* $\mathbf{Y} = \{\mathsf{r}, \mathsf{b}, \mathsf{f}, \mathsf{m}\}$ *pour les classes **red**, **blue**, **female** et **male** respectivement. La connaissance a priori sur la tâche impose que chaque crabe soit d'une et une seule couleur, et d'un et un seul sexe, i.e. :*

$$\kappa := (\mathsf{r} \vee \mathsf{b}) \wedge (\neg \mathsf{r} \vee \neg \mathsf{b}) \wedge (\mathsf{f} \vee \mathsf{m}) \wedge (\neg \mathsf{f} \vee \neg \mathsf{m})$$

Hiérarchique La classification **hiérarchique** sur un ensemble de variables $\mathbf{Y}$ est usuellement représentée par un graphe dirigé acyclique $G = (\mathbf{Y}, E_h)$ où les noeuds sont les variables et les arrêtes E_h exprime l'inclusion d'une classe dans une autre (*e.g.* un chien est un animal). Lorsque le graphe est un arbre (ou une forêt) on parle de classification **taxonomique**. Ce formalisme peut également être enrichi par des arrêtes d'exclusion $H = (\mathbf{Y}, E_h, E_e)$ (*e.g.* une instance ne peut pas être classifiée comme chien et chat simultanément), comme dans les HEX-graphs [15]. Là encore, la traduction en logique propositionnelle vient naturellement :

$$\kappa_H := \left(\bigwedge_{(i,j) \in E_h} Y_i \vee \neg Y_j \right) \wedge \left(\bigwedge_{(i,j) \in E_e} (\neg Y_i \vee \neg Y_j) \right) \quad (10)$$

où la première partie s'assure qu'une classe fille ne peut être *vraie* que si sa classe mère l'est également et la seconde partie empêche deux classes mutuellement exclusives d'être *vraies* en même temps.
Dans le cas hiérarchique il n'y a pas de consensus dans la littérature sur les modules de pertes et d'inférence à utiliser : [36] définit un module de perte hiérarchique pour plus pénaliser les erreurs faîtes sur les plus hautes classes de la hiérarchies (en conservant le module d'inférence de *cim*), [21] affine les *logits* en fonction de la hiérarchie tandis que [15] conditionne la distribution exponentielle par la connaissance hiérarchique.

Exemple 5 (Cifar-100)**.** *Cifar-100 dataset [25] est un jeu de données composé de 60,000 images classifiées en 20 macro-classes (e.g. reptile), chacune divisée en 5 micro-classes (e.g. crocodile, dinosaur, lizard, turtle, et snake). Le domaine d'entrée est l'espace des images RGB de taille* 32×32. *Typiquement utilisé dans un cadre catégorique sur les 100 micro-classes, il peut également être utilisé dans le cadre d'une tâche de classification hiérarchique en incluant les macro-classes. Les relations hiérarchiques d'une classe macro vers ses classes micro, ainsi que les relations d'exclusion entre les macro-classes et entre les micro-classes peuvent être encodées dans un HEX-graphe* $H = (Y, E_h, E_e)$ *et donc exprimées par une formule propositionnelle* κ_H.

Exemple 6 (ImageNet)**.** *Le ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [40] est un challenge de classification d'images, qui a eu lieu annuellement entre 2010 et 2017, et s'est imposé comme un benchmark de référence en vision par ordinateur pour comparer les performances des systèmes neuronaux de classification. En août*

(a) MNIST-Sudoku [4] (b) Warcraft Shortest Path [3]

FIGURE 2 – Exemples d'instances provenant de jeux de données structurés

2014, le jeu de données ImageNet contenait 14,197,122 images annotées classifiées selon 21,841 synsets de la hiérarchie WordNet [35]. Comme pour Cifar-100 (voire Exemple 5), le jeu de données est le plus souvent utilisé dans un cadre catégorique, mais peut être également utilisé dans un cadre hiérarchique en mobilisant les relations d'inclusion entre synsets définies dans WordNet.

Au delà de ces exemples, la logique propositionnelle peut être utilisée pour définir des domaines de sortie structurés très divers : addition de chiffres manuscrits (*i.e.* k-Add-MNIST) [24], solutions d'un Sudoku [4, 14], chemins dans un graphe dirigé [38, 45, 3], classements [45], matching parfait dans un graphe [38, 1], etc.

4 Techniques neurosymboliques probabilistes

L'objectif d'une technique neurosymbolique est de construire un système neuronal de classification automatiquement à partir de la connaissance *a priori* disponible sur le domaine de sortie, généralisant ainsi les travaux menés sur les cas particuliers de la classification multi-classes indépendante, catégorique et hiérarchique au cas général d'une formule propositionnelle.

Comme précisé plus haut, on ne s'intéresse qu'à la spécification des modules de perte et d'inférence. En particulier, nous étudierons un ensemble de techniques qui s'appuient sur le raisonnement probabiliste pour définir leurs modules. Nous utilisons (abusivement) l'appellation *techniques probabilistes* pour les décrire, même si ces techniques ne disposent pas nécessairement d'une interprétation probabiliste.

Régularisation sémantique Une première technique neurosymbolique utilise une approche multi-objectifs : un terme de régularisation mesurant la cohérence des sorties du modèle avec la connaissance *a priori* est ajouté à l'entropie croisée afin d'optimiser les paramètres sur un double objectif de performance et de cohérence. Il existe plusieurs manières de calculer ce terme de régularisation. Introduit dans un premier temps en utilisant de la logique floue [17, 23, 5], une version basée sur le raisonnement probabiliste est présentée par [45].

Définition 4. Un système neuronal performe la **régularisation sémantique** (*rs*) sur κ ssi ses modules de perte et d'inférence sont :

$$\mathsf{L}^{\lambda}_{rs}(\mathbf{a},\mathbf{y}) = \mathsf{L}_{cim}(\mathbf{a},\mathbf{y}) - \lambda.\log(\mathcal{P}(\kappa|\mathbf{a})) \quad (11)$$

$$\mathsf{I}_{cim}(\mathbf{a}) = \underset{\mathbf{y}\in\mathbb{B}^{\mathbf{Y}}}{\operatorname{argmax}}\mathcal{P}(\mathbf{y}|\mathbf{a}) \quad (12)$$

Conditionnement sémantique Suivant l'interprétation probabiliste introduite en Section 2.1, une manière naturelle d'intégrer de la connaissance *a priori* κ dans un système neuronal de classification est de conditionner la distribution $\mathcal{P}(\cdot|\mathsf{M}_\theta)$ par κ. Ce conditionnement affecte les modules de perte et d'inférence, qui reposent tous deux sur la distribution de probabilité paramétrée par la sortie du réseau de neurones, et conduit à la technique neurosymbolique suivante.

Définition 5. Un système neuronal performe le **conditionnement sémantique** (*cs*) sur κ si ses modules de perte et d'inférence sont :

$$\mathsf{L}_{|\kappa}(\mathbf{a},\mathbf{y}) = -\log(\mathcal{P}(\mathbf{y}|\mathbf{a},\kappa)) \quad (13)$$

$$\mathsf{I}_{|\kappa}(\mathbf{a}) = \underset{\mathbf{y}\in\mathbb{B}^{\mathbf{Y}}}{\operatorname{argmax}}\mathcal{P}(\mathbf{y}|\mathbf{a},\kappa) \quad (14)$$

Les modules de pertes et d'inférence présentés pour la classification indépendante et la classification catégorique sont des cas particuliers du conditionnement sémantique. De même, [15] définit le conditionnement sémantique pour les tâches de classification hiérarchiques. Cette technique est généralisée par [3] à un ensemble de distributions de probabilités définies sur des circuits arithmétiques.

Conditionnement sémantique à l'inférence Le conditionnement sémantique à l'inférence est dérivé du conditionnement sémantique, mais applique le conditionnement uniquement sur le module d'inférence (*i.e.* prédit la sortie la plus probable qui satisfait κ) et conserve le module de perte indépendant.

Définition 6. Un système neuronal performe le **conditionnement sémantique à l'inférence** (*csi*) sur κ ssi ses modules de perte et d'inférence sont :

$$\mathsf{L}_{cim}(\mathbf{a},\mathbf{y}) = -\log(\mathcal{P}(\mathbf{y}|\mathbf{a})) \quad (15)$$

$$\mathsf{I}_{|\kappa}(\mathbf{a}) = \underset{\mathbf{y}\in\mathbb{B}^{\mathbf{Y}}}{\operatorname{argmax}}\mathcal{P}(\mathbf{y}|\mathbf{a},\kappa) \quad (16)$$

En plus de conserver des propriétés clé du conditionnement sémantique (voir Section 5), le conditionnement sémantique à l'inférence se démarque des deux autres techniques en n'intégrant la connaissance que dans le module d'inférence, sans impact sur l'entraînement, ce qui présente deux avantages majeurs. Premièrement, tandis que *cs* et *rs* nécessitent de résoudre PQE pour calculer leur module de perte, *csi* ne nécessite que de résoudre MPE pour son module d'inférence, ce qui est plus facile pour certaines classes de formules (*i.e.* matchings parfaits). Deuxièmement, intégrer la connaissance uniquement à l'inférence offre plus de flexibilité. Par exemple, *csi* peut être utilisé dans le cas

où la connaissance n'est pas disponible pendant l'entraînement. C'est une propriété particulièrement importante à l'ère des **modèles sur étagère** et **modèles de fondations** [8], qui sont pré-entraînés sur des grands volumes de données généralistes avant d'être affinés et combinés sur une multitude de tâches hétérogènes, puisque la connaissance spécifique aux différentes tâches ne peut pas être intégrée pendant la majeure partie de l'entraînement.

5 Propriétés

Nous détaillons ci-dessous quelques propriétés théoriques des techniques présentées. Nous renvoyons le lecteur à [30] pour trouver les définitions et les preuves formelles des propriétés énoncées.

Insensibilité syntaxique Une technique neurosymbolique est **insensible à la syntaxe** (*invariant to syntax*) si deux formules équivalentes aboutissent à des modules de pertes et d'inférence identiques. Toutes les techniques qui s'appuient sur le raisonnement probabiliste sont insensibles à la syntaxe (voir Remarque 2), mais ce n'est pas le cas des techniques qui utilisent la logique floue [5].

Consistance Une technique neurosymbolique est consistante (*consistent*) si les prédictions du module d'inférence satisfont nécessairement la connaissance *a priori* κ. Le conditionnement sémantique ainsi que le conditionnement sémantique à l'inférence sont toutes deux consistantes. En revanche, les techniques neurosymboliques qui n'intègrent la connaissance que dans le module de perte (comme la régularisation sémantique) ne peuvent pas garantir la consistance.

Supérieur à l'inférence Un module d'inférence L_1 est dit supérieur à un module d'inférence L_2 ssi quel que soit les scores d'entrée du module, si la prédiction de L_2 est consistante avec la connaissance a priori, alors la prédiction de L_1 est identique à celle de L_2. Cette propriété est intéressante car elle garantit, sous l'hypothèse de consistance du jeu de données, que la performance du module L_1 sera nécessairement meilleure (au sens large) que celle du module L_2 si on les évalue sur le même réseau de neurones (avec les mêmes poids entraînés). En particulier, on peut montrer que le module d'inférence conditionné $\mathsf{I}_{|\kappa}$ est supérieur au module d'inférence indépendant I_{cim}. Cette garantie théorique se traduit expérimentalement, comme montré dans [30].

6 Algorithmes et complexité

Après avoir introduit les principales techniques neurosymboliques probabilistes dans la section précédente, nous nous attaquons dans cette section à la question de leur implémentation et de leur complexité. Alors que de nombreux papiers pointent les problèmes de passage à l'échelle des ces techniques, leur complexité asymptotique n'est pas étudiée de manière systématique dans la littérature neurosymbolique. Ceci mène à plusieurs déboires : certaines techniques sont illustrées sur des tâches pour lesquelles le passage à l'échelle n'est pas possible, tandis que certaines tâches tractables sont considérées intractables.

On remarque premièrement que toutes les techniques mentionnées précédemment s'appuient sur la résolution de problèmes de PQE et MPE sur des distributions exponentielles. Ces problèmes sont malheureusement #P-complet et NP-complet (par réduction de #SAT et SAT respectivement) et sont donc **intractables** en général. Il devient alors naturel de se demander pour quelles familles de formules ces problèmes peuvent être résolus en temps polynomial. Nous appelons ces familles des **familles tractables** et étudions certains cas dans la Section 6.2.
Pour ce faire, nous présentons d'abord deux approches de résolution des problèmes de PQE et MPE, basées sur les modèles graphiques et la compilation de connaissance respectivement. On identifie ensuite quelques familles de formules tractables et intractables.

6.1 Algorithmes

Modèles graphiques Les modèles graphiques [27, 43] permettent de spécifier une famille de distributions de probabilité sur un ensemble de variables à travers une représentation graphique. Le graphe encode un ensemble de propriété (*e.g.* factorisation, indépendance, etc.) que les distributions qui appartiennent à la famille doivent respecter. Ces propriétés peuvent alors être exploitées afin de produire une représentation compressée des distributions et d'exécuter des algorithmes d'inférence efficaces [26]. Dans le contexte des logiques propositionnelles probabilistes, le graphe primaire d'une formule κ indique le modèle graphique auquel appartiennent les distributions exponentielles conditionnées par κ. En particulier, les algorithmes classiques de passage de messages sur un arbre de jonction peuvent être utilisés pour résoudre les problèmes de PQE et MPE sur ces distributions, avec une complexité en temps $\mathcal{O}(k2^{\tau(\kappa)})$ où k est le nombre de variables et $\tau(\kappa)$ est la largeur d'arbre du graphe primaire de κ.

Compilation de connaissances La compilation de connaissances [11] consiste à traduire une formule propositionnelle dans un langage de représentation qui permet d'effectuer certaines opérations de manière tractable. Les Sentential Decision Diagrams (SDD) [12] (voir Figure 3) est un langage de représentation qui permet la négation, la disjonction et la conjonction en temps polynomial (en la taille du circuit), ainsi que le calcul de PQE et MPE dans un temps linéaire (en la taille du circuit). De plus, il a été montré qu'une formule κ en forme normale conjonctive de largeur d'arbre $\tau(\kappa)$ peut être traduite dans un *trimmed and compressed* SDD de taille $\mathcal{O}(k2^{\tau(\kappa)})$ [12]. En raison de ces propriétés, les SDD sont devenus un langage standard de représentation des connaissances pour les systèmes neurosymboliques probabilistes [45, 3].

Programmation Linéaire Binaire Comme souligné dans [37], une formule propositionnelle en forme normale conjonctive peut être facilement compilée dans un programme linéaire équivalent (*i.e.* qui est satisfait par les mêmes états). Ainsi, la tâche de MPE sur cette formule peut alors être résolue en utilisant les nombreux algorithmes combinatoires qui ont été développés pour les problèmes de programmation linéaire binaire ou mixte [7].

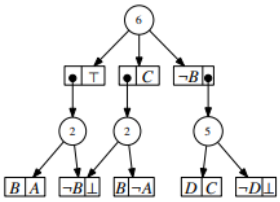


FIGURE 3 – Représentation graphique d'un SDD dans [12]

6.2 Familles tractables

Une familles de formules est dite **tractable** si et seulement si pour toute formule κ de la famille il est possible d'effectuer PQE et MPE sur n'importe quelle distribution exponentielle conditionnée par κ en un temps polynomial (en le nombre de variables de la formule).

Largeur d'arbre bornée Comme énoncé ci-dessus, une condition suffisante de tractabilité d'une famille de formules (et commune aux modèles graphiques et à la compilation de connaissances) est la possibilité de construire en temps polynomial (dans la taille de la signature) une décomposition en arbre de largeur bornée pour toutes les formules de la famille. Ceci implique directement que la famille des formules **taxonomiques** est tractable, puisque de largeur d'arbre 1. Le cas de k-Add-MNIST est intéressant à cet égard : longtemps jugé intractable par la communauté [24], il s'est avéré qu'une représentation légèrement différent du problème (incluant les retenues successives des additions de chiffres) aboutit à une famille de formules de largeur d'arbre bornée (avec une décomposition constructible en temps linéaire) [32].

Énumérable en temps polynomial Un autre type de famille tractable est celui des familles énumérables en temps polynomial, *i.e*. dont on peut énumérer les solutions de chaque formule dans un temps polynomial (en le nombre de variables de la formule). Il est facile de voir pourquoi MPE et PQE sont tous deux calculables en temps polynomial pour une famille énumérable : il suffit de lister chaque état valide et sa probabilité pour compter la probabilité de la formule et trouver son état le plus probable. Cette famille recouvre par exemple le cas des formules **catégoriques** et celui des formules **hiérarchiques** sur un **HEX-graphe en forme d'arbre** et **saturé**.

Cependant, ces deux conditions n'épuisent pas l'ensemble des familles tractable, comme le montre les exemples suivants.

Formules multi-catégoriques La famille des formules multi-catégoriques est un bon exemple des limites des critères de largeur d'arbre et d'énumérabilité pris séparément. En effet, il est facile de voir qu'il s'agit d'une famille de largeur d'arbre non bornée (la largeur d'arbre d'une formule catégorique est son nombre de variables) et non énumérable en temps polynomial (le nombre d'états valides est potentiellement exponentiel). Cependant, il est possible de décomposer chaque formule en composantes indépendantes et énumérables, ce qui rend la famille tractable.

Chemins simples La famille des formules qui encodent les chemins (simples) dans un graphe dirigé acyclique est elle aussi de largeur d'arbre non bornée et non énumérable. Cependant, en suivant un ordre topologique des arrêtes du graphe, il est possible de compiler cette formule en un OBDD [11] de taille polynomiale (dans le nombre d'arêtes) [29]. Cette propriété permet d'assurer la tractabilité de MPE et PQE pour cette famille. Il est de plus intéressant de remarquer que le calcul de MPE pour cette famille peut se ramener facilement à un calcul de plus court chemin, et donc être résolu avec des algorithmes combinatoires classiques (*e.g*. Bellman-Ford [6] ou Dijkstra [16]) : les probabilités sur les arrêtes sont transformés en scores réels par une sigmoïde inverse et multipliés par -1. Cette équivalence avec le problème de plus court chemin nous indique également qu'une famille de formules qui encodent les chemins simples dans un graphe dirigé (potentiellement cyclique) est **intractable**. En effet, le calcul de plus court chemin (avec poids négatifs) est un problème NP-complet, par réduction du problème d'existence d'un cycle Hamiltonien [19]. De plus, le calcul de PQE est #P-complet pour cette famille, par réduction au problème de comptage des chemins simples dans un graphe dirigé [42].

7 Travaux connexes

Logiques alternatives Il existe d'autres langages et sémantiques que la logique propositionnelle pour exprimer de la connaissance *a priori* sur une tâche de classification. Si certains correspondent à un fragment de la logique propositionnelle, comme les HEX-graphs dans [15], d'autres lui sont incommensurables, comme la programmation logique avec sémantique des modèles stables dans [46] ou la programmation par contraintes linéaires dans [37], voire passent à l'ordre supérieur, comme le langage de programmation logique Prolog [13] dans [33] ou la logique de premier ordre dans [5]. Les compromis pour ces différents langages sont principalement entre leur concision, leur expressivité et leur tractabilité. Les conséquences du choix de langage en termes de complexité de calculs ne sont pas encore bien comprises.

Logiques floues De nombreux travaux utilise les **logiques floues** [20, 23, 5] à la place du raisonnement probabiliste comme moyen de faire le pont entre la nature discrète de la connaissance et la nature continue du réseau de neurones.

Logiques pondérées Dans notre formalisme, nous avons fait l'hypothèse que la connaissance *a priori* est une contrainte dure, systématiquement satisfaite dans les jeux d'entraînement et de test. Certains formalismes permettent un langage plus souple qui permettent de représenter de l'incertitude sur la connaissance *a priori*. Des logiques pondérées permettent par exemple d'exprimer quelles formules ont *le plus de chances* d'être satisfaites et peuvent être intégrées dans des systèmes neurosymboliques [10, 34]. Les travaux de la *theory of evidence* [31] ou des *probability kinematics* [18] offrent des outils théoriques qui permettent de combiner de l'information provenant de deux sources incertaines (*e.g*. un réseau de neurones et de la connaissance

a priori probabiliste).

Apprentissage non supervisé De nombreux travaux explorent le potentiel de l'IA neurosymbolique dans un cadre d'apprentissage non *pleinement* supervisé (*i.e*. chaque instance du jeu de données est annotée sur l'ensemble des classes), comme l'apprentissage **faiblement supervisé** (*i.e*. certaines instances ne pas annotées sur toutes les classes) ou **semi-supervisé** (*i.e*. certaines instances ne sont pas annotées). Les techniques de régularisation [2, 45, 23, 5, 17] en particulier sont particulièrement indiquées pour l'apprentissage semi-supervisé et montrent que l'intégration de connaissances *a priori* dans le processus d'apprentissage peut grandement diminuer le besoin en instances annotées tout en augmentant les performances du système. Les techniques de conditionnement sémantique [3, 33, 46] permettent de traiter certaines variables comme des variables latentes en marginalisant la distribution apprise, et ne nécessitent donc pas de labels sur les variables latentes pendant l'apprentissage.

8 Conclusion

Après avoir présenté un formalisme pour la classification supervisée informée par la logique propositionnelle, ce papier réalise une revue de littérature des jeux de données structurés ainsi que des techniques neurosymboliques probabilistes. Enfin, nous détaillons quelques résultats relatifs à la complexité asymptotique des techniques probabilistes, qui manque d'une étude systématique dans la littérature.
Les pistes de recherche pour nos futurs travaux incluent, entre autres : une meilleure cartographie des familles tractables pour les différentes techniques probabilistes, l'utilisation de logiques alternatives, une étude des cadres d'apprentissages non-supervisés en IA neurosymbolique, ou encore l'apprentissage simultané du modèle neuronal et de la structure de la tâche.

Remerciements

Ce travail a été soutenu par le gouvernement français dans le cadre du programme "France 2O30" au sein de l'Institut de Recherche Technologique SystemX.

Références

[1] Kareem AHMED, Kai-Wei CHANG et Guy VAN DEN BROECK. « Semantic Strengthening of Neuro-Symbolic Learning ». In : *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. Sous la dir. de Francisco RUIZ, Jennifer DY et Jan-Willem van de MEENT. T. 206. Proceedings of Machine Learning Research. PMLR, 25–27 Apr 2023, p. 10252-10261. URL : https://proceedings.mlr.press/v206/ahmed23a.html.

[2] Kareem AHMED et al. « Neuro-symbolic entropy regularization ». In : *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*. Sous la dir. de James CUSSENS et Kun ZHANG. T. 180. Proceedings of Machine Learning Research. PMLR, jan. 2022, p. 43-53. URL : https://proceedings.mlr.press/v180/ahmed22a.html.

[3] Kareem AHMED et al. « Semantic Probabilistic Layers for Neuro-Symbolic Learning ». In : *Advances in Neural Information Processing Systems*. Sous la dir. de S KOYEJO et al. T. 35. Curran Associates, Inc., 2022, p. 29944-29959.

[4] Eriq AUGUSTINE et al. « Visual Sudoku Puzzle Classification : A Suite of Collective Neuro-Symbolic Tasks ». In : *International Workshop on Neural-Symbolic Learning and Reasoning*. 2022.

[5] Samy BADREDDINE et al. « Logic Tensor Networks ». In : *Artificial Intelligence* 303 (2022), p. 103649. ISSN : 0004-3702. DOI : https://doi.org/10.1016/j.artint.2021.103649. URL : https://www.sciencedirect.com/science/article/pii/S0004370221002009.

[6] Richard BELLMAN. « On a Routing Problem ». In : *Quarterly of Applied Mathematics* 16.1 (1958). Publisher : Brown University, p. 87-90. ISSN : 0033-569X. URL : https://www.jstor.org/stable/43634538 (visité le 23/04/2024).

[7] M. BENICHOU et al. « Experiments in mixed-integer linear programming ». In : *Mathematical Programming* 1.1 (1er déc. 1971), p. 76-94. ISSN : 1436-4646. DOI : 10.1007/BF01584074. URL : https://doi.org/10.1007/BF01584074 (visité le 02/05/2024).

[8] Rishi BOMMASANI et al. « On the Opportunities and Risks of Foundation Models ». In : *CoRR* abs/2108.07258 (2021). arXiv : 2108.07258. URL : https://arxiv.org/abs/2108.07258.

[9] N. A. CAMPBELL et R. J. MAHON. « A multivariate study of variation in two species of rock crab of the genus Leptograpsus ». In : *Australian Journal of Zoology* 22.3 (1974). Publisher : CSIRO PUBLISHING, p. 417-425. ISSN : 1446-5698. DOI : 10.1071/zo9740417. URL : https://www.publish.csiro.au/zo/zo9740417 (visité le 02/05/2024).

[10] Alessandro DANIELE et Luciano SERAFINI. « Knowledge Enhanced Neural Networks ». In : *PRICAI 2019 : Trends in Artificial Intelligence : 16th Pacific Rim International Conference on Artificial Intelligence, Cuvu, Yanuca Island, Fiji, August 26–30, 2019, Proceedings, Part I*. Berlin, Heidelberg : Springer-Verlag, 26 août 2019, p. 542-554. ISBN : 978-3-030-29907-1. DOI : 10.1007/978-3-030-29908-8_43. URL : https://doi.org/10.1007/978-3-030-29908-8_43 (visité le 19/04/2024).

[11] Adnan DARWICHE. *Modeling and Reasoning with Bayesian Networks*. Cambridge : Cambridge University Press, 2009. ISBN : 978-0-521-88438-9. DOI : 10.1017/CBO9780511811357. URL : https://www.cambridge.org/core/books/modeling-and-reasoning-with-bayesian-networks/8A3769B81540EA93B525C4C2700C9DE6 (visité le 07/08/2023).

[12] Adnan DARWICHE. « SDD : A New Canonical Representation of Propositional Knowledge Bases ». In : *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*. 2011.

[13] Luc DE RAEDT, Angelika KIMMIG et Hannu TOIVONEN. « ProbLog : a probabilistic prolog and its application in link discovery ». In : *Proceedings of the 20th international joint conference on Artifical intelligence*. IJCAI'07. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., jan. 2007, p. 2468-2473. (Visité le 07/08/2023).

[14] Marianne DEFRESNE, Sophie BARBE et Thomas SCHIEX. « Scalable coupling of deep learning with logical reasoning ». In : *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. IJCAI '23. <conf-loc>, <city>Macao</city>, <country>P.R.China</country>, </conf-loc>, 19 août 2023, p. 3615-3623. ISBN : 978-1-956792-03-4. DOI : 10.24963/ijcai.2023/402. URL : https://doi.org/10.24963/ijcai.2023/402 (visité le 23/02/2024).

[15] Jia DENG et al. « Large-Scale Object Classification Using Label Relation Graphs ». In : *Computer Vision – ECCV 2014*. Sous la dir. de David FLEET et al. Springer International Publishing, 2014, p. 48-64. ISBN : 978-3-319-10590-1.

[16] E. W. DIJKSTRA. « A note on two problems in connexion with graphs ». In : *Numerische Mathematik* 1.1 (1er déc. 1959), p. 269-271. ISSN : 0945-3245. DOI : 10.1007/BF01386390. URL : https://doi.org/10.1007/BF01386390 (visité le 21/02/2024).

[17] Michelangelo DILIGENTI, Marco GORI et Claudio SACCÀ. « Semantic-based regularization for learning and inference ». In : *Artificial Intelligence* 244 (mars 2017), p. 143-165. ISSN : 00043702. DOI : 10.1016/j.artint.2015.08.011.

[18] Zoltan DOMOTOR, Mario ZANOTTI et Henson GRAVES. « Probability Kinematics ». In : *Synthese* 44.3 (1980). Publisher : Springer, p. 421-442. ISSN : 0039-7857. URL : https://www.jstor.org/stable/20115538 (visité le 02/05/2024).

[19] Michael R. GAREY et David S. JOHNSON. *Computers and Intractability : A Guide to the Theory of NP-Completeness*. USA : W. H. Freeman & Co., 1979. 338 p. ISBN : 978-0-7167-1044-8.

[20] Francesco GIANNINI et al. « T-norms driven loss functions for machine learning ». In : *Applied Intelligence* 53.15 (fév. 2023), p. 18775-18789. ISSN : 0924-669X. DOI : 10.1007/s10489-022-04383-6. URL : https://doi.org/10.1007/s10489-022-04383-6 (visité le 07/08/2023).

[21] Eleonora GIUNCHIGLIA et Thomas LUKASIEWICZ. « Coherent Hierarchical Multi-Label Classification Networks ». In : *Advances in Neural Information Processing Systems*. T. 33. Curran Associates, Inc., 2020, p. 9662-9673. URL : https://proceedings.neurips.cc/paper/2020/hash/6dd4e10e3296fa63738371ec0d5df818-Abstract.html (visité le 13/09/2023).

[22] Henry A. KAUTZ. « The Third AI Summer : AAAI Robert S. Engelmore Memorial Lecture ». In : *AI Mag.* 43 (2022), p. 93-104.

[23] Emile van KRIEKEN, Erman ACAR et Frank van HARMELEN. « Analyzing Differentiable Fuzzy Logic Operators ». en. In : *Artificial Intelligence* 302 (jan. 2022), p. 103602. ISSN : 0004-3702. DOI : 10.1016/j.artint.2021.103602. URL : https://www.sciencedirect.com/science/article/pii/S0004370221001533 (visité le 07/08/2023).

[24] Emile van KRIEKEN et al. « A-NeSI : A Scalable Approximate Method for Probabilistic Neurosymbolic Inference ». In : *Advances in Neural Information Processing Systems* 36 (13 fév. 2024). URL : https://proceedings.neurips.cc/paper_files/paper/2023/hash/4d9944ab3330fe6af8efb9260aa9f307-Abstract-Conference.html (visité le 21/02/2024).

[25] Alex KRIZHEVSKY. *Learning Multiple Layers of Features from Tiny Images*. 2009.

[26] Frank R. KSCHISCHANG, Brendan J. FREY et Hans Andrea LOELIGER. « Factor graphs and the sum-product algorithm ». In : *IEEE Transactions on Information Theory* 47.2 (2001), p. 498-519. ISSN : 00189448. DOI : 10.1109/18.910572. (Visité le 28/03/2022).

[27] Steffen L. LAURITZEN. *Graphical Models*. Clarendon Press, 1996. ISBN : 978-0-19-852219-5.

[28] Yann LECUN et al. « Gradient-based learning applied to document recognition ». In : *Proceedings of the IEEE* 86 (11 1998), p. 2278-2323. ISSN : 00189219. DOI : 10.1109/5.726791.

[29] Arthur LEDAGUENEL, Céline HUDELOT et Mostepha KHOUADJIA. *Complexity of Probabilistic Reasoning for Neurosymbolic Classification Techniques*. 2024. arXiv : 2404.08404 [cs.AI].

[30] Arthur LEDAGUENEL, Céline HUDELOT et Mostepha KHOUADJIA. *Improving Neural-based Classification with Logical Background Knowledge*. 2024. arXiv : 2402.13019 [cs.AI].

[31] Jianbing MA et al. « Bridging jeffrey's rule, agm revision and dempster conditioning in the theory of evidence ». In : *International Journal on Artificial Intelligence Tools* 20.4 (août 2011). Publisher : World Scientific Publishing Co., p. 691-720. ISSN : 0218-2130. DOI : 10.1142/S0218213011000401. URL : https://www.worldscientific.com/doi/10.1142/S0218213011000401 (visité le 02/05/2024).

[32] Jaron MAENE et Luc DE RAEDT. « Soft-Unification in Deep Probabilistic Logic ». In : *Advances in Neural Information Processing Systems* 36 (15 déc. 2023). URL : https://papers.nips.cc/paper_files/paper/2023/hash/bf215fa7fe70a38c5e967e59c44a99d0-Abstract-Conference.html (visité le 21/02/2024).

[33] Robin MANHAEVE et al. « Neural probabilistic logic programming in DeepProbLog ». en. In : *Artificial Intelligence* 298 (sept. 2021), p. 103504. ISSN : 0004-3702. DOI : 10.1016/j.artint.2021.103504. URL : https://www.sciencedirect.com/science/article/pii/S0004370221000552 (visité le 07/08/2023).

[34] Giuseppe MARRA et Ondřej KUŽELKA. « Neural markov logic networks ». In : *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*. Sous la dir. de Cassio de CAMPOS et Marloes H. MAATHUIS. T. 161. Proceedings of Machine Learning Research. PMLR, 27–30 Jul 2021, p. 908-917. URL : https://proceedings.mlr.press/v161/marra21a.html.

[35] George A. MILLER. « WordNet ». In : *Communications of the ACM* 38 (11 nov. 1995), p. 39-41. ISSN : 15577317. DOI : 10.1145/219717.219748. URL : https://dl.acm.org/doi/10.1145/219717.219748.

[36] Bruce R. MULLER et W. SMITH. « A Hierarchical Loss for Semantic Segmentation ». In : *VISIGRAPP*. 2020. URL : https://api.semanticscholar.org/CorpusID:215791996.

[37] Mathias NIEPERT, Pasquale MINERVINI et Luca FRANCESCHI. « Implicit MLE : Backpropagating Through Discrete Exponential Family Distributions ». In : *Advances in Neural Information Processing Systems*. T. 34. Curran Associates, Inc., 2021, p. 14567-14579. URL : https://proceedings.neurips.cc/paper_files/paper/2021/hash/7a430339c10c642c4b2251756fd1b484-Abstract.html (visité le 17/01/2024).

[38] Marin Vlastelica POGANČIĆ et al. « Differentiation of Blackbox Combinatorial Solvers ». en. In : sept. 2019. URL : https://openreview.net/forum?id=BkevoJSYPB (visité le 27/10/2023).

[39] Laura von RUEDEN et al. « Informed Machine Learning – A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems ». In : *IEEE Transactions on Knowledge and Data Engineering* 35.1 (jan. 2023). Conference Name : IEEE Transactions on Knowledge and Data Engineering, p. 614-633. ISSN : 1558-2191. DOI : 10.1109/TKDE.2021.3079836.

[40] Olga RUSSAKOVSKY et al. « ImageNet Large Scale Visual Recognition Challenge ». In : *International Journal of Computer Vision* 115 (3 déc. 2015), p. 211-252. ISSN : 15731405. DOI : 10.1007/s11263-015-0816-y.

[41] Stuart RUSSELL et Peter NORVIG. *Artificial Intelligence A Modern Approach (4th Edition)*. Pearson Higher Ed, 2021. Chap. 7, p. 208-250.

[42] Leslie G. VALIANT. « The Complexity of Enumeration and Reliability Problems ». In : *SIAM Journal on Computing* 8.3 (1er août 1979), p. 410-421. ISSN : 0097-5397. DOI : 10.1137/0208032. URL : https://doi.org/10.1137/0208032 (visité le 21/02/2024).

[43] M J WAINWRIGHT et al. « Graphical Models, Exponential Families, and Variational Inference ». In : *Foundations and Trends R in Machine Learning* 1.2 (2008), p. 1-305. DOI : 10.1561/2200000001. (Visité le 07/04/2022).

[44] Wenguan WANG, Yi YANG et Fei WU. *Towards Data-and Knowledge-Driven Artificial Intelligence : A Survey on Neuro-Symbolic Computing*. 2023. arXiv : 2210.15889 [cs.AI]. URL : https://arxiv.org/abs/2210.15889.

[45] Jingyi XU et al. « A Semantic Loss Function for Deep Learning with Symbolic Knowledge ». In : *35th International Conference on Machine Learning, ICML 2018*. T. 12. International Machine Learning Society (IMLS), 2018, p. 8752-8760. ISBN : 9781510867963.

[46] Zhun YANG, Adam ISHAY et Joohyung LEE. « NeurASP : Embracing Neural Networks into Answer Set Programming ». In : *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. California : International Joint Conferences on Artificial Intelligence Organization, juill. 2020, p. 1755-1762. ISBN : 978-0-9992411-6-5. DOI : 10.24963/ijcai.2020/243.

Un cadre pour la planification consciente d'un observateur sous observabilité partielle

Salomé Lepers, Vincent Thomas, Olivier Buffet

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

prénom.nom@loria.fr

Résumé

Dans cet article, nous nous intéressons à des problèmes de planification où l'agent est conscient de la présence d'un observateur et où cet observateur est en situation d'observabilité partielle. L'agent doit donc choisir sa stratégie dans le but d'optimiser les informations qu'il transmet à travers les observations. Nous proposons un cadre qui permet de traiter ce type de problème et de travailler avec différentes propriétés telles que la prédictibilité, la lisibilité et l'explicabilité. Notre travail s'appuie sur le cadre des processus de décision markoviens conscients d'un observateur (OAMDP). Étendre les OAMDP en observabilité partielle permet d'une part de travailler sur des problèmes plus réalistes (des situations où l'observateur n'aurait pas accès à l'ensemble des données de l'environnement), mais permet aussi de considérer des variables cibles dynamiques. Ces types dynamiques permettent de traiter la prédictibilité telle que présentée dans les pOAMDP (predictable OAMDP) ainsi que des problèmes de lisibilité à objectifs multiples où l'objectif de l'agent pourrait changer au cours du temps.

Mots-clés

Planification probabiliste, observabilité partielle, explicabilité, prédictibilité, lisibilité

Abstract

In this article, we are interested in planning problems where the agent is aware of the presence of an observer, and where this observer is in a partial observability situation. The agent has to choose its strategy to optimize the information transmitted by observations. We build a framework to handle those kinds of problems and work with various properties such as predictability, legibility and explicability. Our work is based on the Observer Aware Markov Decision Process (OAMDP) framework. The extension of OAMDPs to partial observability can handle more realistic problems (situations where the observer doesn't have access to all of the environment information) but also allows to consider dynamic types. Those dynamic target variables allow to work with predictability as presented in the pOAMDP (predictable OAMDP) framework and with legibility problems with multiple goals where the goal might change during the task.

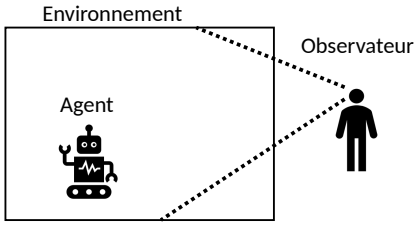


FIGURE 1 : Agent dans son environnement en présence d'un observateur passif

Keywords

Probabilistic planning, partial observability, explicability, predictability, legibility

1 Introduction

Dans des situations de collaboration homme-robot, certaines propriétés du comportement du robot peuvent être appréciées de l'humain, voire permettre une meilleure collaboration. Divers travaux récents ont porté sur l'obtention automatique de comportements dotés de telles propriétés, en particulier dans le cas où l'humain ne fait qu'observer l'agent dans son environnement, et où l'agent, conscient de cet observateur, cherche à adopter un comportement qui permette de contrôler au mieux les informations acquises par l'humain (cf. figure 1).

CHAKRABORTI, KULKARNI, SREEDHARAN et al. [1] proposent une taxonomie des différents concepts rencontrés dans ces travaux, certains cherchant 1. à transmettre de l'information, tels que la *lisibilité* (lorsque l'agent essaye de communiquer son but à travers ses choix d'actions), l'*explicabilité* (un comportement explicable est conforme aux attentes de l'observateur), et la *predictabilité* (un comportement est prédictible si il est facile de deviner la fin d'une trajectoire en cours), ou 2. d'autres à cacher de l'information, par exemple l'*obscurcissement*, quand le comportement vise à cacher la tâche réelle de l'agent. Ils formalisent aussi ces différents problèmes de manière unifiée sous l'hypothèse que les transitions sont déterministes, raisonnant donc principalement sur des plans (une séquence d'actions induisant une unique séquence d'états). Dans leur approche, le robot modélise l'humain comme ayant un cer-

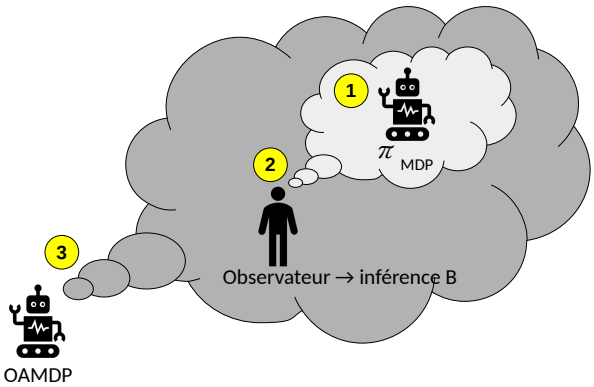


FIGURE 2 : Un agent OAMDP (3) fait l'hypothèse que l'observateur s'attend (2) à ce que l'agent se comporte de manière à accomplir une certaine tâche (1).

tain modèle du système robot+environnement (y compris de la ou les tâches possibles du robot), et pouvant ainsi anticiper les comportements possibles du robot. Chacune de ces propriétés peut être intéressante dans certaines situations et transmet différentes informations à l'observateur.

MIURA et ZILBERSTEIN [2], pour leur part, proposent un formalisme générique analogue (voir figure 2), mais sous l'hypothèse de transitions stochastiques, d'où le nom de *processus de décision markovien conscient d'un observateur* (OAMDP pour *observer-aware Markov decision process*). Ils font l'hypothèse que, du point de vue de l'observateur, l'agent effectue sa tâche en ignorant la présence de l'observateur. C'est une hypothèse réaliste dans un grand nombre de contextes. En outre, faire l'hypothèse contraire (l'observateur suppose que l'agent essaye d'aider son inférence) induirait un problème de poule et d'œuf, les deux cherchant à se modéliser l'un l'autre. Entre autres choses, ils travaillent aussi sur l'explicabilité, la lisibilité et la prédictibilité.

LEPERS, THOMAS et BUFFET [3] ont plus récemment proposé une nouvelle façon de modéliser la prédictibilité en s'inspirant du cadre OAMDP, et en proposant une approche plus adaptée aux environnements incertains. La variable cible n'étant plus un type statique, mais la prochaine action ou le prochain état de l'agent, donc une variable dynamique, il a fallu introduire un nouveau cadre, celui des pOAMDP (predictable OAMDP). L'objectif de cet article est de proposer un modèle qui permette de traiter à la fois des problèmes avec un type statique (lisibilité, explicabilité pour les OAMDP), des problèmes avec des variables cibles dynamiques (prédictibilité des pOAMDP) et des problèmes en observabilité partielle. Dans cette dernière situation, l'observateur n'a alors plus forcement accès à l'état et à l'action de l'agent mais à une observation liée à la transition suivie par le système. L'introduction d'observabilité partielle permet de travailler sur des problèmes plus divers et plus proches de la réalité tels que des situations où l'observateur n'aurait pas accès à l'ensemble des informations de l'environnement. Nous pouvons par exemple considérer des situations où l'agent PO-OAMDP n'est pas toujours visible et doit choisir d'utiliser certains passages pour être vu par l'observateur et lui permettre de mieux inférer la situation courante.

La section 2 introduit des pré-requis sur le processus de décision markoviens (MDP) et les MDP conscients d'un observateur. Notre approche général est décrite en section 3, la résolution des PO-OAMDP est décrite en section section 4 avant de conclure en section 5.

2 Pré-requis

2.1 Processus de décision markovien

Un *processus de décision markovien* (MDP) est un 6-uplet $\langle \mathcal{S}, \mathcal{A}, T, R, \gamma, \mathcal{S}_f \rangle$ où :

- $\mathcal{S}$ est l'ensemble des états ;
- $\mathcal{A}$ est l'ensemble des actions ;
- $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0; 1]$, la fonction de transition, donne la probabilité $T(s, a, s')$ d'aller dans un état s' depuis un état s en exécutant l'action a ;
- $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, la fonction de récompense, donne la récompense reçue $R(s, a, s')$ lors d'une transition (s, a, s') ;
- $\gamma \in [0, 1]$ est le facteur d'actualisation ; et
- $\mathcal{S}_f \subset \mathcal{S}$ est l'ensemble des états terminaux : pour tout $s, a \in \mathcal{S}_f \times \mathcal{A}$, $T(s, a, s) = 1$ et $R(s, a, s) = 0$.

Une politique $\pi_{\text{OBS}} : \mathcal{S} \to \Delta(\mathcal{A})$ détermine un comportement en associant à chaque état une action à effectuer. Elle peut éventuellement être stochastique, $\pi_{\text{OBS}}(a|s)$ étant alors la probabilité d'effectuer a dans l'état s. Considérant un *MDP actualisé*, c'est-à-dire tel que $\gamma < 1$, la valeur d'une politique π_{OBS} en un état s est l'espérance de la somme des récompenses actualisées sur un horizon infini :

$$V^{\pi_{\text{OBS}}}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\pi_{\text{OBS}}} \left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t) | S_0 = s \right].$$

Il existe toujours au moins une politique $\pi_{\text{OBS}}{}^*$, dite optimale, telle que, pour tout s, $V^{\pi_{\text{OBS}}*}(s) = \max_{\pi_{\text{OBS}}} V^{\pi_{\text{OBS}}}(s)$. L'algorithme d'*itération sur la valeur* (VI) calcule cette fonction de valeur optimale, notée V^*, en itérant le calcul suivant jusqu'à atteindre une précision suffisante (où k désigne l'itération courante) :

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s, a, s') \cdot (R(s, a, s') + \gamma V_k(s')).$$

On interrompt les calculs quand le *résidu de Bellman* est inférieur à un seuil fonction de l'erreur ϵ souhaitée et de γ :

$$\underbrace{\max_s |V_{k+1}(s) - V_k(s)|}_{\text{résidu de Bellman}} \leq \frac{1-\gamma}{\gamma}\epsilon,$$

une politique déterministe ϵ-optimale étant alors obtenue en agissant de "manière gourmande" dans tout état s avec :

$$\pi_{\text{OBS}}{}^*(s) \leftarrow \arg\max_a \sum_{s'} T(s,a,s') \cdot (R(s,a,s') + \gamma V^*(s')) \,.$$

Les propriétés ci-dessus restent valident avec $\gamma = 1$ si

1. $\mathcal{S}_f$ non vide ; et
2. R est telle qu'il existe des politiques atteignant $\mathcal{S}_f$ avec probabilité 1 depuis tout état s, et que la valeur des autres politiques diverge vers $-\infty$ dans les états depuis lesquels on ne peut pas être sûr de pouvoir atteindre un état terminal.

On parle alors de problème de type *chemin stochastique le plus court* (SSP). On a un SSP en particulier, si, pour tout $(s,a,s') \in (\mathcal{S} \setminus \mathcal{S}_f) \times \mathcal{A} \times \mathcal{S}$, on a $r(s,a,s') < 0$, c'est-à-dire si on cherche à atteindre un état terminal à "moindre coût" (en moyenne).
Note : On peut transformer tout MDP actualisé en un SSP dans lequel, à chaque instant, on a une probabilité $1-\gamma$ de transiter vers un état terminal. Le cas SSP est donc plus général.

2.2 Processus de décision markovien conscient d'un observateur

Un *MDP conscient d'un observateur* (OA-MDP pour *observer-aware MDP*) décrit une situation dans laquelle un agent interagit avec son environnement en ayant conscience de la présence d'un observateur, et en cherchant à maximiser un critère de performance lié aux croyances de cet observateur. Il est défini formellement par un 8-uplet $\langle \mathcal{S}, \mathcal{A}, T, \gamma, \mathcal{S}_f, \Theta, B, R \rangle$ où :

- $\langle \mathcal{S}, \mathcal{A}, T, \gamma, \mathcal{S}_f \rangle$ est un MDP sans fonction de récompense ;
- Θ est un ensemble fini de *types* possibles de l'agent, représentant une caractéristique de celui-ci telle que sa tâche réelle ou ses capacités ;
- $B : H^* \to \Delta^{|\Theta|}$ donne la croyance que l'observateur a sur le type de l'agent (la *croyance* sur une variable aléatoire est la distribution de probabilité sur ses valeurs possibles étant données les informations disponibles) en fonction de l'historique des états et des actions ($H \stackrel{\text{def}}{=} \mathcal{S} \times \mathcal{A}$) ;
- $R : \mathcal{S} \times \mathcal{A} \times \Delta^{|\Theta|} \to \mathbb{R}$ est la fonction de récompense.

Dans la plupart des cas considérés par MIURA et ZILBERSTEIN, B est obtenue en s'appuyant sur la définition de la mise-à-jour de croyance bayésienne BST de BAKER, SAXE et TENENBAUM, c'est-à-dire en considérant que, du point de vue de l'agent, l'observateur modélise le comportement de l'agent pour une tâche donnée à travers un MDP :

1. en utilisant une fonction de récompense R_{OBS} appropriée ;
2. en résolvant le MDP $\langle \mathcal{S}, \mathcal{A}, T, \gamma, \mathcal{S}_f \rangle$ (où tous les composants, exceptée la fonction de récompense R_{OBS}, émanent de la définition de l'OAMDP) pour obtenir V^*_{OBS} ; et
3. en construisant une politique "softmax", c'est-à-dire telle que, pour chaque couple (s,a),

$$\pi_{\text{OBS}}(a|s) = \frac{e^{\frac{1}{\tau} Q^*_{\text{OBS}}(s,a)}}{\sum_{a'} e^{\frac{1}{\tau} Q^*_{\text{OBS}}(s,a')}}, \text{ où}$$
$$Q^*_{\text{OBS}}(s,a) = \sum_{s'} T(s,a,s') \cdot (r(s,a,s') + \gamma V^*_{\text{OBS}}(s')) \,,$$

$\tau > 0$ représentant le niveau de rationalité de l'agent (considéré par l'observateur) afin de pouvoir raisonner sur des politiques plus ou moins proches de la politique optimale.

La croyance de l'observateur sur les types peut ensuite être obtenue par inférence bayésienne en utilisant π_{OBS}.
MIURA et ZILBERSTEIN formalisent ainsi, entre autres, des problèmes de lisibilité, d'explicabilité, et de prédictibilité.
Note : Comme déjà fait ci-dessus, on indicera souvent par "OBS" les quantités liées au point de vue de l'observateur (tel que perçu par l'agent). Entre autres, certaines probabilités seront calculées du point de vue de l'observateur, et notées P_{OBS}. Aussi, on écrira parfois une fonction $f(X,Y)$ décrivant une distribution de probabilité conditionnelle sous la forme $f(Y|X)$ pour faire ressortir les dépendances entre variables.

3 Contribution : MDP conscient d'un observateur en observabilité partielle

Comme vu en introduction, on souhaite proposer un modèle OAMDP en observabilité partielle, lequel permettrait à la fois de traiter plus de scénarios (situations où l'observateur ne voit pas toujours le robot) et traiter des propriétés qui nécessitent l'utilisation de variables cibles dynamiques.

3.1 Formalisme

Dans le cadre PO-OAMDP, l'agent a accès à l'état complet du système, et l'observateur n'en a désormais qu'une perception partielle. Sa construction part de l'ajout au formalisme OAMDP d'un ensemble d'observations et d'une fonction d'observation. Nous allons expliquer cette construction avant de donner une définition formelle. Dans ce contexte, le type est désormais nommé *variable cible* et peut évoluer au cours du temps, contrairement au type statique des OAMDP. Afin de rester souple et générique dans la définition de ce qu'est la variable cible, sa valeur à chaque pas de temps est le résultat d'une fonction prenant en entrée la transition suivie par le système. La variable cible peut donc être juste une sous-partie de l'état du système (par exemple une variable non observable par l'observateur), mais elle peut aussi être liée à l'action émise par l'agent (pour des problèmes de prédictibilité) ou à l'évolution de l'état plus qu'à l'état lui même. Évidemment, cette

variable cible peut regrouper en son sein plusieurs variables différentes. Nous ne parlons que d'une unique variable que par commodité mais sans perte de généralité.
En outre, on suppose que, en plus de l'état complet du système, l'agent a accès aux observations reçues par l'observateur (ce qui reste réaliste dans de nombreuses situations, en particulier si le processus d'observation est déterministe, auquel cas les observations reçues par l'observateur sont facilement prédictibles). L'agent peut ainsi construire l'état interne de l'observateur au fur et à mesure de l'exécution de son comportement.
En ayant accès à toutes les informations du problème (l'état du système, les actions effectuées et les observations perçues par l'observateur), l'agent va planifier ses actions pour chercher à contrôler l'inférence faite par l'observateur sur sa variable cible.
Formellement, un PO-OAMDP est ainsi défini par un n-uplet $\langle \mathcal{S}, \mathcal{A}, T, \gamma, \mathcal{S}_f, R_{\text{OBS}}, \Theta, \Omega, O, B, R_{\text{AG}}, \phi \rangle$, où :

- $\langle \mathcal{S}, \mathcal{A}, T, \gamma, \mathcal{S}_f, R_{\text{OBS}} \rangle$ est un MDP ;
- Θ désigne une *variable cible* (dynamique), mais aussi l'ensemble fini de valeurs qu'elle peut prendre ; nous changeons de terminologie pour souligner la différence entre cette variable (dynamique) et la variable *type* des OAMDP ;
- $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \Theta$ est une fonction qui donne la valeur de la variable cible en fonction de la transition : $\theta_t = \phi(s_t, a_t, s_{t+1})$;
- Ω est un ensemble fini d'observations ;
- $O : \mathcal{A} \times \mathcal{S} \to \Omega$ est la fonction d'observation ; $O(a, s', o)$ est la probabilité d'émission d'une observation o si l'action a conduit dans l'état s' ;
- $B : \Omega^* \to \Delta^{|\mathcal{S}|}$ donne la croyance que l'observateur a sur l'état en fonction de l'historique des observations ; la croyance sur la variable cible pourra en être déduite (voir section 3.2) ; on notera $b_t \stackrel{\text{def}}{=} B(o_1, \ldots, o_{t-1})$, en appelant croyance initiale $b_0 \stackrel{\text{def}}{=} B()$;
- $R_{\text{AG}} : \mathcal{S} \times \Delta^{|\Theta|} \times \mathcal{A} \times \mathcal{S} \times \Delta^{|\Theta|} \to \mathbb{R}$ est la fonction de récompense de l'agent sous sa forme la plus générale : $R_{\text{AG}}(s_t, \beta_t, a_t, s_{t+1}, \beta_{t+1})$.

Alors que dans un POMDP, la mise à jour de la croyance dépend nécessairement des actions effectuées choisies par l'agent lui même (incluses dans l'historique), dans un PO-OAMDP, les actions ne sont pas forcément connues de l'observateur. Les observations peuvent inclure les actions selon le scénario considéré.
Nous ferons ici l'hypothèse que l'observation reçue permet à l'observateur de savoir à chaque instant si un état terminal a été atteint ou non, sans nécessairement indiquer duquel il s'agit (ce qui pourra dépendre du problème).
On notera que le modèle PO-OAMDP repose sur un seul MDP sous-jacent, contrairement au modèle OAMDP, qui associe un MDP à chaque type possible. Dans notre cadre à observabilité partielle, n'employer qu'un seul MDP sous-jacent n'est toutefois pas restrictif. On pourrait montrer formellement que tout OAMDP peut se ré-écrire comme un PO-OAMDP.
On notera aussi que l'introduction de la variable cible n'a rien de nécessaire. On pourrait obtenir un formalisme équivalent en écrivant une fonction de récompense directement sur la croyance sur les états au lieu de la croyance sur la variable cible. L'introduction de la variable cible est une commodité pour afficher le lien avec les OAMDP et pour faciliter la modélisation des problèmes.
La suite de cette section décrit comment la croyance de l'observateur (sur l'état) est mise à jour, et comment en déduire la croyance sur la variable cible, laquelle permet de calculer la récompense de l'agent lors d'une transition. Elle illustre ensuite les usages possibles du cadre PO-OAMDP sur différents scénarios.

3.2 Calcul des croyances sur l'état et la variable cible

Comme dans le cadre OAMDP, l'observateur calcule la politique de l'agent étant donnée la fonction de récompense qu'il connaît, R_{OBS}. L'observateur modélise le comportement de l'agent pour une tâche donnée à travers un MDP :

1. en utilisant une fonction de récompense R_{OBS} ;
2. en résolvant le MDP sous-jacent ; et
3. en construisant une politique softmax.

On peut noter que, étant données la dynamique (transitions+observations) du PO-OAMDP et la politique π_{OBS} de l'agent, l'observateur fait face à un HMM : il résout un problème de *filtrage* en devant estimer la croyance sur l'état s_t en fonction de l'historique d'observation $o_{1:t}$.
La croyance de l'agent peut ensuite être construite à partir de la politique de l'agent, de la fonction de transition du MDP sous-jacent et de la fonction d'observation :

$$
\begin{aligned}
B(s_{t+1}|o_{1:t+1}) = P(s_{t+1}|o_{1:t+1}) &= \frac{P(s_{t+1}, o_{1:t}, o_{t+1})}{P(o_{1:t+1})} \\
&= \frac{P(s_{t+1}, o_{1:t+1})}{\sum_{s_{t+1}} P(s_{t+1}, o_{1:t+1})} \\
&= \frac{K(s_{t+1}, o_{1:t+1})}{\sum_{s_{t+1}} K(s_{t+1}, o_{1:t+1})}, \text{ avec} \\
K(s_{t+1}, o_{1:t+1}) \stackrel{\text{def}}{=} \sum_{a_t} O(o_{t+1}|a_t, s_{t+1}) &\sum_{s_t} T(s_{t+1}|s_t, a_t) \cdot \\
&\pi_{\text{OBS}}(a_t|s_t) \cdot B(s_t|o_{1:t}).
\end{aligned}
$$

Croyance sur la variable cible Pour déterminer la récompense reçue lors d'une transition, il est aussi nécessaire de calculer la croyance β sur la valeur que va prendre la variable cible : $\Theta_t = \phi(S_t, A_t, S_{t+1})$. Cela peut être réalisé en partant de la croyance b de l'observateur sur l'état cou-

rant, S_t, comme suit :

$$\begin{aligned}\beta(\theta) &= \sum_{s,a,s'} \mathbb{1}_{\theta=\phi(s,a,s')} \cdot P_{obs}(s,a,s'|b) \\ &= \sum_{s,a,s'} \mathbb{1}_{\theta=\phi(s,a,s')} \cdot P_{obs}(s'|s,a) \cdot P_{obs}(a|s) \cdot P_{obs}(s|b) \\ &= \sum_{s,a,s'} \mathbb{1}_{\theta=\phi(s,a,s')} \cdot T(s,a,s') \cdot \pi_{\text{OBS}}(a|s) \cdot b(s). \qquad (1)\end{aligned}$$

3.3 Mise en œuvre sur divers scénarios

Le modèle PO-OAMDP permet de construire différents comportements en faisant varier Θ et R et donc d'aborder différents types de problèmes.
Nous allons commencer par montrer comment des OAMDP peuvent être reformulés comme des PO-OAMDP. Dans un OAMDP, l'agent a un *type* statique qui le caractérise. Dans le cadre PO-OAMDP, cela peut se traduire par une variable d'état (cachée) dont la valeur est extraite par $\theta = \phi(s)$.

3.3.1 Expression de B et R' pour différentes propriétés

Lisibilité La lisibilité réduit l'ambiguïté sur les buts possibles de l'agent. Dans cette situation, l'agent a plusieurs buts possibles inconnus de l'observateur. Un comportement lisible transmet le but (ou, plus généralement, le critère de performance) de l'agent à travers ses choix d'actions.

Θ **:** Dans le cas de la propriété de lisibilité, le type va donc caractériser ce critère de performance parmi un ensemble fini de critères possibles. Cela va se traduire typiquement par le fait que la fonction de récompense dépend de ce type, mais pas nécessairement la fonction de transition ou la fonction d'observation.

R_{AG} **:** Pour la fonction de récompense, MIURA et ZILBERSTEIN utilisent l'opposé de la distance euclidienne à la "croyance idéale". La croyance idéale étant définie par : $\beta^*(s) = (0,\dots,0,1,0,\dots,0)$ (avec un 1 en composante $\theta = \phi(s)$), on a alors :

$$R_{\text{AG}}(s,\beta,a,s',\beta') \stackrel{\text{def}}{=} -\sqrt{\|\beta - \beta^*(s)\|_2}.$$

Explicabilité Un comportement explicable est un comportement cohérent avec les attentes de l'observateur.

Θ **:** Pour traduire cette idée, MIURA et ZILBERSTEIN (suivant SREEDHARAN, KULKARNI, CHAKRABORTI et al. [5]) proposent de minimiser la probabilité que le comportement observé soit celui d'un comportement aléatoire, même si plusieurs autres comportements restent probables. Ils introduisent ainsi un type "virtuel" θ_0 qui représente un comportement (une politique) aléatoire en plus des autres types (réels).

R_{AG} **:** Pour traduire le critère d'explicabilité susmentionné, on va prendre

$$R_{\text{AG}}(s,\beta,a,s'\beta') \stackrel{\text{def}}{=} -\beta(\theta_0).$$

Prédictibilité Un comportement prédictible est un comportement dont la fin de trajectoire est plus facile à prédire pour l'observateur. On propose ici une définition de la prédictibilité inspirée des travaux de LEPERS, THOMAS et BUFFET [3], mais mieux fondée d'un point de vue théorique. On essaye donc de prédire soit l'action, soit l'état de l'agent,
Si nous partons comme MIURA et ZILBERSTEIN et d'autres de cette définition, nous nous inspirons plutôt des travaux de LEPERS, THOMAS et BUFFET [3], lesquels sont plus adaptés aux problèmes à dynamique stochastique, mais en faisant ici une proposition dont la sémantique est plus claire.

Θ **:** L'idée de départ est que l'observateur cherche, à chaque instant, à prédire la prochaine action ou le prochain état, d'où deux types de prédictibilité différents. Pour la prédictibilité sur l'action, on pose $\Theta = A$ et $\phi(s,a,s') = a$. Pour la prédictibilité sur l'état, on pose $\Theta = S$ et $\phi(s,a,s') = s'$. Dans les deux cas, pour agir optimalement, l'observateur doit parier sur une des prochaines valeurs cibles les plus probables, et donc choisir une valeur dans l'ensemble

$$\psi_\Theta(\beta_t) \stackrel{\text{def}}{=} \arg\max_\theta \beta_t(\theta).$$

R_{AG} **:** On considère que l'observateur échantillonne sa prédiction de façon uniforme, on peut alors définir :

$$\text{pred}(\theta|\beta_t) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{|\psi_\Theta(\beta_t)|} & \text{si } \theta \in \psi_\Theta(\beta_t), et \\ 0 & \text{sinon.} \end{cases}$$

En définissant

$$R_{\text{AG}}(s,\beta,a,s',\beta') \stackrel{\text{def}}{=} \begin{cases} \text{pred}(a|\beta) - 1 & \text{si } \Theta = A, \\ \text{pred}(s'|\beta) - 1 & \text{si } \Theta = S, \end{cases}$$

la récompense immédiate est l'opposé de la probabilité que le pari (d'un observateur rationnel) échoue : $R_{\text{AG}}(s,\beta,a,s',\beta') = -P(\text{pari perdu})$.

Obscurcissement La problématique inverse peut également être considérée. L'agent essaye alors de cacher des informations telles que son but à l'observateur. Dans cette situation, l'agent a plusieurs buts possibles et essaye de ne pas révéler son "vrai" but à l'observateur. L'obscurcissement avec le modèle PO-OAMDP présente les mêmes difficultés que celles rencontrées par le modèle OAMDP :

- si l'objectif ne porte que sur l'obscurcissement, mais pas sur la réalisation de la tâche, l'agent peut simplement ne rien faire pour dissimuler son but, et
- pour construire la croyance de l'observateur sur les buts, on fait l'hypothèse que l'observateur ne sait pas qu'on essaye de le tromper.

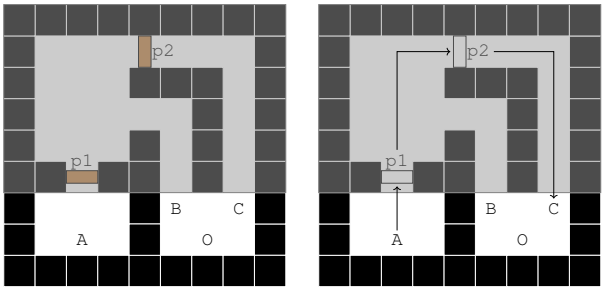


FIGURE 3 : Représentation d'un environnement avec des portes pouvant être verrouillées. Les murs sont représentés pas des cases noires, les portes par des rectangles bruns. La zone grisée correspond à une zone non visible de l'observateur. L'agent débute en A et doit se diriger en O.

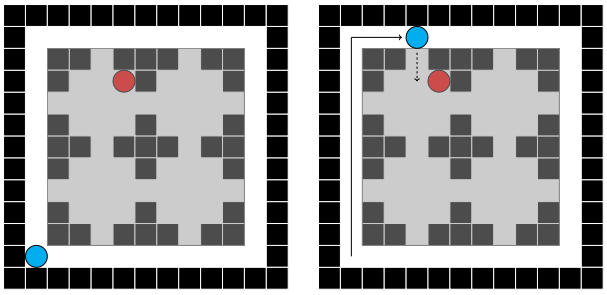


FIGURE 4 : Environnement à grille avec un intrus dont l'observateur cherche à connaître la localisation. Les murs sont représentés par des cases noires. La zone grisée correspond à une zone non visible de l'observateur. L'intrus est représenté par un cercle rouge et l'agent par un cercle bleu.

3.3.2 Élargissement des types de problèmes couverts

Les sections précédentes ont montré comment étendre les problèmes déjà modélisés dans les cadres OAMDP et p-OAMDP en considérant l'observabilité partielle de l'observateur. Il faut cependant noter que les PO-OAMDP permettent la formalisation de nouveaux problèmes dans lesquels l'agent ne va pas simplement chercher à exhiber un comportement prédictible, lisible ou explicable, mais pourra chercher à transmettre au mieux de l'information sur l'état du monde partiellement observé par l'observateur.

Scénario 1 : Si on considère un environnement de type bureau (cf. figure 3) avec des portes soit ouvertes, soit fermées à clef, on peut imaginer un agent cherchant à faire comprendre à un observateur extérieur l'état des portes à travers ses actions. Cela peut bien entendu se faire en ouvrant des portes visibles pour l'observateur mais aussi en se montrant dans certaines zones qui ne peuvent être atteintes que par l'ouverture de certaines portes. Ainsi, même si ces portes ne sont jamais vues par l'observateur, la présence de l'agent peut lui permettre d'inférer que certaines portes sont ouvertes. Dans l'exemple représenté en figure 3, en choisissant un chemin plus long dans la zone cachée et en redevenant visible en C, il informe l'observateur que les portes p1 et p2 ne sont pas verrouillées. Se rendre visible en B permettrait d'atteindre l'objectif et augmenterait un peu la probabilité que la porte p2 soit fermée. Résoudre ce problème nécessite que l'agent raisonne sur 1. les conséquences de ses actions, 2. sa visibilité en fonction de sa localisation, 3. les inférences que pourra faire l'observateur et 4. les portes dont l'observateur cherche à estimer l'état.

Scénario 2 : Dans une seconde situation, on peut considérer un agent en charge de détecter des intrus dans un environnement inaccessible pour l'observateur (cf. figure 4). En utilisant un modèle que l'observateur a de son comportement, cet agent peut tirer parti des attentes de l'observateur pour agir, se rendre visible à certains endroits et faire comprendre à l'observateur la présence effective d'un intrus et sa localisation. Dans l'exemple illustré par la figure 4, l'observateur estime que l'agent cherche à se rapprocher de l'intrus. L'agent peut ainsi informer l'observateur de la position de l'intrus en choisissant, parmi les trajectoires possibles l'amenant au plus proche de l'intrus, un chemin souvent visible de l'observateur.

Scénario 3 : Enfin, dans des tâches complexes qui requièrent plusieurs étapes intermédiaires, l'agent peut chercher à transmettre l'état de la tâche en cours en empruntant des chemins plus longs mais 1. qui sont partiellement visibles par l'observateur et 2. qui laissent moins d'ambiguïté sur son objectif intermédiaire.

En cherchant à faciliter l'inférence des objectifs intermédiaires qu'il cherche à atteindre, l'agent peut ainsi transmettre l'état de la tâche à réaliser, ce qui peut s'avérer crucial dans une situation collaborative (qui attendrait en retour une action particulière de l'humain).
Ces différentes situations montrent que le cadre des OAMDP permet d'élargir les problèmes à d'autres types d'échanges d'information que simplement des informations sur le comportement de l'agent lui-même. Le cadre permet de modéliser des problèmes proches des problèmes de recherche active d'information tels que formalisés par les ρ-POMDP [6]. Dans le cadre ρ-POMDP, un agent est dans un environnement partiellement observé et doit agir au mieux pour acquérir des observations pertinentes et maximiser une mesure d'information sur des variables cibles (par exemple sa localisation). La principale différence avec le cadre PO-OAMDP est que, dans ce dernier, c'est l'information acquise par un tiers (l'observateur) que l'agent cherche à contrôler, pas la sienne propre (qui est complète). Cela nécessite en particulier un modèle de l'observateur dont l'agent va tirer parti pour chercher à contrôler indirectement les croyances de l'observateur.

4 Résolution

Nous allons maintenant discuter des approches possibles pour la résolution de PO-OAMDP, c'est-à-dire de politiques maximisant (au moins à $\epsilon > 0$ près) la somme des récompenses atténuées. On fait l'hypothèse dans cette section qu'on dispose d'un couple état-croyance initial, les algo-

rithmes discutés en faisant tous usage.
Dans la suite, on considère d'abord les solutions prenant la forme naturelle de politiques dépendant de l'historique, avant de voir que l'on peut aussi raisonner avec des croyances à la place de ces historiques.

4.1 Recherche d'une politique historique-dépendante

4.1.1 Données pertinentes pour la politique

Au premier abord, à un instant donné, le choix d'action de l'agent dépend au plus des données spécifiques à sa trajectoire actuelle, c'est-à-dire l'historique des états, actions et observations.
On a toutefois besoin pour prendre des décisions que d'informations nécessaires à la prédiction des récompenses. Or, par définition, la récompense reçue à un instant t dépend de la transition $s_t, b_t, a_t, s_{t+1}, b_{t+1}$ (la croyance sur la variable cible se déduisant de la croyance sur l'état comme on l'a déjà vu). Comme, dans ce tuple, 1. l'état s_t évolue de manière markovienne (indépendamment de l'historique d'états et d'actions antérieurs), et 2. la croyance sur l'état b_t ne dépend que des observations passées, alors le choix d'action ne dépend que de l'état courant s_t et de l'historique d'observations $o_1, \ldots, o_{t-1}$. On va donc pouvoir ne considérer que les politiques de la forme $\pi_{\text{AG}} : \mathcal{S} \times (\Omega)^* \to \mathcal{A}$ (en notant que, dans cet espace "d'états", parmi les politiques optimales, certaines sont déterministes).

4.1.2 Résolution

Dans le cas $\gamma < 1$**,** étant donné $\epsilon > 0$, on peut trouver une solution ϵ-optimale en se ramenant à un problème à horizon fini et en employant un algorithme tel que la programmation dynamique, AO* [7] ou MCTS [8] (comme l'ont fait MIURA et ZILBERSTEIN [2] pour les OAMDP). L'opérateur d'optimalité de Bellman va alors s'écrire, pour $t < H - 1$,

$$V_t^*(s, o_{1:t}) = \max_a \sum_{s', o_{t+1}} T(s'|a, s) \cdot O(o_{t+1}|s', a) \cdot [R'(s, b_{[o_{1:t}]}, a, s', b_{[o_{1:t+1}]}) + \gamma V_{t+1}^*(s', b_{[o_{1:t+1}]})],$$

et

$$V_H^*(s, o_{1:H}) = 0.$$

Dans le cas $\gamma = 1$**,** sauf cas particulier, on ne peut pas se ramener à un horizon temporel fini. Les politiques historique-dépendantes ne paraissent donc pas adaptées.

4.2 Recherche d'une politique croyance-dépendante

Nous allons voir ici que la résolution d'un PO-OAMDP est équivalente à celle d'un MDP dans un espace continu particulier. Cela va permettre d'envisager l'emploi de politiques croyances-dépendantes.

4.2.1 Belief-MDP équivalent

Statistique suffisante Nous avons vu précédemment comment calculer l'état de croyance de l'observateur sur l'état, au vu de la séquence d'observations qu'il a reçues, par filtrage bayésien. En fait, l'état d'information (s, b) (l'état courant couplé avec la croyance de l'observateur sur l'état) constitue une statistique suffisante pour la planification puisqu'elle

1. est markovienne (on peut prédire son évolution sans avoir recours à des informations antérieures) puisque l'état s est markovien par définition, et la croyance b l'est aussi d'après nos calculs (elle peut être mise à jour en ne connaissant que la dernière observation reçue) ; et

2. permet d'estimer la récompense à chaque pas de temps, par définition de la fonction de récompense dans un PO-OAMDP.

Formalisation du belief-MDP On obtient donc un MDP valide $\langle \mathcal{I}, \mathcal{A}, T', R', \gamma, \mathcal{I}_f \rangle$, où :

- $\mathcal{I} \stackrel{\text{def}}{=} \mathcal{S} \times B$ est l'ensemble des états ;
- $\mathcal{A}$ est l'ensemble des actions, identique à l'ensemble des actions du PO-OAMDP ;
- $T' : \mathcal{I} \times \mathcal{A} \times \mathcal{I} \to [0; 1]$ est une nouvelle fonction de transition (voir plus bas) ;
- $R' : \mathcal{I} \times \mathcal{A} \times \mathcal{I} \to \mathbb{R}$ est une nouvelle fonction de récompense (voir plus bas) ;
- $\gamma \in [0, 1]$ est le facteur d'actualisation ; et
- $\mathcal{I}_f \subset \mathcal{I}$ est l'ensemble des éléments $\iota \equiv (s, b)$ de $\mathcal{I}$ tels que $s \in \mathcal{S}_f$; on pourra vérifier ultérieurement que, pour tout $\iota, a \in \mathcal{I} \times \mathcal{A}$, $T(\iota, a, \iota) = 1$ et $R'(\iota, a, \iota) = 0$.

Fonction de transition T' **:** La fonction de transition du belief-MDP est définie par :

$$T'(\iota'|a, \iota) \stackrel{\text{def}}{=} T(s', b'|a, s, b) = \sum_o \mathbb{1}_{b'=B(b,o)} O(o|s', a) P(s'|a, s).$$

Fonction de récompense R' **:** La fonction de récompense du belief-MDP est définie par :

$$R'(s, b, a, s', b') \stackrel{\text{def}}{=} R_{\text{AG}}(s, \beta(b), a, s', \beta(b')),$$

où $\beta(b)$ dénote la croyance β que l'on peut dériver de b comme vu dans l'équation (1).

4.2.2 Résolution

Ce nouveau MDP est défini sur un espace d'états continu. Sauf cas particuliers, l'ensemble des états accessibles depuis l'état initial est donc infini.

Dans le cas $\gamma < 1$**,** on pourrait à nouveau se ramener à un problème à horizon fini, avec l'avantage que deux trajectoires différentes peuvent conduire à la même paire (ι, t),

ce qui permettrait de réduire la quantité de calculs. L'opérateur d'optimalité de Bellman s'écrit alors (en développant $\iota = (s, b)$)

$$\begin{aligned} V_t^*(s,b) &= \max_a \sum_{s'} \int_{b'} T(s', b'|a, s, b) \cdot \\ &\quad [R'(s, b, a, s', b') + \gamma V_{t+1}^*(s', b')] db' \\ &= \max_a \sum_{s',o} O(o|s', a) \cdot T(s'|a, s) \cdot \\ &\quad [R'(s, b, a, s', b^o) + \gamma V_{t+1}^*(s', b^o)], \end{aligned}$$

où b^o est la croyance sur l'état mise à jour après observation de o, et

$$V_H^*(s, b) = 0.$$

On peut aussi se demander si, comme dans le cadre des POMDP résolus via des bMDP, la fonction de valeur optimale a des propriétés de continuité particulières (en l'occurrence de convexité) qui permettraient d'approcher celle-ci. Des résultats préliminaires montrent toutefois qu'il peut y avoir des discontinuités sur les bords de la fonction de valeur d'une politique, autour de points dont l'état de croyance est impossible. On ne pourra donc pas retranscrire directement les approches POMDP reposant sur des approximateurs généralisants de V^* (tels que HSVI [9], PBVI [10] et SARSOP [11]). Mais des versions spécifiques tenant compte des localisations des discontinuités sont peut-être envisageables.

Dans le cas $\gamma = 1$, un premier problème est de savoir si le problème obtenu est un SSP valide. Il faudrait ainsi vérifier, par exemple, si c'est toujours le cas avec les fonctions de récompenses proposées pour les propriétés de lisibilité, explicabilité et prédictibilité.
À supposer que le SSP obtenu soit valide, on se retrouve sur des problèmes proches des POSSP (ou Goal-POMDP) pour lesquels peu de travaux ont été développés à part, par exemple, ceux de PATEK [12] ou, plus récemment, de HORÁK, BOŠANSKÝ et CHATTERJEE [13].

Approche bi-critère Un problème déjà présent dans la résolution des OAMDP est que, si l'on emploie un critère lié à une fonction de récompense "observer-aware", il est possible que la performance liée à la fonction de récompense classique du MDP sous-jacent soit fortement dégradée. Une première approche peut être de combiner linéairement deux tels critères, mais cela soulève la question de la bonne pondération de ceux-ci. Une autre approche, aussi évoquée par MIURA et ZILBERSTEIN [2], est d'optimiser un critère observer-aware sous la contrainte que le critère "classique" doit atteindre au minimum une certaine valeur [14]-[18]. Il peut alors être nécessaire que la politique optimale soit stochastique.

5 Conclusion

Nous avons introduit un nouveau formalisme, celui des OAMDP en observabilité partielle (PO-OAMDP), lequel permet de travailler sur des problèmes de lisibilité, d'explicabilité et de prédictibilité en observabilité partielle. La complexité des PO-OAMDP est au moins celle des OAMDP telle qu'étudiée par MIURA et ZILBERSTEIN [2]. Selon eux, il n'est pas nécessairement bénéfique de se ramener un POMDP pour le résoudre. Différents ensembles de variables cibles et fonctions de récompense ont été proposés pour construire des comportements avec des propriétés différentes. Ce nouveau modèle permet aussi de traiter des problèmes plus proches de la réalité où un agent agit en prenant en compte un observateur qui peut ne pas avoir accès à l'ensemble des informations de l'environnement (porte fermée, champ de vision bloqué). Nous avons également discuté de la résolution des PO-OAMDP.
Comme expliqué dans la partie résolution, nous proposons de transformer les PO-OAMDP en belief-MDP équivalents pour les résoudre avec des approches génériques comme MCTS. Une première perspective est d'étudier les performances de ces algorithmes pour résoudre les PO-OAMDP. Une seconde direction de travail est d'étudier les propriétés théoriques du modèle PO-OAMDP pour proposer des algorithmiques tirant parti des propriétés de ce cadre. Une étape sera de proposer des problèmes pertinents pour tester notre modèle et évaluer les algorithmes proposés.

Références

[1] T. CHAKRABORTI, A. KULKARNI, S. SREEDHARAN, D. E. SMITH et S. KAMBHAMPATI, "Explicability ? legibility ? predictability ? transparency ? privacy ? security ? the emerging landscape of interpretable agent behavior", in *Proceedings of the Twenty-Ninth International Conference on Automated Planning and Scheduling (ICAPS)*, 2019. adresse : https://ojs.aaai.org/index.php/ICAPS/article/view/3463.

[2] S. MIURA et S. ZILBERSTEIN, "A unifying framework for observer-aware planning and its complexity", in *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, t. 161, juil. 2021, p. 610-620. adresse : https://proceedings.mlr.press/v161/miura21a.html.

[3] S. LEPERS, V. THOMAS et O. BUFFET, "Comment rendre des comportements plus prédictibles", in *JIAF-JFPDA - Journées d'Intelligence Artificielle Fondamentale*, juil. 2023. adresse : https://hal.science/hal-04212452.

[4] C. L. BAKER, R. SAXE et J. B. TENENBAUM, "Action understanding as inverse planning", *Cognition*, t. 113, n° 3, p. 329-349, déc. 2009. DOI : 10.1016/j.cognition.2009.07.005.

[5] S. SREEDHARAN, A. KULKARNI, T. CHAKRABORTI, D. E. SMITH et S. KAMBHAMPATI, *A Bayesian account of measures of interpretability in human-AI interaction*, 2020. arXiv : 2011.10920 [cs.AI].

[6] M. ARAYA-LÓPEZ, O. BUFFET, V. THOMAS et F. CHARPILLET, "A POMDP extension with belief-dependent rewards", in *Advances in Neural Information Processing Systems 23*, Vancouver, Canada, 2010.

[7] N. NILSSON, *Principles of Artificial Intelligence*. Morgan Kaufmann Publishers, 1980.

[8] L. KOCSIS et C. SZEPESVARI, "Bandit based Monte-Carlo planning", in *Proceedings of the Sixteenth European Conference on Machine Learning*, 2006.

[9] T. SMITH et R. G. SIMMONS, "Point-based POMDP algorithms : improved analysis and implementation", in *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, 2005, p. 542-549.

[10] J. PINEAU, G. GORDON et S. THRUN, "Anytime point-based approximations for large POMDPs", *Journal of Artificial Intelligence Research*, t. 27, p. 335-380, 2006.

[11] H. KURNIAWATI, D. HSU et W. S. LEE, "SARSOP : efficient point-based POMDP planning by approximating optimally reachable belief spaces", in *Robotics : Science and Systems IV*, 2008.

[12] S. PATEK, "On partially observed stochastic shortest path problems", in *Proceedings of the 40th IEEE Conference on Decision and Control*, t. 5, 2001, p. 5050-5055. DOI : 10.1109/CDC.2001.981011.

[13] K. HORÁK, B. BOŠANSKÝ et K. CHATTERJEE, "Goal-HSVI : heuristic search value iteration for goal-POMDPs", in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018, p. 4764-4770.

[14] E. ALTMAN, *Constrained Markov Decision Processes*. Chapman et Hall/CRC, 1999.

[15] D. KIM, J. LEE, K.-E. KIM et P. POUPART, "Point-based value iteration for constrained POMDPs", in *IJCAI*, t. 11, 2011, p. 1968-1974.

[16] F. DUFOUR et T. PRIETO-RUMEAU, "Stochastic approximations of constrained discounted Markov decision processes", *Journal of Mathematical Analysis and Applications*, t. 413, n° 2, p. 856-879, 2014. DOI : 10.1016/j.jmaa.2013.12.016.

[17] F. TREVIZAN, S. THIÉBAUX, P. SANTANA et B. WILLIAMS, "Heuristic search in dual space for constrained stochastic shortest path problems", *Proceedings of the International Conference on Automated Planning and Scheduling*, t. 26, n° 1, mar. 2016. DOI : 10.1609/icaps.v26i1.13768.

[18] J. LEE, G.-H. KIM, P. POUPART et K.-E. KIM, "Monte-Carlo tree search for constrained POMDPs", *Advances in Neural Information Processing Systems*, t. 31, 2018.

Combinatorial Games with Incomplete Information*

Junkang Li[1,2], Bruno Zanuttini[2], Véronique Ventos[1]

[1] NukkAI, Paris, France
[2] Normandie Univ.; UNICAEN, ENSICAEN, CNRS, GREYC, 14 000 Caen, France

junkang.li@nukk.ai ; bruno.zanuttini@unicaen.fr ; vventos@nukk.ai

Résumé

Les jeux à information incomplète modélisent des interactions multi-agents dans lesquelles les joueurs n'ont pas connaissance commune du jeu auquel ils jouent. Nous proposons une généralisation minimale du formalisme des jeux combinatoires afin d'incorporer l'information incomplète : les jeux combinatoires à information incomplète (CGII). La caractéristique la plus importante des CGII est que toutes les actions sont publiques, ce qui permet de mieux visualiser la connaissance et l'information incomplète de chaque joueur. Pour motiver davantage l'étude de ce nouveau formalisme, nous montrons que le calcul des stratégies optimales pour les CGII a exactement la même complexité algorithmique que pour les jeux généraux sous forme extensive.

Mots-clés

Jeux à information incomplète, jeux sous forme extensive, théorie de la complexité

Abstract

Games with incomplete information model multi-agent interaction in which players do not have common knowledge of the game they play. We propose a minimal generalisation of combinatorial games to incorporate incomplete information, called combinatorial game with incomplete information (CGII). The most important feature of CGIIs is that all actions are public, which allows better visualisation of each player's knowledge and incomplete information. To further motivate the study of this new formalism, we show that computing optimal strategies for CGIIs has the same computational complexity as for general extensive-form games.

Keywords

Games with incomplete information, extensive-form games, computational complexity theory

1 Introduction

Game theory is a mathematical framework for studying multi-agent interactions. We focus on *extensive-form games* (EFG), in which the interaction between agents takes place sequentially, i.e. every agent takes turns to make a move. Prominent examples of such games are Chess and Go.

Of particular interest to us is the notion of games with *incomplete information*, which are games in which agents do not have common knowledge of the game they play. For instance, an agent does not know the number of participants in an auction, or how much these participants value the object to be sold ; a Poker player does not see the cards in their opponent's hidden hands, hence cannot know for sure the exact consequence (i.e. payoff) of calling and raising bets ; a Bridge or Hearts player does not know the cards that their opponent can play during a trick since this depends on their hidden hand ; etc.

The notion of (in)complete information is frequently confused with the one of *(im)perfect information*. Complete information describes situations in which the whole structure of a game (the number of players, the game tree, the information sets of each player, the owner of each node, the payoff for each player at each leaf node, etc.) is common knowledge among all the players of the game. On the other hand, perfect information is a more stringent requirement than complete information. Not only the structure of the game is common knowledge, but all players have full observability and perfect recall of the history (which is essentially a record of every decision made by every player so far). In other words, players always know their exact position in the game tree when asked to make the next decision. To summarise, incomplete information is an example of imperfect information ; see Faliszewski *et al.* [15, Sec. 2.4.2].

We propose a new and minimal formalism for EFGs with incomplete information that we call combinatorial games with incomplete information (CGIIs). In such a game, Nature picks a world from a universe according to some common prior ; each player may have different observability of this world. Then, the game proceeds sequentially, during which there is no chance factor and all moves by the players are publicly observable. This formalism is designed to be a minimal generalisation of the notion of combinatorial games (which are Boolean games of no chance and with perfect information ; see Siegel [39]) and to closely capture the epistemic aspect of games with incomplete information. For such games, we are interested in knowing how much reward an agent or a team of agents can guarantee for themselves ; this corresponds to the notion of maxmin value, well known in optimisation under uncertainty, in which we aim

*This article is to be published in the proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI 2024). A long version with proofs of all claims is available at `https://hal.science/hal-04568854`.

to ensure that the worst possible outcome is not too bad.

By design, our new formalism seems particularly restrictive when compared to general EFGs, where hidden actions and arbitrary chance nodes are allowed. However, we show that the computational complexity of computing optimal strategies (with respect to the maxmin value) for CGIIs is as hard as for EFGs, which allows concluding that the difficulty of playing games comes from incomplete information/knowledge alone, not from hidden actions or mid-game chance factors. This also justifies that restricting algorithmic studies to CGIIs is without loss of generality. We also give a construction to enforce coordination between players in CGIIs under the constraint of public actions, which allows modelling situations similar to concurrent actions.

2 Related Work

Game theory. The study of games with incomplete information was pioneered by Harsanyi [22, 23, 24], who proposes a formalism to model games of incomplete information as EFGs with imperfect information. This formalism, called the Harsanyi model of incomplete information, introduces *types* of players, or equivalently, a universe of worlds for which each player has a potentially different partial observability (also called the Aumann model of incomplete information). For detailed and formal definitions, see textbooks on game theory, e.g. Maschler *et al.* [31, Chapter 9].

Combinatorial games, the inspiration of our formalism of CGII, are studied in the field of combinatorial game theory, established in the 70s by two books by Conway [13] and Berlekamp *et al.* [3, 4, 5, 6]. For recent advances in this field, see Nowakowski [33, 34, 35], Albert and Nowakowski [1], and Larsson [29].

Our formalism is also inspired by Frank and Basin [16], who, in order to model the card play phase of the card game Bridge, propose a game with public actions and one-sided incomplete information in which the opponent has complete information. Frank and Basin [17] show that finding optimal pure strategies for these games is NP-complete. Ginsberg [20] proposes the first exact algorithm for these games, and implements it for Bridge robots. Parallelly, Chu and Halpern [11] study a model of games with incomplete information with common payoffs, and only one round of concurrent interaction after Nature picks the world; they show that it is NP-complete to play such games optimally.

Like us, Kovarík *et al.* [27] highlight the distinction between public and private actions. They also argue that this distinction, essential for recent search algorithms, is partially lost when we model sequential multi-agent interaction with EFGs, which do not *explicitly* tell whether an action is public or not. They propose an alternative model for stochastic games that makes this distinction prominent, and show how to transform such models to augmented EFGs and vice versa.

Complexity of games. Most work in the literature on the computational complexity of games concerns the complexity of finding Nash equilibria, especially for normal-form games [18, 14]. For more references, see Conitzer and Sandholm [12], who also show that it is NP-complete to decide whether Nash Equilibria with certain natural properties exist.

Koller and Megiddo [25], Koller *et al.* [26], and von Stengel [41] make seminal contributions to understanding the complexity of two-player zero-sum EFGs. They also give polynomial-time algorithms for computing behaviour maxmin strategies of EFGs with perfect recall, based on linear programming.

Maxmin for a team of players with common payoffs is called team maxmin equilibrium (TME) in the literature, and was first proposed by von Stengel and Koller [42]. Basilico *et al.* [2] and Celli and Gatti [8] propose another notion called TMECor ("Cor" stands for "correlation"), which allows agents in the same team to access a correlation device in order to coordinate their mixed strategies. Building on these works, Gimbert *et al.* [19] and Zhang *et al.* [43] study the complexity of TME and TMECor, thereby yielding a relatively complete picture of the complexity of behaviour and mixed maxmin for two-team EFGs.

The complexity of other models of decision making have also been extensively studied, e.g. Markov decision process [32, 7, 21], propositional planning [37], graph games [10, 9]. Similar to these works, we confirm the intuition that partial observability and multi-agent coordination increases the difficulty of optimal decision making.

3 Combinatorial Games with Incomplete Information

3.1 Definitions

Combinatorial games are EFGs of no chance and with perfect information. To generalise this formalism minimally to allow incomplete information, we propose the following definition.

Definition 3.1 (CGII)**.** *A* combinatorial game with incomplete information *(CGII) is a tuple of the following elements :*

- *An Aumann model* $\langle U, A, (\mathcal{R}_i)_{i\in A}, \rho\rangle$*, where* U *is a finite set of* worlds *called* universe*,* A *is a set of* agents*,* $\mathcal{R}_i$ *is an equivalence relation over* U *for each agent* $i \in A$*, and* $\rho \in \Delta(U)$ *is a probability distribution over the universe called* common prior*;*
- *A tree* T *called* public tree*, the nodes of which* $(\mathrm{N}(T))$ *are partitioned into* $\{\mathrm{N}_i(T)\}_{i\in A} \cup \mathrm{L}(T)$*, where* $\mathrm{L}(T)$ *is the set of all leaves;*
- *A reward function* $u_i : \mathrm{L}(T) \times U \to \mathbb{R}$ *for each* $i \in A$*.*

Note that the children of a node (available actions at that node) do not depend on the real (and partially observable) world ω; only the rewards depend on ω.

A CGII is said to be *Boolean* if all its reward functions have values in $\mathbb{B}$.[1] The Aumann model of a CGII defines each agent's observability over the universe, which characterises their incomplete information.

1. In Boolean games, the rewards 0 and 1 are interpreted as a loss and a win, respectively.

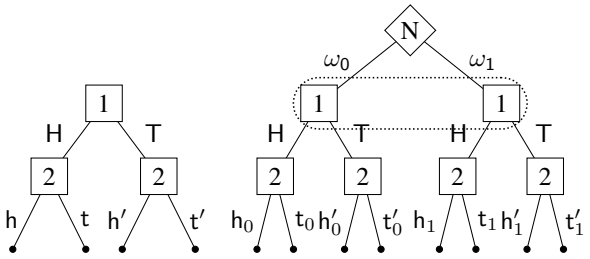


FIGURE 1 – The public tree of a CGII and the game tree of its corresponding EFG, both with rewards omitted.

Pure strategies in a CGII. A CGII as an EFG with incomplete information proceeds as follows. First, Nature picks the *real world* $\omega \in U$ according to ρ. Then the *state game* in ω proceeds from the root of the public tree T ; agents take turns to pick a child of the current node, depending on their equivalence class of the real world. This continues until a leaf l is reached, and agent i receives a payoff $u_i(l, \omega)$.

Definition 3.2 (Pure strategy)**.** *A* pure strategy *of an agent $i \in A$ is a mapping $s_i : \mathrm{N}_i(T) \times U \to \mathrm{N}(T)$ such that for all $v \in \mathrm{N}_i(T)$:*

- *For all $\omega \in U$, $s_i(v, \omega)$ is a child of v ;*
- *$\forall \omega, \omega' \in U, \omega \mathcal{R}_i \omega' \implies s_i(v, \omega) = s_i(v, \omega')$.* [2]

The set of all pure strategies of agent i is denoted by Σ_i^{P}.

From the definition of a strategy, one can see that the actions of every agent are indeed public : when making a decision at a node, an agent knows perfectly where the node is in the public tree, which in particular means they observe and remember the actions picked by every agent in the past, starting from the root of the public tree. In addition, compared to general games with incomplete information, the state games of a CGII have the particularity that they share the same game tree T, which does not have chance nodes. These three features (public actions, unique game tree, and no chance) are the defining features of our formalism CGII. Each CGII describes an EFG in which Nature picks the real world at the root, and information sets are determined by the players' observability of the world.

Example. *Consider the following CGII : the public tree is shown in figure 1 (on the left) ; the universe reads $\{\omega_0, \omega_1\}$; agent 2 can distinguish these two worlds while agent 1 cannot. This CGII models a variant of Matching Pennies with incomplete information. The game tree of its corresponding EFG is also shown in figure 1 (on the right).*
On the right, since agent 1 cannot observe the real world, they must play H *in both state games, or* T *in both. This constraint is respected by the notion of strategies in a CGII : on the left, agent 1 only has two pure strategies* H *and* T *since ω_0 and ω_1 are indistinguishable by agent 1.*
Similarly, on the right, agent 2 can pick between heads or tails, depending on the choices of Nature and agent 1. On the left, agent 2 can again pick between heads or tails, depending on agent 1's choice and the real world, since agent 2 can distinguish between ω_0 and ω_1 ; note that the latter point is not reflected by the public tree, but by the Aumann model.

Teams and Information in a CGII. In a CGII, agents i and j are said to be in the same *team* if $u_i = u_j$.

Definition 3.3 (Team)**.** *A* team *is an inclusion-wise maximal group of agents with the same reward function.*

We now define a team's *degree of incomplete information*.

- *Multi-agent incomplete information* (MA-II) : an arbitrary team.
- *Single-agent incomplete information* (SA-II) : a team of agents with the same equivalence relation (i.e. $\mathcal{R}_i = \mathcal{R}_j$ for all agents i and j in the team).
- *Complete information* (CI) : a team whose agents all have the finest equivalence relation (i.e. $\mathcal{R}_i = \{(\omega, \omega) \mid \omega \in U\}$ for all agents i in the team).

In particular, CI implies SA-II, which implies MA-II. Intuitively, a team is a group of decentralised agents with shared interests working cooperatively. In a CGII, a team with SA-II can be regarded as one single agent since every agent in this team has the same information and all actions are public.

Example. *In the CGII in figure 1, if the two agents have the same reward function, then they are in a team with MA-II ; otherwise, each is a (single-agent) team with SA-II.* [3]

Due to the public actions property, there is a close link between the degree of incomplete information of a team in a CGII and the degree of imperfect information of the corresponding team in the EFG defined by the CGII :

- a team with CI in the CGII is a player with perfect information in the EFG ;
- a team with SA-II in the CGII can be seen as a single player with perfect recall in the EFG ;
- a team with MA-II in the CGII is a team of players who all have perfect recall in the EFG.

This correspondence will allow us to establish upper bounds on the complexity of solving CGIIs.

Team maxmin in a CGII. Let $(s_1, \ldots, s_n) \in \Sigma_1^{\mathrm{P}} \times \cdots \Sigma_n^{\mathrm{P}}$, where $n = |A|$, be a pure *strategy profile*. We write $(s_1, \ldots, s_n)(\omega)$ for the unique leaf reached under this profile when the real world is ω.

Definition 3.4 (Expected utility)**.** *The* expected utility *for an agent $i \in A$ under a pure strategy profile $(s_1, \ldots, s_n)$ is defined to be :*

$$\mathcal{U}_i(s_1, \ldots, s_n) = \sum_{\omega \in U} \rho(\omega) u_i\big((s_1, \ldots, s_n)(\omega), \omega\big).$$

Let $\mathcal{T} \subseteq A$ be a team. Notice that all agents in a team share the same expected utility function, which we denote by $\mathcal{U}_{\mathcal{T}}$. A pure strategy of the team is uniquely defined by the pure strategy of each of its players. In particular, the

2. This means agent i must pick the same child for a node in any two worlds indistinguishable by them.

3. The team of agent 2 even has CI.

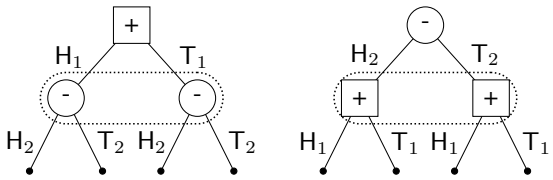


FIGURE 2 – Two EFGs for Matching Pennies.

set of pure strategies of a team $\mathcal{T}$, denoted by $\Sigma^{\mathrm{P}}_{\mathcal{T}}$, is in bijection with $\prod_{i\in A}\Sigma^{\mathrm{P}}_i$. In the following, we also write $\Sigma^{\mathrm{P}}_{-\mathcal{T}} = \prod_{i\notin\mathcal{T}}\Sigma^{\mathrm{P}}_i$, the set of pure strategy profiles of the players not in $\mathcal{T}$.

Definition 3.5 (Pure maxmin for a team)**.** *The* pure maxmin *value for a team $\mathcal{T} \subseteq A$ is defined to be*

$$\underline{v}_{\mathcal{T}} \coloneqq \max_{s_{\mathcal{T}}\in\Sigma^{\mathrm{P}}_{\mathcal{T}}} \min_{s_{-\mathcal{T}}\in\Sigma^{\mathrm{P}}_{-\mathcal{T}}} \mathcal{U}_{\mathcal{T}}(s_{\mathcal{T}}, s_{-\mathcal{T}}).$$

Intuitively, this value is the largest expected reward that a team can guarantee to get by playing a pure strategy.

The notion of behaviour/mixed strategy can be defined similarly to the one for EFGs : a mixed strategy of an agent i is a probability mixture of pure strategies of i ; a behaviour strategy of i picks, at each node and for each equivalence class of $\mathcal{R}_i$, a probability mixture of children (instead of just a child as for pure strategies). Hence, expected utility with respect to behaviour/mixed strategy profiles and behaviour/mixed maxmin for a team can be similarly defined. [4]

In the following, we focus on zero-sum two-team CGIIs. We call the two teams *player MAX* and *player MIN*, and denote them by $+$ and $-$, respectively.

3.2 Motivation for CGIIs

Our motivations for introducing CGII as a subclass of games with incomplete information are multiple. First and foremost, the formalism of CGII aims to be a minimal generalisation of combinatorial games to allow incomplete information. Indeed, it is clear that a CGII with a singleton universe is a combinatorial game. This new formalism allows modelling a number of card games, notably Bridge. [5]

But more importantly, the formalism of CGII also aims to minimally capture the notion of knowledge and incomplete information. Due to the public actions property, the only source of the imperfect (in particular, incomplete) information of every agent comes from their partial observability of the real world, drawn at the beginning of a game.

In contrast, we argue that the distinction between perfect and imperfect information does not completely capture the essence of players' knowledge. For example, in the game Matching Pennies, MAX and MIN pick a side of a coin concurrently ; this can be modelled by two different EFGs, shown in figure 2. In the EFG on the left, MAX has perfect information while MIN has imperfect information but perfect recall ; and in the EFG on the right, the situation is reversed. However, the roles of MAX and MIN are symmetric in Matching Pennies ; MAX also has exactly the same information/knowledge in both EFGs when they need to choose an action. Hence, considering CGIIs allows one to focus on an unambiguous notion of knowledge of the players, as captured by the Aumann model and the initial drawing of a world.

Expressiveness. At first sight, the requirements of public actions and no chance seem particularly restrictive : many popular tabletop games with incomplete information allow private actions (e.g. concealed Kong in Mahjong, pass in Hearts) or have randomness and chance factors besides the initial drawing (e.g. dice rolls during a game). One may worry that, due to these restrictions, CGII is not expressive enough to be conceptually or algorithmically interesting. However, we argue that this impression is not correct.

First, an initial drawing over the universe is actually quite expressive. For example, for the dice rolls we evoke above, if their number and occasions are fixed in advance, then their results can be encoded into the initial drawing of worlds. [6] Another example is given by video games, which typically use a random seed as the sole source of randomness for all procedurally generated levels and random events during a playthrough. Similar ideas have been investigated in automated planning [36, Sec. 10].

Second, even with only public actions, we show in sous-section 4.1 that we can still design a game to force a team of players to coordinate their actions. This means that we can essentially encode concurrent actions (as in standard Matching Pennies) using only public actions (and no chance except the initial drawing).

All in all, we suggest that at least as far as computation of optimal strategies is concerned, CGII, rather than EFG, be the right model for studying sequential multi-agent interactions depending on each player's knowledge. Moreover, as we will show, CGIIs are as hard to solve as EFGs, which confirms our intuition that the difficulty of a game actually comes from the incomplete information of a player, and not from their inability to observe the moves made by the other players.

4 Complexity of PURE MAXMIN for CGIIs

The decision problem PURE MAXMIN is defined as follows.

Definition 4.1 (PURE MAXMIN)**.** *Let $\mathcal{G}$ be a class of zero-sum CGIIs. Then* PURE MAXMIN($\mathcal{G}$) *is the following decision problem.*

INPUT *A CGII $G \in \mathcal{G}$ and a rational number m.*

OUTPUT *Decide whether the pure maxmin value for team MAX in G satisfies $\underline{v}_+(\Sigma^{\mathrm{P}}_+, \Sigma^{\mathrm{P}}_-) \geq m$.*

4. Behaviour maxmin and mixed maxmin are commonly known as TME and TMECor in the literature [8].

5. The card play of Bridge can be described as a CGII in which MAX has SA-II and MIN has MA-II.

6. This will only enlarge the game tree by a polynomial factor.

MAX \ MIN	CI	SA-II	MA-II
CI	P	**NP-c**	$\mathbf{\Sigma_2^P}$**-c**
SA-II	NP-c	NP-c	$\mathbf{\Sigma_2^P}$**-c**
MA-II	NP-c	NP-c	$\mathbf{\Sigma_2^P}$**-c**

TABLE 1 – Complexity of PURE MAXMIN for CGIIs.

We study the complexity of PURE MAXMIN for CGIIs depending on the degrees of incomplete information for MAX and MIN : complete information (CI), single-agent incomplete information (SA-II), multi-agent incomplete information (MA-II). For complexity analyses, we consider the parameters $|T|$ (number of nodes in the public tree), $|U|$ (number of worlds), and possibly the number of bits to encode the utilities, the common prior, and the threshold m.
The complexity of PURE MAXMIN is summarised in tableau 1. By definition, the complexity of each case is increasingly monotone in both MAX's and MIN's degree of incomplete information (CI/SA-II/MA-II). Hence, only a few hardness results have to be proved to establish the table. The results written in bold font are new from this work ; the others can be directly deduced from the literature.
The membership results in tableau 1 follow from results by Koller and Megiddo [25, Sec. 3.3] ; in particular, memberships in NP and in Σ_2^P follow from the fact that given a strategy of MAX, computing MIN's best response is a problem in coNP, and even linear time when MIN has perfect recall.
Hence, we focus on hardness results. The following result is by Frank and Basin [17, Sec. 6]. [7]

Proposition 4.2. PURE MAXMIN *is* NP*-hard for Boolean CGIIs in which MAX has SA-II and MIN has CI.*

The symmetric case does not trivially follow from this result (since the minimax theorem does not hold for pure strategies) and necessitates a proof :

Proposition 4.3. PURE MAXMIN *is* NP*-hard for Boolean CGIIs in which MAX has CI and MIN has SA-II.*

Proof sketch. By a reduction from VERTEX COVER. Given a graph (V, E), consider the universe $U = \{\omega_e \mid e \in E\}$; the worlds are observable by MAX but not by MIN. During the game, MAX picks a vertex $v \in V$, then MIN picks an edge $e' \in E$. In a world $\omega_e \in U$, MAX gets a payoff of 1 if v covers e and MIN does not correctly guess this edge (i.e. $e' \neq e$) ; otherwise, MAX gets 0. One can verify that the pure maxmin value of this game is at least $1 - k/|E|$ if and only if the graph has a vertex cover of size at most k. □

4.1 Multi-Agent Coordination in CGIIs

Coordination game. Now we turn our attention to CGIIs with multi-agent teams. We first show how to construct CGIIs to impose a perfect coordination between agents from the same team (à la Matching Pennies).
Consider the following Boolean CGII, which we call *coordination game*. This game has two agents of MAX, referred to as MAX 1 and MAX 2, and no agent of MIN ; its universe has 4 worlds and reads $U = \{(b_1, b_2) \mid b_1, b_2 \in \mathbb{B}\}$; its Aumann model has the uniform common prior and is such that for $i = 1, 2$, MAX i only observes b_i ; its public tree is shown in figure 3 ; the reward for MAX is 1 if and only if $a_1 \oplus b_1 = a_2 \oplus b_2$, where $\oplus$ is the exclusive or of two bits and a_i is the action chosen by MAX i.
We refer to b_i as the *hidden bit* of MAX i since it is only observable by MAX i. The coordination game is designed in such a way that MAX 1 and MAX 2 must perfectly coordinate their answer in order to win. Intuitively, MAX 1 and MAX 2 need to agree on the same answer $A \in \mathbb{B}$, then stick to it during the game by playing $a_i = A \oplus b_i$. Indeed, if they employ this strategy, then they guarantee a win since

$$a_1 \oplus b_1 = (A \oplus b_1) \oplus b_1 = A = (A \oplus b_2) \oplus b_2 = a_2 \oplus b_2.$$

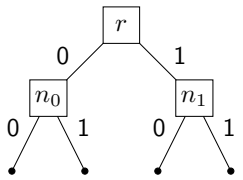


FIGURE 3 – The public tree of the coordination game.

Remark. *Under these two winning strategies (one for each value of A), both MAX 1 and 2 pick the actions* 0 *and* 1 *with equal probability. Indeed, once the common answer A is fixed, which action to play by MAX i is dictated by their hidden bit b_i. Hence, the bits b_1 and b_2 act as the keys of a one-time pad to* encrypt/mask *the intended answer (i.e. A) of MAX 1 and 2. This is the key element to ensure that MAX 1 and 2 must cooperate without cheating.*

Proposition 4.4. *In a coordination game, the only winning pure strategies of team MAX are of the following form : for some $A \in \mathbb{B}$, MAX* 1 *plays $A \oplus b_1$ and MAX* 2 *plays $A \oplus b_2$.*

Démonstration. Notice that the pure strategies of MAX 1 can be written in the form (a_1^0, a_1^1), which means they choose a_1^0 if $b_1 = 0$ and a_1^1 if $b_1 = 1$. As for MAX 2, they have the right to pick a_2 as a function of a_1 and b_2. If MAX 1 plays (A, A) (i.e. they play A regardless of b_1) for some $A \in \mathbb{B}$, then MAX 2 has no winning strategy, since the winning condition $A \oplus b_1 = a_2 \oplus b_2$ cannot be satisfied for both values of b_1. Now if MAX 1 plays $(A, A \oplus 1)$ for some $A \in \mathbb{B}$, then to satisfy the winning condition, MAX 2 is forced to play $a_2 = A \oplus b_2$; hence $a_i = A \oplus b_i$. □

The same reasoning also shows that these pure strategies are also the only winning behaviour strategies, and that the winning mixed strategies are exactly the mixtures of them.

Remark. *From this proof, one can see that if MAX 1 cheats by using their hidden bit b_1 incorrectly (i.e. does not use b_1*

7. In their setting, there is no prior over the worlds ; they are interested in the strategies that win in the greatest number of worlds. This is equivalent to finding maxmin strategies with respect to the uniform prior in our setting.

to encrypt their intended answer and always picks the same action), then MAX 2 cannot cooperate perfectly since they cannot observe the value of b_1.
In addition, when MAX 1 plays correctly (i.e. chooses a strategy of the form $(A, A \oplus 1)$*), then MAX 2 must also pick* A *as their intended answer and mask it with their own bit* b_2 *in order to win. Notice that in this case, the action picked by these two agents are uniformly and independently distributed. This is an important feature since agents (of MAX or MIN) in the following part of the game tree cannot deduce any information about the intended answer of these two agents by observing only their actions.*

Interrogation game. We now generalise the coordination game to the following situation : we have a finite set of questions Q, and MAX has a Boolean answer for each question $\{A_q \in \mathbb{B}\}_{q \in Q}$, or equivalently a mapping from Q to $\mathbb{B}$. We wish to verify whether MAX's mapping satisfies some given binary constraints $\{C_{qq'} \subseteq \mathbb{B}^2 \mid q, q' \in Q, q \neq q'\}$: MAX's mapping is said to be *valid* if it satisfies all the constraints, that is, $(A_q, A'_q) \in C_{qq'}$ for all $C_{qq'}$.

Example. *For cliques of a given graph, the questions are the vertices of this graph; MAX's mapping induces a subgraph (MAX's answer to a vertex corresponds to whether to include this vertex in their intended subgraph); the binary constraints impose the requirement that all vertices in this subgraph be connected. Then MAX's mapping is valid if and only if it describes a clique of the graph.*

To model this situation, consider the following Boolean CGII, which we call *interrogation game* : two agents of MAX (MAX 1 and MAX 2), and no agent of MIN ; the universe reads $U = \{(q_1, b_1, q_2, b_2) \mid q_1, q_2 \in Q, b_1, b_2 \in \mathbb{B}\}$; the Aumann model is such that for $i = 1, 2$, MAX i only observes q_i and b_i ; the common prior is uniform ; the public tree is the same one as for the coordination game (i.e. the one in figure 3) ; MAX loses if and only if either (1) $q_1 = q_2$ but $a_1 \oplus b_1 \neq a_2 \oplus b_2$ or (2) $q_1 \neq q_2$ but $(a_1 \oplus b_1, a_2 \oplus b_2) \notin C_{q_1 q_2}$.
This CGII has size $\mathcal{O}(|Q|^2)$: the universe has size $\mathcal{O}(|Q|^2)$, while the public tree has size $\mathcal{O}(1)$. Notice that a coordination game is just an interrogation game with only one question (hence no binary constraint). We refer to (q_i, b_i) as the *hidden information* of MAX i. Inspired by the coordination game, we propose the following definition.

Definition 4.5 (Perfect coordination)**.** *In an interrogation game, a* perfect coordination *of team MAX is a pure strategy of MAX of this form : there is a set* $\{A_q \in \mathbb{B}\}_{q \in Q}$ *such that for all* i*, MAX* i *will play the action* $a_i = A_{q_i} \oplus b_i$ *in all worlds in which their hidden information is* (q_i, b_i)*. For such a strategy, the set* $\{A_q\}_{q \in Q}$ *is called the* intended mapping *or* intended answer *of the perfect coordination.*

By a similar argument to the one for the coordination game, the reward condition (1) ensures that MAX 1 and 2 have an incentive to implement a perfect coordination, which is a dominant strategy. In other words, (1) imposes non-adaptivity of MAX's answers. As for the reward condition (2), it ensures that all binary constraints are satisfied by the intended mapping of a perfect coordination, since by (1) we have $a_i \oplus b_i = A_{q_i}$ for all i. In summary, we have established the following result.

Proposition 4.6. *In an interrogation game, a pure strategy of team MAX is winning if and only if it is a perfect coordination with a valid intended mapping.*

It is straightforward to construct interrogation games involving team MIN such that if MIN does not cooperate, MAX receives a large reward. Similarly, we can also extend the construction above to allow k-ary constraints with $k \geq 2$. The interrogation game will then involve k agents of MAX, each with their hidden information (q_i, b_i), and has size $\mathcal{O}(2^k |Q|^k)$. Such an interrogation game can be used to encode problems such as k-SAT. [8]

4.2 Hardness for Two-Team CGIIs

With the gadgets of interrogation game, it is straightforward to show that PURE MAXMIN is Σ_2^{P}-hard for CGIIs in which both MAX and MIN are multi-agent teams, for instance by a reduction from the canonical problem $\exists\forall$3SAT. However, we provide a stronger result : Σ_2^{P}-hardness holds even when MAX has complete information.

Proposition 4.7. PURE MAXMIN *is* Σ_2^{P}*-hard for CGIIs in which MAX has CI and MIN has MA-II.*

Proof sketch. By a reduction from the Σ_2^{P}-complete problem SUCCINCT SET COVER [40] : given a collection of 3-DNF formulae and an integer k, decide whether there is a subset S of size at most k the disjunction of which is a tautology.
We design a game in which Nature draws a DNF formula from the collection, 3 variables, and 4 hidden bits, according to the uniform common prior. The DNF is known to MAX, who plays 1 or 0 according to whether it should be in S. This answer is masked (to MIN) by the hidden bit of MAX, as in a coordination game. Then MIN chooses either to verify the size of S or to verify that the disjunction of S is a tautology. The other 3 hidden bits are used in the latter verification : MIN plays an interrogation game over the 3 variables, with the constraint to falsify the disjunction of S.
Finally, since MAX is designed to have CI, they know the variables and the hidden bits of MIN. To ensure that MAX does not play a strategy that depends on MIN's information, we introduce one additional agent of MIN whose role is to punish MAX whenever MAX plays such a strategy. □

Remark. *The construction shows something stronger :* Σ_2^{P}*-hardness holds even when MIN has* joint *complete information (i.e. if the agents of MIN could pool their information, then they would have complete information).*

8. Contrastingly, we leave open the problem of constructing an interrogation game in which MAX's answers are not binary.

MAX \ MIN	CI	SA-II	MA-II
CI	P	P	**coNP-c**
SA-II	P	P	**coNP-c**
MA-II	**NP-c**	**NP-c**	$\mathbf{\Sigma_2^P}$**-c/**$\mathbf{\Delta_2^P}$**-c**

TABLE 2 – Complexity of BEHAVIOUR MAXMIN and of MIXED MAXMIN for CGIIs.

5 Complexity of BEHAVIOUR MAXMIN and MIXED MAXMIN

The decision problems BEHAVIOUR MAXMIN and MIXED MAXMIN can be defined similarly to PURE MAXMIN, the only difference being that MAX can use behaviour/mixed strategies instead of just pure strategies. [9]

The complexity of BEHAVIOUR MAXMIN and MIXED MAXMIN is summarised in tableau 2. Again, the complexity is increasingly monotone in both dimensions, and results written in bold font are new. The only case where the complexity differs between behaviour and mixed strategies is the case in which both MAX and MIN have MA-II; in this case, BEHAVIOUR MAXMIN and MIXED MAXMIN are respectively Σ_2^P- and Δ_2^P-complete.

The membership results follow from those for EFGs, which are superclasses of CGIIs : the results for P are by Koller and Megiddo [25, Sec. 3.5], and the others by Zhang *et al.* [43, Appx. C].

Therefore, we only have to establish the hardness results when MAX and/or MIN have MA-II. We first adapt a reduction from 3-SAT by Chu and Halpern [11].

Proposition 5.1. *Both* BEHAVIOUR MAXMIN *and* MIXED MAXMIN *are* NP*-hard for Boolean CGIIs in which MAX has MA-II and MIN has CI.*

Démonstration. For a 3-CNF with N clauses, consider the following CGII. The universe consists of the clauses, which are observable by MAX 1 but not by MAX 2, and with the uniform prior. During the game, MAX 1 picks a variable, then MAX 2 observes this variable and picks a truth value. They win if and only if the variable picked by MAX 1 is in the clause picked by Nature, and the truth value picked by MAX 2 for this variable renders this clause true.

Since there is no agent of MIN in this game, playing behaviour or mixed strategies is no better than pure ones. It is also straightforward to verify that MAX can guarantee an expected payoff of 1 if the 3-CNF is satisfiable; otherwise, the maxmin value for MAX is at most $1 - 1/N$. □

Now, coNP-hardness for the symmetric case (when MAX has CI and MIN has MA-II) essentially follows from this result. For mixed strategies, the minimax theorem ensures that when switching the roles of MIN and MAX, and negating the utilities in the game from the proof of Proposition 5.1, the maxmin value for MAX is at least $-(1-1/N)$ if the 3-CNF is unsatisfiable, and -1 otherwise. The hardness for BEHAVIOUR MAXMIN follows from the fact that mixed maxmin and behaviour maxmin have the same value due to MAX's perfect recall.

Proposition 5.2. BEHAVIOUR MAXMIN *is* Σ_2^P*-hard for CGIIs in which both MAX and MIN have MA-II.*

Proof sketch. By a reduction from $\exists\forall$3SAT (for a 3-DNF formula $\varphi(x, y)$, decides whether $\exists x \forall y\, \varphi(x, y)$ holds) which is known to be Σ_2^P-hard [38]. Given such a formula, we construct a CGII with 3 agents of MAX and 3 agents of MIN. The worlds consist of one existential (resp. universal) variable and one hidden bit for each agent of MAX (resp. of MIN); the common prior is uniform; each agent only observes their variable and hidden bit. During the game, the agents of MAX take turns to choose between 0 and 1, then so do the agents of MIN. The total payoff for MAX is computed as follows : (1) an inconsistency among the agents of MAX (in the sense of an interrogation game) yields $-N$ for MAX, where N is a large real number; (2) an inconsistency among the agents of MIN yields $+N$ for MAX; (3) if at least one term in $\varphi(x, y)$ is satisfied by the assignment picked by the agents of MAX and MIN, then MAX receives $+1$.

By choosing N large enough, agents of MAX have an incentive to perform a perfect coordination, and the same goes for agents of MIN. In particular, MAX has no incentive to play non-pure behaviour strategies, which would cause inconsistency to happen with a non-zero probability. It is then straightforward to verify that $\exists x \forall y\, \varphi(x, y)$ holds if and only if MAX can guarantee an expected utility of at least $+1/n^3$, where n is the maximum between the number of existential variables and the number of universal ones. □

Proposition 5.3. MIXED MAXMIN *is* Δ_2^P*-hard for CGIIs in which both MAX and MIN have MA-II.*

Proof sketch. By a reduction from LAST SAT (for a 3-CNF, decide whether the lexicographically maximum satisfying assignment has value 1 for the last variable), which is Δ_2^P-hard [28]. The construction is very similar to the last proof. Given a 3-CNF, we write the variables as $x_1, \ldots, x_n$, and we construct a CGII with 3 agents of MAX and 3 agents of MIN. The worlds consist of one variable and one hidden bit for each agent of MAX or MIN; the common prior is uniform; each agent only observes their variable and hidden bit. During the game, the agents of MAX take turns to choose between 0 and 1, then so do the agents of MIN. The total payoff for MAX is computed as follows : (1) an inconsistency among the agents of MAX or a clause violated by their assignment yields $-2N$ for MAX, where N is a large real number; (2) an inconsistency among the agents of MIN or a clause violated by their assignment yields $+N$

9. In our definition for all these decision problems, MIN only uses pure strategies, which is without loss of generality. Indeed, MIN is a team of agents with perfect recall, hence every MIN's behaviour strategy has an equivalent mixed strategy [31, Theorem 6.11]. In addition, the best responses in mixed strategies are no better than the best responses in pure strategies due to the linearity of expected utility with respect to mixtures of strategies.

for MAX ; (3) for the first agent of MAX (resp. of MIN), if their hidden variable and bit are x_k and b, and they pick $1 \oplus b$, then MAX receives $+2^{n-k}$ (resp. -2^{n-k}) ; (4) MAX receives a bonus $+1$ if the variable x_n is assigned $1 \oplus b_1^+$ by the first agent of MAX.

By choosing N large enough, both MAX and MIN have an incentive to perform a perfect coordination (which can be pure or mixed for MAX) with a satisfying assignment. Let $x = (x_1, \dots x_n)$ be the lexicographically maximum satisfying assignment (if there is no such assignment, then MAX is bound to get a large negative expected utility). If $x_n = 1$, then MAX can guarantee an expected utility of $+1/n$ by choosing this assignment for their perfect coordination ; the best MIN can do is to choose this assignment. If $x_n = 0$, MAX has an expected utility of at most 0 when MIN plays this assignment : MAX gets 0 by playing the same assignment, and possibly less when playing other satisfying (hence lexicographically smaller) assignments with a non-zero probability. □

6 Conclusion

We have proposed a new formalism for extensive-form games with incomplete information that we name combinatorial games with incomplete information. Compared to EFGs, CGIIs only have public actions and one chance node at the beginning of the game, thereby putting better emphasis on the aspect of incomplete information/knowledge of the players.

Apart from the conceptual simplicity, the interests in this new formalism are also justified by the complexity results. Indeed, all the upper bounds for CGIIs are provided by membership results that also hold for EFGs, while all the lower bounds, proven by hardness results, coincide with the upper bounds. In particular, for every degree of observability, CGIIs have the same complexity as EFGs.

We have also shown how to model binary concurrent actions to enforce multi-agent coordination in CGIIs. We leave to future work how to model other types of hidden actions, in particular non-binary concurrent actions. Future work also includes tightening the complexity results to show that hardness holds even for Boolean CGIIs with a minimum number of agents and distributed knowledge of the real world for each team ; designing a generic polynomial transformation from an arbitrary two-team EFG into a CGII ; and extending the study to general-sum multi-team CGIIs with respect to solution concepts that generalise maxmin (e.g. strategies to commit to [30]). Algorithmic studies adapted to CGIIs will also be of interest, with the long-term goal to implement better AIs for games such as Bridge.

Références

[1] Michael H. Albert and Richard J. Nowakowski, editors. *Games of No Chance 3*, volume 56 of *Mathematical Sciences Research Institute Publications*. Cambridge University Press, 2009.

[2] Nicola Basilico, Andrea Celli, Giuseppe De Nittis, and Nicola Gatti. Team-maxmin equilibrium : Efficiency bounds and algorithms. In Satinder Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI 2017)*, pages 356–362. AAAI Press, 2017.

[3] Elwyn R Berlekamp, John H Conway, and Richard K Guy. *Winning Ways for Your Mathematical Plays, Volume 1*. CRC Press, 2nd edition, 2001.

[4] Elwyn R Berlekamp, John H Conway, and Richard K Guy. *Winning Ways for Your Mathematical Plays, Volume 2*. CRC Press, 2nd edition, 2003.

[5] Elwyn R Berlekamp, John H Conway, and Richard K Guy. *Winning Ways for Your Mathematical Plays, Volume 3*. CRC Press, 2nd edition, 2003.

[6] Elwyn R Berlekamp, John H Conway, and Richard K Guy. *Winning Ways for Your Mathematical Plays, Volume 4*. CRC Press, 2nd edition, 2004.

[7] Daniel S. Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of markov decision processes. *Math. Oper. Res.*, 27(4) :819–840, 2002.

[8] Andrea Celli and Nicola Gatti. Computational results for extensive-form adversarial team games. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)*, pages 965–972. AAAI Press, 2018.

[9] Krishnendu Chatterjee, Laurent Doyen, and Thomas A. Henzinger. A survey of partial-observation stochastic parity games. *Formal Methods Syst. Des.*, 43(2) :268–284, 2013.

[10] Krishnendu Chatterjee and Thomas A. Henzinger. A survey of stochastic ω-regular games. *J. Comput. Syst. Sci.*, 78(2) :394–413, 2012.

[11] Francis C. Chu and Joseph Y. Halpern. On the np-completeness of finding an optimal strategy in games with common payoffs. *Int. J. Game Theory*, 30(1) :99–106, 2001.

[12] Vincent Conitzer and Tuomas Sandholm. New complexity results about Nash equilibria. *Games Econ. Behav.*, 63(2) :621–641, 2008.

[13] John H. Conway. *On Numbers and Games*. A K Peters/CRC Press, 2nd edition, 2000.

[14] Constantinos Daskalakis, Paul W. Goldberg, and Christos H. Papadimitriou. The complexity of computing a Nash equilibrium. *SIAM J. Comput.*, 39(1) :195–259, 2009.

[15] Piotr Faliszewski, Irene Rothe, and Jörg Rothe. Noncooperative game theory. In Jörg Rothe, editor, *Economics and Computation, An Introduction to Algorithmic Game Theory, Computational Social Choice, and Fair Division*, Springer texts in business and economics, pages 41–134. Springer, 2016.

[16] Ian Frank and David A. Basin. Search in games with incomplete information : A case study using bridge card play. *Artif. Intell.*, 100(1-2) :87–123, 1998.

[17] Ian Frank and David A. Basin. A theoretical and empirical investigation of search in imperfect information games. *Theor. Comput. Sci.*, 252(1-2) :217–256, 2001.

[18] Itzhak Gilboa and Eitan Zemel. Nash and correlated equilibria : Some complexity considerations. *Games and Economic Behavior*, 1(1) :80–93, 1989.

[19] Hugo Gimbert, Soumyajit Paul, and B. Srivathsan. A bridge between polynomial optimization and games with imperfect recall. In Amal El Fallah Seghrouchni, Gita Sukthankar, Bo An, and Neil Yorke-Smith, editors, *Proceedings of the Nineteenth International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020)*, pages 456–464. International Foundation for Autonomous Agents and Multiagent Systems, 2020.

[20] Matthew L. Ginsberg. GIB : imperfect information in a computationally challenging game. *J. Artif. Intell. Res.*, 14 :303–358, 2001.

[21] Judy Goldsmith and Martin Mundhenk. Competition adds complexity. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems (NIPS 2007)*, pages 561–568. Curran Associates, Inc., 2007.

[22] John C. Harsanyi. Games with incomplete information played by "Bayesian" players, Part I. The Basic Model. *Management Science*, 14(3) :159–182, 1967.

[23] John C. Harsanyi. Games with incomplete information played by "Bayesian" players, Part II. Bayesian Equilibrium Points. *Management Science*, 14(5) :320–334, 1968.

[24] John C. Harsanyi. Games with incomplete information played by "Bayesian" players, Part III. The Basic Probability Distribution of the Game. *Management Science*, 14(7) :486–502, 1968.

[25] Daphne Koller and Nimrod Megiddo. The complexity of two-person zero-sum games in extensive form. *Games and Economic Behavior*, 4(4) :528–552, 1992.

[26] Daphne Koller, Nimrod Megiddo, and Bernhard von Stengel. Efficient computation of equilibria for extensive two-person games. *Games and Economic Behavior*, 14(2) :247–259, 1996.

[27] Vojtech Kovarík, Martin Schmid, Neil Burch, Michael Bowling, and Viliam Lisý. Rethinking formal models of partially observable multiagent decision making. *Artif. Intell.*, 303 :103645, 2022.

[28] Mark W. Krentel. The complexity of optimization problems. *J. Comput. Syst. Sci.*, 36(3) :490–509, 1988.

[29] Urban Larsson, editor. *Games of No Chance 5*, volume 70 of *Mathematical Sciences Research Institute Publications*. Cambridge University Press, 2019.

[30] Joshua Letchford and Vincent Conitzer. Computing optimal strategies to commit to in extensive-form games. In David C. Parkes, Chrysanthos Dellarocas, and Moshe Tennenholtz, editors, *Proceedings of the Eleventh ACM Conference on Electronic Commerce (EC-2010)*, pages 83–92. ACM, 2010.

[31] Michael Maschler, Eilon Solan, and Shmuel Zamir. *Game Theory*. Cambridge University Press, 2nd edition, 2020.

[32] Martin Mundhenk, Judy Goldsmith, Christopher Lusena, and Eric Allender. Complexity of finite-horizon markov decision process problems. *J. ACM*, 47(4) :681–720, 2000.

[33] Richard J. Nowakowski, editor. *Games of No Chance*, volume 29 of *Mathematical Sciences Research Institute Publications*. Cambridge University Press, 1996.

[34] Richard J. Nowakowski, editor. *More Games of No Chance*, volume 42 of *Mathematical Sciences Research Institute Publications*. Cambridge University Press, 2002.

[35] Richard J. Nowakowski, editor. *Games of No Chance 4*, volume 63 of *Mathematical Sciences Research Institute Publications*. Cambridge University Press, 2015.

[36] Héctor Palacios and Hector Geffner. Compiling uncertainty away in conformant planning problems with bounded width. *J. Artif. Intell. Res.*, 35 :623–675, 2009.

[37] Jussi Rintanen. Complexity of planning with partial observability. In Shlomo Zilberstein, Jana Koehler, and Sven Koenig, editors, *Proceedings of the Fourteenth International Conference on Automated Planning and Scheduling (ICAPS 2004)*, pages 345–354. AAAI, 2004.

[38] Marcus Schaefer and Christopher Umans. Completeness in the polynomial-time hierarchy : A compendium. *SIGACT news*, 33(3) :32–49, 2002.

[39] Aaron N Siegel. *Combinatorial game theory*. American Mathematical Society, 2013.

[40] Christopher Umans. Hardness of approximating Σ_2^p minimization problems. In Paul Beame, editor, *Proceedings of the Fourtieth Annual Symposium on Foundations of Computer Science (FOCS 1999)*, pages 465–474. IEEE Computer Society, 1999.

[41] Bernhard von Stengel. Efficient computation of behavior strategies. *Games and Economic Behavior*, 14(2) :220–246, 1996.

[42] Bernhard von Stengel and Daphne Koller. Team-maxmin equilibria. *Games and Economic Behavior*, 21(1) :309–321, 1997.

[43] Brian Hu Zhang, Gabriele Farina, and Tuomas Sandholm. Team belief DAG : generalizing the sequence form to team games for fast computation of correlated team max-min equilibria via regret minimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the Fortieth International Conference on Machine Learning (ICML 2023)*, volume 202 of *Proceedings of Machine Learning Research*, pages 814–839. PMLR, 2023.

sETL: Outils ETL pour la construction de graphes de connaissances en exploitant la sémantique implicite des schémas de données

S. Ouelhadj[1,2], P. Champin[1], J. Gaillard[2]

[1] Univ Lyon, UCBL, CNRS, INSA Lyon, Centrale Lyon, Univ Lyon 2, LIRIS, UMR5205, F-69622 Villeurbanne, France
[2] Métropole de Lyon, Lyon, France

{firstname.lastname}@liris.cnrs.fr; jegaillard@grandlyon.com

Résumé

Nous présentons sETL, une nouvelle approche pour la construction de graphes de connaissances (GCs) en utilisant les technologies du Web Sémantique (WS). Nous abordons les défis rencontrés par la Métropole de Lyon - une collectivité territoriale française - pour assurer l'interopérabilité des données ouvertes de sa plateforme `data.grandlyon.com`*. sETL exploite la sémantique implicite des schémas de données, et fournit une boîte à outils aux praticiens de données pour enrichir sémantiquement leurs données sans requérir des connaissances spécifiques en WS. Ce travail formalise un Modèle Sémantique, introduit le concept de Bundles pour la transformation des données, et présente une implémentation Python d'opérateurs ETL de haut niveau. L'approche est comparée à celles de l'état de l'art, mettant en évidence ses caractéristiques uniques, et sa capacité à répondre aux exigences spécifiques identifiées dans l'étude.*

Mots-clés

Enrichissement sémantique, ETL, Graphe de connaissances, Schéma de données.

Abstract

We introduces sETL, a novel approach to build knowledge graphs (KGs) using Semantic Web (SW) technologies. It addresses the challenges faced by the Metropolis of Lyon - a French local authority - in ensuring interoperability of open data from its `data.grandlyon.com` *platform. sETL leverages the implicit semantics of schemas, and provides a toolkit for data practitioners to semantically lift their data without requiring specific SW knowledge. The paper formalizes a Semantic Model, introduces the concept of Bundles for data transformation, and presents a Python implementation of high-level ETL operators. The approach is compared to related works, highlighting its unique features and ability to meet specific requirements identified in the study.*

Keywords

Semantic Lifting, ETL, Knowledge Graph, Data Schema.

1 Introduction

Les données sont continuellement produites et publiées sur le web conformément à des normes et standards, afin d'améliorer leur compréhensibilité, interopérabilité et intégration [25]. La Métropole de Lyon - une collectivité territoriale française - et ses partenaires ont embrassé le mouvement des données ouvertes, et ont mis en place un point d'accès central aux données locales qu'ils produisent, appelé `data.grandlyon.com`. Les producteurs de données de la Métropole de Lyon font des efforts considérables pour améliorer continuellement la qualité des données. Parmi ces efforts figure la participation à la production et à l'utilisation de schémas de données partagés afin de normaliser les données.

Les schémas de données permettent de décrire le modèle de données des jeux de données. Ils fournissent des descriptions précises et non ambiguës des différents champs qui composent un jeu de données : les valeurs possibles, les types, le caractère obligatoire ou non du remplissage de ces champs, etc. La production de données conformes à un schéma présente de nombreux avantages tels que la validation des données, l'amélioration de leur interopérabilité et croisement, la génération automatique de documentations, et la pérennité des modèles de données [1]. Pour cette raison, plusieurs communautés de producteurs de données, de réutilisateurs, d'experts métiers et techniques ont été constituées pour le développement de schémas de données partagés [2].

Toutefois, les défis liés à l'amélioration de la compréhensibilité, l'interopérabilité et l'intégration des données sont toujours d'actualité, car la sémantique de ces schémas de données est largement implicite. En effet, si les schémas de données capturent une sémantique partagée, notamment dans les descriptions textuelles de leurs champs qui peuvent

1. https://guides.data.gouv.fr/publier-des-donnees/guide-qualite/maitriser-les-schemas-de-donnees/comprendre-les-benefices-dutiliser-un-schema-de-donnees
2. par ex. https://schema.data.gouv.fr, https://smart-data-models.github.io/data-models, https://www.futurocite.be/standardiser-les-donnees-ouvertes/

être comprises par les producteurs et consommateurs de données, cette sémantique n'est pas accessible aux machines, et est donc moins prometteuse pour l'interopérabilité et l'intégration des données à grande échelle. Un moyen de relever ces défis est de construire des graphes de connaissances (GCs) en utilisant les technologies du Web Sémantique (WS) [2], et en exploitant la sémantique implicite des schémas de données déjà disponibles et produits par les producteurs de données. Les technologies du WS nécessitent l'utilisation de RDF [27] comme modèle de données basé sur des graphes, et d'ontologies [4] pour définir formellement la sémantique. Cela peut constituer un obstacle pour les praticiens de données qui ne sont pas familiers avec les technologies du WS, mais qui manipulent généralement des données dans des formats (semi-)structurés (ex. JSON, CSV), en utilisant des schémas de données [1, 17] au lieu d'ontologies. Ces schémas sont basés sur des spécifications techniques telles que JSON Schema [32] ou Table Schema [31].

Par conséquent, notre objectif dans le contexte de la Métropole de Lyon, et plus largement pour toute organisation productrice de données, est de permettre aux praticiens de données de construire des GCs à partir de données dans des formats (semi-)structurés. Pour ce faire, nous avons identifié 5 exigences (Ri) basées sur les conclusions d'un atelier mené avec les producteurs de données de la Métropole de Lyon [24]. Ces exigences sont les suivantes :

(R1) exploiter la sémantique implicite des schémas de données existants : les participants de notre atelier connaissent bien les schémas de données, certains s'appuient déjà sur des dictionnaires de données internes, et veulent développer de bonnes pratiques pour améliorer la qualité des données ;

(R2) impliquer des praticiens de données sans compétences approfondies en WS : aucun des participants de notre atelier n'était familier avec les concepts du WS ;

(R3) être applicable à des structures de données hétérogènes : les données qu'ils produisent sont dans des formats variés (ex. CSV, JSON, GeoJSON) ;

(R4) être en mesure de s'aligner avec les ontologies existantes : ils s'intéressent aux vocabulaires partagés développés par les instituts nationaux et les organisations gouvernementales tels que l'IGN [3], afin de fournir une compréhension commune de la signification des termes utilisés (harmonisation des termes), et de relier les données en interne et en externe ;

(R5) être en mesure de produire des ontologies manquantes lorsqu'aucune n'est disponible pour décrire les données en question : ils sont moins intéressés par les vocabulaires à usage général tels que schema.org [4] et DBpedia, dont les termes sont souvent définis vaguement. Une nouvelle ontologie légère, basée sur les éléments du schéma, est considérée comme plus souhaitable pour leurs objectifs.

A partir de ces exigences, nous proposons dans cet article la boîte à outils sETL, pour « semantic ETL » (*Extract Transform Load*). Elle est basée sur des concepts et technologies d'ingénierie des données bien connus (UML, ETL, Python, Pandas dataframes) intégrés en une nouvelle approche pour tenter d'abaisser la barrière des compétences en WS requises dans la construction de GCs, permettant ainsi une exploitation plus efficace des capacités des technologies du WS. Les contributions de ce travail sont les suivantes :

1. la formalisation d'un *Modèle Sémantique* (MS) qui exprime la sémantique implicite des données à partir des schémas fournis, et l'expose en vue d'un raffinement ultérieur ;
2. la définition du concept de *Bundle* qui permet de décomposer le jeu de données et d'en associer chaque partie avec l'élément du MS lui correspondant ;
3. la mise en œuvre d'un ensemble d'opérateurs ETL de haut niveau qui permettent d'affiner les données et leur sémantique correspondante au niveau du bundle, afin de s'adapter aux contraintes des ontologies cibles et de gagner en expressivité, avant de charger ces bundles en un GC.

Dans la suite de cet article, nous commençons par présenter un exemple pour illustrer les caractéristiques de sETL, basé sur un jeu de données ouvert existant. Dans la section 2, nous détaillons la boîte à outils ETL proposée (sETL) où nous définissons les concepts de Modèle Sémantique, Bundle, Graphes de Bundles, et comment nous appliquons dans sETL les trois phases du paradigme ETL. Ensuite, dans la section 3, nous décrivons les aspects de mise en œuvre de la boîte à outils. Puis, dans la section 4, nous passons en revue l'état de l'art des approches de construction de GCs, et les comparons à sETL à travers nos 5 critères ci-dessus. Enfin, dans la section 5, nous présentons nos conclusions et les travaux futurs possibles.

Exemple fil conducteur

Cet exemple est basé sur un jeu de données ouvert à partir du Point d'Accès National français aux données de transport. Le jeu de données décrit l'emplacement géographique et les caractéristiques techniques des Infrastructures de Recharge pour Véhicules Electriques (IRVE) [5]. Il est publié au format CSV, avec un schéma [6] exprimé selon la spécification Table Schema [31]. Ce jeu de données contient 40 colonnes. Par souci de concision, nous ne considérons que 7 colonnes : nom_operateur, contact_operateur, telephone_operateur, id_station_local, nom_station, adresse_station, implantation_station. Selon le schéma fourni, implantation_station a un ensemble défini de valeurs autorisées : "Voirie", "Parking public", "Parking privé à usage public", "Parking privé réservé à la clientèle",

3. http://data.ign.fr/data.html
4. https://schema.org/
5. https://transport.data.gouv.fr/datasets/fichier-irve-gireve?locale=fr
6. https://schema.data.gouv.fr/schemas/etalab/schema-irve-statique/2.2.0/schema-statique.json

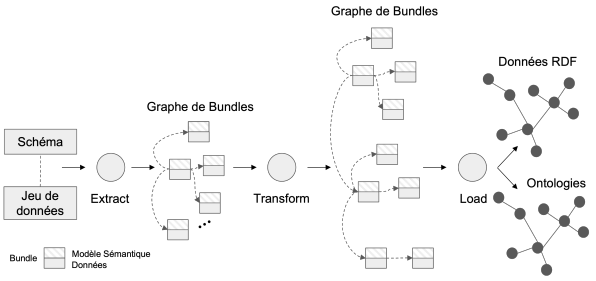


FIGURE 1 – Vue d'ensemble du processus de construction d'un graphe de connaissances avec la boîte à outils sETL.

"Station dédiée à la recharge rapide". Un échantillon du jeu de données est fourni dans le Tableau 1. Le jeu de données complet et son schéma sont disponibles dans le répertoire du projet [7].

2 sETL

La boîte à outils sETL vise à exploiter les schémas dans un processus de construction de GCs suivant le paradigme ETL. La Figure 1 présente une vue d'ensemble de ce processus. Au cours de ce processus, nous manipulons un nouveau type d'objets, appelés bundles, organisés au sein d'un graphe de bundles. Un bundle regroupe une partie des données avec son Modèle Sémantique. Ci-après, nous décrivons d'abord le Modèle Sémantique dans la section 2.1. Ensuite, nous définissons le Bundle et le Graphe de Bundles dans la section 2.2. Enfin, nous décrivons les trois phases du paradigme ETL dans le contexte de sETL. La phase *Extract* (Section 2.3) convertit un jeu de données et son schéma en un graphe de bundles. La phase *Transform* (Section 2.4) affine le graphe de bundles afin d'améliorer la sémantique des données. La phase *Load* (Section 2.5) exporte le graphe de bundles en une ontologie et des données RDF.

2.1 UML annoté : le Modèle Sémantique

Le modèle sémantique adopté dans sETL, également appelé UML annoté, est basé sur le diagramme de classes UML avec des ajouts détaillés ci-dessous. Les diagrammes de classes UML offrent un large éventail de types de composants. Pour cette première version de la boîte à outils, nous n'utilisons que les types de composants suivants : les classes (et leurs attributs), les énumérations (et leurs valeurs énumérées), et les associations. La Figure 2, partie C, illustre ces types de composants, dans un diagramme de classes UML capturant la sémantique implicite de notre exemple fil conducteur.

Ajouts de l'UML annoté

L'UML annoté étend les diagrammes de classes UML en permettant à chaque composant (classe, énumération, association, attribut et valeur énumérée) d'être annoté par un *IRI* et une *documentation textuelle* de ce composant. Les IRIs permettent de faire le lien entre l'UML annoté et les ontologies du WS. En outre, une énumération peut se voir attribuer un *type de données*, et chaque valeur énumérée peut être *alignée* avec des entités des GCs externes (ex. Wikidata, DBpedia).

Sur la base de ces considérations, nous définissons les notations suivantes pour les composants de l'UML annoté. Un Modèle Sémantique MS est un ensemble $\{MS_i \| \ i \in [0, n[\}$, où chaque MS_i est soit une $classe$, soit une $enumeration$, et où :

- chaque $classe$ a la forme $(nom,\ IRI,\ definition,\ attributs,\ associations)$;
- chaque $enumeration$ a la forme $(nom,\ IRI,\ definition,\ type,\ valeurs\_enumerees)$;
- chaque $attribut$ a la forme $(nom,\ IRI,\ definition,\ type,\ estIdentifiant)$;
- chaque $association$ a la forme $(nom,\ IRI,\ definition,\ destination)$, où $destination$ est une référence vers une $classe$ ou une $enumeration$ du MS ;
- chaque $valeur\_enumeree$ a la forme $(nom,\ IRI,\ definition,\ alignements)$;
- chaque $alignement$ a la forme $(autre\_entite,\ relation)$.

Comme tout diagramme de classes UML, un Modèle Sémantique (MS) peut être considéré comme un graphe orienté labellisé, dont les nœuds sont des classes et des énumérations, et dont les arêtes sont des associations.

2.2 Bundle et graphe de bundles

Dans sETL, les données ne sont jamais manipulées ou transformées de manière isolée : elles sont constamment liées à un Modèle Sémantique, dans des unités appelées *bundles*. En reliant chaque nœud (classe ou énumération) du Modèle Sémantique aux données correspondantes, nous obtenons une nouvelle structure : le *graphe de bundles*.

Considérons un bundle $b_i = (MS_i,\ D|MS_i)$ où MS_i est une classe ou une énumération du Modèle Sémantique, et $D|MS_i$ sont les données décrivant les instances de MS_i (notation inspirée de [3]). Un bundle peut être soit un bundle-classe ($b_i \in B_{class}$) si MS_i est une classe, soit un bundle-énumération ($b_i \in B_{enum}$) si MS_i est une énumération. Par conséquent, un graphe de bundles G_B est un graphe orienté labellisé où $B = \{(b_0, b_1, \cdots, bn);\ b_i = (MS_i,\ D|MS_i)\} = B_{class} \cup B_{enum}$.

Le Modèle Sémantique MS du graphe de bundles final représente l'ontologie sous-jacente du jeu de données en entrée. Ainsi, l'export de données RDF à partir du graphe de bundles final revient à peupler l'ontologie sous-jacente avec les données contenues dans chaque bundle du graphe de bundles. Par conséquent, nous remarquons que le graphe de bundles fournit une double vue, en faisant abstraction de la plupart des concepts du WS : une vue ontologique du jeu de données en entrée par le biais de l'UML annoté (MS), et une vue compacte et tabulaire des données RDF à produire par le biais des données contenues dans chaque bundle $D|MS_i$.

7. https://github.com/Sarra-Ouelhadj/SemanticLifting/tree/ic2024/Examples/Charging_Stations

TABLE 1 – Échantillon de données à partir de l'exemple fil conducteur

id_station_local	nom_station	adresse_station	nom_operateur	contact_operateur	telephone_operateur	implantation_station
756453	BornEco/ 63dcef1cde53 c3ec2928c1e	3 Place Maurice de Sully, Sully-sur-Loire 45600 France	Borneco | FR*BHM	technique.borneco @gmail.com	33123456789	Voirie
510419	WattzHub/ 625fc63fb907 c5cc90734800	40 Allée de la Mare Jodoin, Gif-sur-Yvette 91190 France	WattzHub | FR*SMI	contact @wattzhub.com	33185412867	Parking public

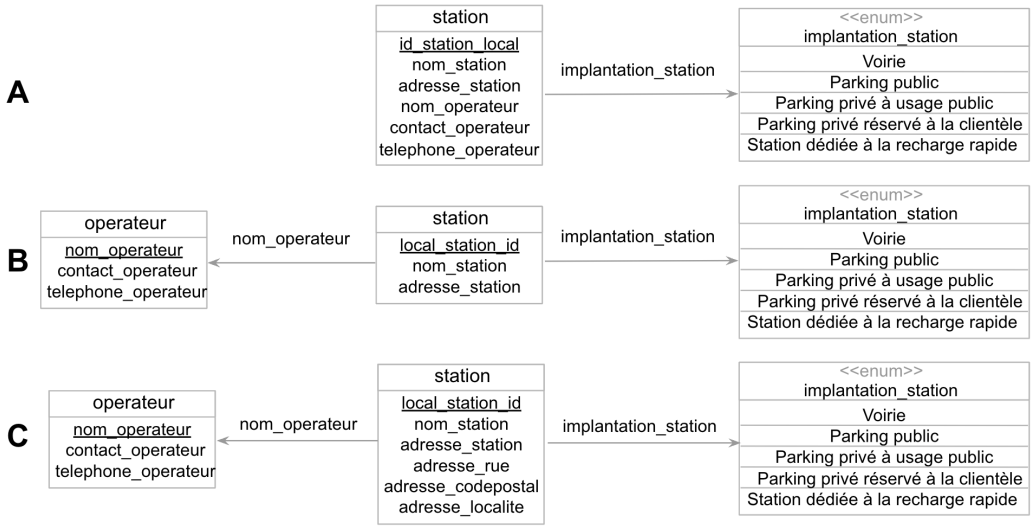


FIGURE 2 – Modèle Sémantique de notre exemple fil conducteur, à 3 phases du processus (A, B, et C).

2.3 Extracteurs

sETL extrait un graphe de bundles initial par le biais d'opérateurs appelées extracteurs à partir d'un schéma en entrée et de son jeu de données conforme. Les jeux de données dans différents formats de données sont conformes à des schémas définis à l'aide de différentes spécifications (ex. JSON Schema pour JSON, Table Schema pour CSV, etc.) Nous définissons donc un extracteur pour chaque paire de spécification de schéma et de format de données. Chaque extracteur met en correspondance chaque type d'élément du schéma source avec une construction correspondante dans notre Modèle Sémantique (MS).

Par exemple, notre extracteur Table Schema CSV appliqué sur l'exemple fil conducteur donne le résultat en Figure 2, partie A. Cet extracteur crée une classe unique pour l'ensemble de la table. Les colonnes déclarées par le schéma avec un type énuméré deviennent une association entre cette classe et une nouvelle énumération UML, également nommée d'après la colonne ("implantation_station" dans notre exemple). Toutes les autres colonnes deviennent des attributs de la classe. La $definition$ de chaque élément du MS est extraite à partir des descriptions textuelles fournies par le schéma (le cas échéant) ; il en va de même pour d'autres aspects du MS (par exemple $type$ ou $estIdentifiant$ pour les attributs). Les données liées à la classe sont la table entière. Les données liées à chaque énumération sont la colonne unique à partir de laquelle elle a été générée.

2.4 Transformeurs

Les transformeurs sont des opérateurs de raffinement disponibles pour l'utilisateur afin d'affiner le MS en un modèle plus complexe et riche sur le plan sémantique. Ces transformeurs sont appliqués sur les bundles, et sont listés dans le Tableau 2 où nous donnons de brèves descriptions de leurs effets. Par manque de place, nous ne pouvons pas donner une description détaillée de chacun d'entre eux, mais à titre d'exemple, nous donnons l'Algorithme 1 pour le transformeur *split*.

Dans l'exemple fil conducteur, nous appliquons 3 transformeurs : *split*, *apply*, et *annotate*. Premièrement, comme le bundle-classe "station" contient des données sur les opérateurs, qui sont sémantiquement différents des stations, nous divisons le bundle "station" afin d'ajouter un bundle-classe "opérateur" qui lui est lié. Les paramètres du transformeur *split* comprennent le nom (« opérateur » dans notre exemple) de la nouvelle classe, l'attribut servant d'identifiant à la nouvelle classe ("nom_operateur"), et les autres attributs et associations qui doivent être déplacés vers la nouvelle classe ("contact_operateur", et "telephone_operateur"). Le MS résultant est illustré dans la Figure 2, partie B. Les données associées à chaque bundle sont la projection correspondante du Tableau 1. Deuxièmement, nous affinons le contenu de l'attribut

TABLE 2 – Liste des transformeurs de sETL

Appliqué sur	Nom et Description
$B_{class} \cup B_{enum}$	**annotate** : assigner un IRI à un composant du MS_i. **document** : donner une définition textuelle à un composant du MS_i. **reconcile** : trouver les IRIs correspondant aux valeurs des colonnes spécifiées de chaque ligne de $D\|MS_i$ à partir des GCs externes (ex. Wikidata). **rename** : changer le nom d'un composant du MS_i, et mettre à jour les colonnes de $D\|MS_i$ en conséquence.
B_{class}	**apply** : appliquer une fonction aux valeurs des colonnes spécifiées de chaque ligne de $D\|MS_i$ et mettre à jour le MS_i en conséquence. Le(s) résultat(s) de la fonction est (sont) utilisé(s) pour remplir une (ou plusieurs) nouvelle(s) colonne(s) dans $D\|MS_i$, et le MS_i est enrichi des attributs correspondants. **mark_identifier** : marquer un attribut comme identifiant. **split** : diviser en 2 bundles-classes reliés entre eux par une association. **transform_attr_to_enum** : transformer un attribut en une énumération peuplée de toutes les valeurs de cet attribut dans $D\|MS_i$ en tant que B_{enum}.
B_{enum}	**add_value** : ajouter une valeur énumérée.

Algorithme 1 $b_1 = b_0.split(id,\ attributs,\ associations,\ nouv\_nom\_class)$

Entrées:

$b_0 = (MS_0,\ D_0|MS_0) \in B_{Class}$
$nouv\_nom\_class \in String$
$id \in MS_0.attributs$
$attributs \subset MS_0.attributs$
$associations \subseteq MS_0.associations$

soit MS_1 = class-vide
soit $D_1|MS_1$ = dataset-vide
$MS_1.nom \leftarrow nouv\_nom\_class$
$MS_1.definition \leftarrow id.definition$
$MS_0.attributs$.remove(id)
$id.estIdentifiant \leftarrow True$
$MS_1.attributs$.append(id)
$D_1|MS_1$.add_column($id.nom$)
pour $elem$ dans $attributs$ **faire**
    $MS_1.attributs$.append($elem$)
    $MS_0.attributs$.remove($elem$)
    $D_1|MS_1$.add_column($elem.nom$)
    $D_0|MS_0$.pop_column($elem.nom$)
pour $asso$ dans $associations$ **faire**
    $MS_1.associations$.append($asso$)
    $MS_0.associations$.remove($asso$)
    $D_1|MS_1$.add_column($asso.nom$)
    $D_0|MS_0$.pop_column($asso.nom$)
$b_1 \leftarrow$ BundleClass($MS_1,\ D_1|MS_1$)
$MS_0.associations$.append((nom = $id.nom$, definition = $id.definition$, destination = b_1))
retourne b_1

"adresse_station". Le schéma précise que les valeurs de cette colonne doivent contenir (1) le numéro et le nom de la rue, (2) le code postal et (3) la localité. Nous écrivons une fonction simple qui sépare ces trois composants et la transmettons au transformeur *apply*, avec les noms des trois colonnes à créer : "adresse_rue", "adresse_codepostal", "adresse_localite" (Figure 2, partie C).

Troisièmement, nous remarquons que plusieurs attributs de la classe "station" ont des termes correspondants dans l'ontologie schema.org. Nous utilisons l'opérateur *annotate* pour attacher l'IRI approprié à chacun de ces attributs. Il peut y avoir autant d'opérations de transformation que l'utilisateur le juge nécessaire. Dans notre exemple, l'utilisateur pourrait vouloir transformer les attributs de l'adresse postale en une classe distincte "adresse_postale" (en utilisant "adresse_station" comme clé primaire).

2.5 Loadeurs

Les loadeurs exportent le graphe de bundles final en un graphe de connaissances RDF. sETL fournit un loadeur d'ontologie et un loadeur de données RDF. Ces loadeurs attendent comme paramètres des espaces de noms, qui sont utilisés pour forger des IRIs pour les classes, les énumérations et les valeurs énumérées, et les instances.

Loadeur d'Ontologie. Lorsqu'un composant – classe, énumération, association, attribut ou valeur énumérée – n'est pas annoté explicitement par un IRI provenant d'une ontologie existante, un nouveau terme est créé à l'aide d'un template préconfiguré d'une ontologie (exemple dans Figure 3).

Loadeur de Données. Chaque ligne des données d'un bundle-classe représente une instance de cette classe. L'IRI de la classe est celui du MS s'il existe, sinon l'IRI forgé par le loadeur d'ontologie. Un triplet est généré pour chaque cellule du tableau des données du bundle-classe. Un exemple de triplets RDF est présenté dans le Listing 1. Le sujet de chaque ligne est l'IRI obtenu par la concaténation de l'espace de noms des instances avec la valeur de

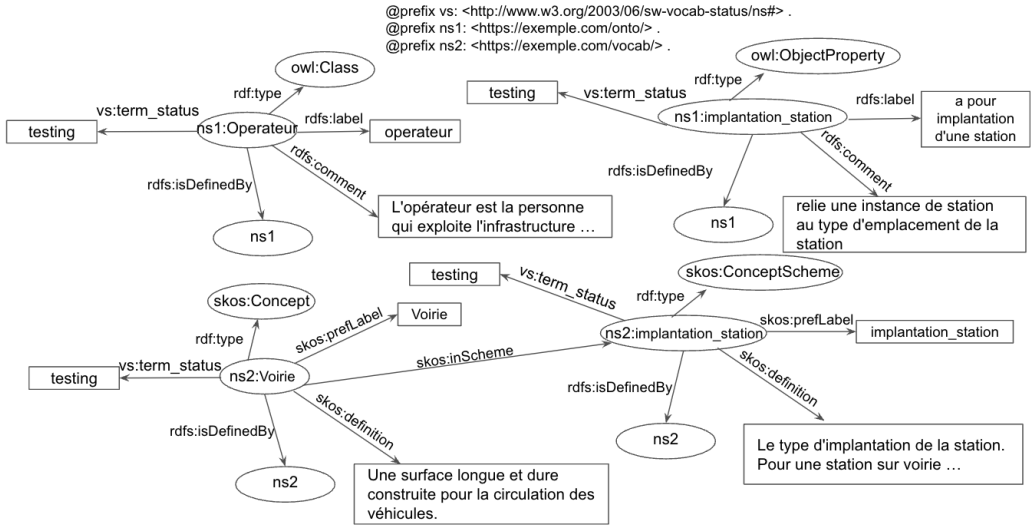


FIGURE 3 – Une partie de l'ontologie générée à partir du Modèle Sémantique du graphe de bundles illustré dans la Figure 2, partie C

```
@prefix ns1: <https://exemple.com/onto/> .
@prefix ns2: <https://exemple.com/vocab/> .
@prefix schema: <https://schema.org/> .

<https://exemple.com/id/station/756453> a ns1:
    Station ;
    schema:identifier "756453" ;
    schema:name "BornEco/63dcef1cde530c3ec2928c1e
    " ;
    schema:address "3  Place Maurice de Sully,
    Sully-sur-Loire  45600  France" ;
    schema:streetAddress "3  Place Maurice de
    Sully" ;
    schema:postalCode "45600" ;
    schema:addressLocality "Sully-sur-Loire" ;
    ns1:implantation_station ns2:Voirie ;
    ns1:nom_operateur <https://exemple.com/id/
    operateur/Borneco%20%7C%20FR%2ABHM> .

<https://exemple.com/id/operateur/Borneco%20%7C
    %20FR%2ABHM>
    a ns1:Operateur ;
    schema:name "Borneco | FR*BHM" ;
    schema:email "technique.borneco@gmail.com" ;
    schema:telephone  "33123456789" .
```

Listing 1 – Un exemple de triplets RDF générés

la colonne clé correctement échappée. Le prédicat est l'IRI de l'attribut ou de l'association correspondant à cette colonne, tel qu'annoté dans le MS ou forgé par le loadeur d'ontologie. Si la colonne correspond à un attribut, l'objet est un littéral, qui est la valeur de la colonne. Si la colonne correspond à une association, et que le bundle associé est un bundle-classe, l'objet est l'IRI identifiant l'instance correspondante au bundle-classe associé. Si la colonne correspond à une association, et que le bundle associé est un bundle-énumération, l'objet est l'IRI associé à la valeur énumérée dans la cellule actuelle.

3 Implémentation

La boîte à outils sETL est développée comme une bibliothèque Python. Les utilisateurs enrichissent leurs données en écrivant de simples scripts Python, en s'appuyant sur les classes (pour les bundles) et les fonctions (pour les opérateurs ETL) fournies par sETL. En interne, les données des bundles sont manipulées à l'aide de la bibliothèque Pandas [18] ; les loadeurs de données RDF et d'ontologies utilisent la bibliothèque RDFLib [8]. Un autre avantage de Python est la possibilité d'utiliser sETL dans une configuration interactive avec les Notebooks Jupyter [9], où les utilisateurs peuvent visualiser le graphe de bundles sous forme de diagrammes UML et inspecter les données de n'importe quel bundle, et à n'importe quelle étape de la transformation. Le code source est disponible en ligne [10], ainsi que des exemples de notebooks.

Nous avons développé 2 extracteurs jusqu'à présent : un pour les jeux de données GeoJSON conformes à JSON Schema, et un pour les jeux de données CSV conformes à Table Schema [31].

Nous avons testé avec succès sETL avec 3 jeux de données provenant de plateformes de données ouvertes françaises, chacun étant conforme à un schéma différent : les aménagements cyclables (GeoJSON/JSON Schema), lieux de stationnement (CSV/Table Schema), tous deux provenant de `schema.data.gouv.fr`, et les Infrastructures de Recharge pour Véhicules Électriques (CSV/Table Schema) provenant de `transport.data.gouv.fr`. Les résultats et pipelines correspondants sont disponibles dans le dépôt [11].

8. https://rdflib.readthedocs.io/
9. https://jupyter.org/
10. https://github.com/Sarra-Ouelhadj/SemanticLifting/
11. https://github.com/Sarra-Ouelhadj/SemanticLifting/tree/ic2024/Examples

4 Comparaison à l'état de l'art

De nombreuses approches d'enrichissement sémantique ont été proposées dans la littérature pour construire des graphes de connaissances (GCs). Ces approches peuvent être classées comme automatiques ou semi-automatiques. Les approches semi-automatiques impliquent une intervention humaine au cours d'étapes déterminées du processus de construction de GCS, tandis que les approches automatiques sont entièrement autonomes.

4.1 Approches semi-automatiques

Nous distinguons ici deux catégories : les approches déclaratives et les approches interactives. Dans la première, le mapping est fournie dans un langage déclaratif, tandis que dans la seconde, les utilisateurs construisent un mapping en interagissant avec le modèle de données.

Approches déclaratives

JSON-LD [12] et CSVW [29] permettent d'interpréter les jeux de données JSON et CSV, respectivement, en tant que RDF par le biais de fichiers de contexte. RML [7], qui étend la recommandation R2RML du W3C [6] pour prendre en charge des sources de données (semi-)structurées hétérogènes, expose les données en RDF par l'intermédiaire de mappings exprimés eux-mêmes en RDF. Le Linked data Modeling Language (LinkML) [21] est un framework de modélisation de données orienté objets basé sur une syntaxe de schéma YAML personnalisée inspirée de la structure de diagramme de classe UML. Il permet de schématiser une grande variété de formats de données en entrée, simplifiant ainsi la production de données RDF.

Au cours du processus d'enrichissement sémantique avec des approches déclaratives, des valeurs de données individuelles doivent parfois être transformées afin de s'adapter aux contraintes des ontologies cibles (ex. conversion d'unités, division ou fusion de plusieurs valeurs, etc.) Nous appelons cela une transformation fine. The Function Ontology (FnO) [20] est une description sémantique des fonctions. Elle décrit les paramètres d'une fonction, sa valeur de retour, et le problème qu'elle résout, ce qui permet de réutiliser la fonction. Les descriptions FnO s'intègrent parfaitement dans les représentations des mappings R2RML. La fonction FnO peut être utilisée comme une transformation fine des données ou comme une condition dans RML lors de l'exécution des mappings. SPARQL-Generate [16] étend le langage de requête SPARQL recommandé par le W3C [8] pour transformer des données (semi-)structurées hétérogènes en RDF à l'aide de modèles de graphes. Le mapping et les transformations fines sont exprimés grâce à l'expressivité de SPARQL. D-REPR [30] est un langage de mapping basé sur une syntaxe YAML personnalisée pour transformer des données (semi-)structurées hétérogènes en RDF. Il intègre des transformations fines de données pour le pré-traitement des données par le biais de fonctions personnalisables écrites en Python avec des paramètres codés en dur.

À l'exception de LinkML, aucune de ces approches n'exploite le schéma auquel les données en entrée peuvent se conformer (R1), ni n'est utilisable sans une bonne compréhension des technologies du WS (R2). Elles ne permettent pas non plus de générer des ontologies manquantes (R5). LinkML est différent en ce sens qu'il est basé sur un langage de schéma, qui pourrait en principe être généré à partir de schémas existants. Basé sur les concepts UML et la syntaxe YAML, il ne nécessite pas beaucoup de compétences en WS. Toutefois, les schémas LinkML sont étroitement liés à la structure des données d'origine (contrairement à nos graphes de bundles), ce qui rend difficile pour les praticiens de données de les adapter afin d'expliciter leur sémantique implicite. Par conséquent, LinkML ne satisfait que partiellement R1 et R2.

Approches interactives

OntoRefine [23] et Karma [14] permettent de faire des transformations fines des données de manière interactive à travers une représentation intermédiaire tabulaire des données en entrée. Bien que ce type de représentation intermédiaire soit familier aux praticiens de données qui n'ont pas d'expérience dans le domaine du WS, sa sémantique est pauvre par rapport aux représentations hiérarchiques ou en réseau qui mettent en évidence les concepts et leurs relations (R3), car souvent une seule ligne de structures tabulaires peut représenter plusieurs entités (voir notre exemple fil conducteur). Datalift [26], Csv2rdf4lod-automation [15], UnifiedViews [10], et LinkedPipes ETL [13] forment un autre groupe d'outils de manipulation de données qui convertissent systématiquement les données en entrée en RDF brut, puis appliquent d'autres processus d'enrichissement sémantique. Comme il faut manipuler des données RDF, l'interactivité de ces approches ne profite pas aux praticien des données qui n'ont pas de connaissances en WS (R2). En outre, aucune de ces approches n'exploite le schéma des données (R1), et ne permet de générer de nouvelles ontologies (R5).

4.2 Approches automatiques

Contrairement aux approches présentées ci-dessus, les approches automatiques visent à convertir les données sans aucune intervention humaine. Elles se répartissent en trois catégories : les mappings *hard-codés*, les mappings basés sur des schémas, et les mappings inférés.

Mappings hard-codés

Les outils dont les mappings sont hard-codés ciblent les données conformes à une spécification précise dont la sémantique a été codée en dur dans le système. gtfs-csv2rdf [12] prend en charge les données en entrée conformes à GTFS (General Transit Feed Specification) [13], et les convertit selon une ontologie prédéfinie (Linked GTFS vocabulary) [14] développée spécifiquement à cette fin. Guid-O-Matic [15] convertit les données en entrée conformes au stan-

12. https://github.com/OpenTransport/gtfs-csv2rdf
13. https://gtfs.org/
14. https://github.com/OpenTransport/linked-gtfs/blob/master/spec.md
15. https://github.com/baskaufs/guid-o-matic

dard Darwin Core [16] en RDF.

Mappings basées sur des schémas

Les outils avec mappings basés sur des schémas automatisent la génération de règles de mapping à partir de schémas. MIRROR[19], AutoMap4OBDA[28] et BOOTOX[11] sont des exemples de ces outils qui ciblent les bases de données relationnelles et, à notre connaissance, il n'existe aucune solution prenant en charge d'autres types de schémas largement utilisés.

Mappings inférés

Les solutions avec mappings inférés (également connues sous le nom d'annotation sémantique) automatisent la construction de GCs sans utiliser des règles de mapping. Au lieu de cela, elles infèrent automatiquement la correspondance des données avec un référentiel prédéfini. Par exemple, les travaux concourant au défi SemTab [17] [5, 22, 9] mettent en correspondance des éléments de données tabulaires (*i.e.*, cellules, colonnes, lignes) à des éléments sémantiques (*i.e.*, entités, classes, propriétés) provenant de GCs collaboratifs et généralistes (ex. Wikidata, DBpedia).

Aucune de ces approches automatiques n'implique d'intervention humaine (R2), ni n'est applicable à des structures de données hétérogènes (R3) puisqu'elles se concentrent sur une structure de données spécifique. Elles ne permettent pas non plus de générer de nouvelles ontologies (R5). De plus, les solutions avec des mappings inférés n'exploitent pas le schéma des données (R1). Cependant, les solutions basées sur des schémas satisfont partiellement R1 car elles sont basées sur les schémas sous-jacents des bases de données relationnelles.

4.3 Notre proposition

Nous montrons maintenant comment l'approche que nous proposons, sETL, répond aux 5 exigences présentées dans la Section 1. Premièrement, sETL tire son originalité de l'utilisation des schémas de données dans l'enrichissement sémantique des données (R1). Elle extrait le graphe de bundles initial - la représentation intermédiaire des données dans le système - à partir d'un jeu de données en entrée et de son schéma. Le Modèle Sémantique (MS) de ce graphe initial représente une transcription directe de la sémantique explicite du schéma. Deuxièmement, afin d'affiner la sémantique des données, sETL fournit un ensemble complet de transformeurs que les praticiens de données peuvent appliquer de manière interactive au graphe de bundles. Ces transformeurs garantissent la conformité continue entre le MS et les données de chaque bundle. Étant basée sur des concepts et des technologies d'ingénierie de données bien connus (UML, ETL, Python, Pandas), sETL a une faible barrière à l'entrée pour les praticiens n'ayant pas d'expérience en WS (R2). Troisièmement, le modèle de graphe de bundles est indépendant de tout format de jeux de données en entrée ou de toute spécification de schéma. Par conséquent, le système peut être étendu pour prendre en charge n'importe quelle structure de données en entrée (R3), ce qui convient à la nature hétérogène des données. Quatrièmement, sETL permet la réutilisation d'ontologies existantes en annotant le MS (R4), lorsque des vocabulaires avec la sémantique correspondante ont été identifiés, ou la génération de nouvelles ontologies légères (R5), dans les cas où aucun tel vocabulaire n'est disponible. Après l'initialisation du premier graphe de bundles au cours de la phase *Extract*, les données RDF et l'ontologie peuvent être matérialisées à tout moment. Combiné à l'utilisation interactive de sETL, cela permet un processus d'enrichissement sémantique incrémental, qui peut être utile pour des jeux de données complexes.

16. https://dwc.tdwg.org/
17. SW Challenge on Tabular Data to Knowledge Graph Matching

5 Conclusions et perspectives

Dans cet article, nous avons proposé la boîte à outils sETL qui relève le défi de l'enrichissement sémantique des données dans le contexte des organisations productrices de données n'ayant pas d'expertise étendue en WS. Cette approche a réussi à répondre aux 5 exigences que nous avons identifiées lors d'un travail antérieur réalisé avec un groupe de producteurs de données de la Métropole de Lyon. sETL permet la spécification d'un workflow complet d'enrichissement sémantique, et la production de données RDF qui en résultent. Elle comprend (1) un Modèle Sémantique étendant les diagrammes UML, pour capturer la sémantique implicite des schémas de données existants, (2) la notion de Bundle, qui groupe les éléments du MS avec leurs données correspondantes, et (3) des opérateurs ETL de haut niveau, qui font abstraction de la plupart des concepts du WS, pour le bénéfice des praticiens de données qui n'ont pas de connaissances approfondies en WS. Nous avons pu appliquer sETL à trois jeux de données complets provenant de plateformes de données ouvertes françaises.

Dans les travaux futurs, nous souhaitons étendre les fonctionnalités de la boîte à outils proposée à différents niveaux. Tout d'abord, nous souhaitons ajouter de nouveaux extracteurs pour d'autres spécifications de schémas et de formats de données (ex. les schémas de bases de données relationnelles), permettant aux utilisateurs de traiter des jeux de données plus variés et complexes. Nous souhaitons également enrichir les transformeurs. En particulier, nous souhaitons tirer parti, dans l'opérateur *reconcile*, des méthodes existantes de mapping inférés (annotation sémantique), pour faciliter la tâche d'alignement des valeurs dans les données tabulaires avec les concepts de GCs ouverts tels que Wikidata. Nous souhaitons également étudier l'ajout d'une interface graphique à sETL afin de faciliter son utilisation par les personnes qui ne maîtrisent pas l'écriture de scripts ou workflow de base sous forme de code. Enfin, nous souhaitons étudier l'intégration de sETL dans des outils existants (ex. des outils de nettoyage de données en amont, et des bases de données orientées graphe en aval) afin d'étendre les workflows existants, et de rendre sETL plus utilisable dans un environnement de production.

Références

[1] G. Aldebert and A. Augusti. schema.data.gouv.fr - An Open Data Schema Catalog for France, May 2020.

[2] S. Auer. Semantic integration and interoperability. In *Designing Data Spaces*, chapter 12, pages 195–210. Springer, Cham, 2022.

[3] F. Bariatti. *Mining Tractable Sets of Graph Patterns with the Minimum Description Length Principle*. Theses, Université de Rennes 1, November 2021.

[4] B. Chandrasekaran and et al. What are ontologies, and why do we need them ? *IEEE Intelligent Systems*, 14(01) :20–26, jan 1999.

[5] V. Cutrona and et al. Results of SemTab 2021. In *SemTab 2021*, volume 3103 of *CEUR-WS.org*, October 2021.

[6] S. Das and et al. R2RML : RDB to RDF Mapping Language. Technical report, W3C, 2012.

[7] A. Dimou and M. Vander Sande. RDF Mapping Language (RML). Technical report, RML.io, 2020.

[8] S. Harris and A. Seaborne. SPARQL 1.1 Query Language. Technical report, W3C, 2013.

[9] V-P. Huynh and et al. DAGOBAH : Table and Graph Contexts For Efficient Semantic Annotation Of Tabular Data. In *SemTab 2021*, volume 3103 of *CEUR-WS.org*, October 2021.

[10] K. Janowicz and et al. Unifiedviews : An etl tool for rdf data management. *Semantic Web*, 9(5) :661–676, jan 2018.

[11] E. Jiménez-Ruiz and et al. BootOX : Practical Mapping of RDBs to OWL 2. In *ISWC 2015*, pages 113–132. Springer, 2015.

[12] G. Kellogg and et al. JSON-LD 1.1. Technical report, W3C, July 2020.

[13] J. Klímek and et al. LinkedPipes ETL : Evolved Linked Data Preparation. In *Proc. ESWC 2016*. Springer, 2016.

[14] C. A. Knoblock and et al. Semi-automatically Mapping Structured Sources into the Semantic Web. In *Proc. ESWC 2012*, pages 375–390, Germany, 2012. Springer.

[15] T. Lebo and G.T. Williams. Converting governmental datasets into linked data. In *Proc. I-SEMANTICS 2010*, USA, 2010. ACM.

[16] M. Lefrançois and et al. A SPARQL extension for generating RDF from heterogeneous formats. In *Proc. ESWC 2017*, volume 10249, Slovenia, May 2017. Springer.

[17] Local Government Association. Open data | LG Inform Plus - Schemas, April 2023.

[18] W. McKinney. Data Structures for Statistical Computing in Python. In *Proc. SciPy*, 2010.

[19] L. F. de Medeiros and et al. MIRROR : Automatic R2RML Mapping Generation from Relational Databases. In *Proc. ICWE 2015*, pages 326–343. Springer, 2015.

[20] B. De Meester and et al. Implementation-independent function reuse. *Future Generation Computer Systems*, 110, 2020.

[21] S. Moxon and et al. The Linked Data Modeling Language (LinkML) : A General-Purpose Data Modeling Framework Grounded in Machine-Readable Semantics. In *Proc. ICBO 2021*, volume 3073 of *CEUR-WS.org*, 2021.

[22] P. Nguyen and et al. MTab4DBpedia : Semantic Annotation for Tabular Data with DBpedia. *Semantic Web Journal*, 2022.

[23] OntoText. Ontotext Refine, February 2023.

[24] S. Ouelhadj and et al. Méthode pour enrichir sémantiquement les données en utilisant l'UML annoté. EGC 2023, January 2023. Poster.

[25] M. Page and et al. *Open data maturity report 2023*. Publications Office of the European Union, LU, 2023.

[26] F. Scharffe and et al. Enabling linked data publication with the Datalift platform. In *AAAI Workshop on Semantic Cities*, Canada, July 2012. AAAI.

[27] G. Schreiber and Y. Raimond. RDF 1.1 Primer. W3C Working Group Note, W3C, June 2014.

[28] Á. Sicilia and G. Nemirovski. AutoMap4OBDA : Automated Generation of R2RML Mappings for OBDA. In *EKAW 2016*, page 577–592. Springer, 2016.

[29] J. Tennison. CSV on the Web : A Primer. Technical report, W3C, 2016.

[30] B. Vu and et al. D-repr : A language for describing and mapping diversely-structured data sources to rdf. In *Proc. K-CAP 2019*, USA, 2019. ACM.

[31] P. Walsh and R. Pollock. Table Schema. Technical report, Frictionless Standards, October 2021.

[32] A. Wright and et al. JSON Schema : A Media Type for Describing JSON Documents. Internet Draft 01, IETF, June 2022.

CapData Opéra : faciliter l'interopérabilité des données des maisons d'opéra

E. Peyre[1], F. Amarger[2], N. Chauvat[2]

[1] Réunion des Opéras de France, ROF

[2] Logilab

eudes.peyre@rof.fr ; {prenom.nom}@logilab.fr

Résumé

Le projet CapData Opéra, mené à l'initiative de la ROF[1] *utilise les technologies du Web sémantique comme fondement d'une solution de structuration et de diffusion des données culturelles capable de répondre aux enjeux de développement des publics, de soutien à la création artistique et d'accès à la culture.*

Cette solution de mutualisation permet d'interroger les données produites par plusieurs acteurs du domaine pour, par exemple, connaître la programmation et la circulation d'une œuvre ou d'une production entre plusieurs maisons d'opéra.

Pour faciliter la publication des données produites par chaque maison d'opéra, la ROF propose une ontologie du domaine, des référentiels, une infrastructure de publication, des outils et de l'accompagnement humain.

Dans cet article, nous présentons les objectifs et les moyens mis en œuvre pour fédérer des données hétérogènes, nous faisons un retour sur expérience en abordant les aspects techniques et la gestion, et nous décrivons les résultats actuels et les perspectives de ces travaux.

Mots-clés

ROF, Opéras, RDF, Interopérabilité

Abstract

The "CapData Opéra" project, initiated by ROF (Réunion des Opéras de France - French Opera Association) and supported by the French Ministry of Culture, uses Semantic Web technologies to share cultural data with the public and the artistic community.

The aim is to aggregate data produced by various domain actors to make it globally searchable. This highlights previously invisible data, such as the exchange of creative works and performances between opera houses. To achieve this, an ontology has been designed to define a common vocabulary and implement data interoperability objectives. This ontology is aligned with schema.org, and we are working to align additional models. A set of SHACL rules has been created to validate the data before publication. A dedicated tool, Rodolf, has been developed to monitor the RDF publishing process. This tool is used to execute the process and track which sources have been uploaded to the SPARQL endpoint, including upload times and any errors encountered. Exporting RDF data can be challenging for institutions unfamiliar with Semantic Web technologies, so a dedicated Software Development Kit (SDK) has been developed to assist web developers in exporting CapData RDF data even if they lack experience in this area.

In this presentation, we aim to share with the SWIB community the objectives and solutions we have found to federate heterogeneous data. We will present feedback on this project, focusing on technical and management aspects, and then describe the results we have achieved and the future of this project.

Keywords

ROF, Opera, RDF, Interoperability

1 Introduction

Les structures et les établissements culturels du spectacle vivant ont connu à la fin des années 1990 une profonde révolution liée aux transformations induites par le développement d'Internet et du numérique. Deux décennies plus tard, les enjeux liés à la diffusion et aux échanges des données produites prennent une importance majeure.

Si les stratégies et solutions développées par les politiques culturelles et les acteurs de ce secteur, dont les maisons d'opéra, ont été menées de manière relativement homogène et précoce en réponse aux enjeux de démocratisation, de création, de valorisation et de médiation auprès des publics, les problématiques liées à la gestion et au partage des données produites n'ont émergé que plus récemment.

Pour mettre en valeur leur programmation et interagir efficacement avec leur public, les services au sein des maisons d'opéra sont amenés à manipuler quotidiennement de nombreux outils numériques, tel que des CMS (*Content Management System*) avec lesquels sont gérés les informations diffusées sur leur site web, des logiciels dédiés à la gestion de billetterie, à la production ou encore de multiples réseaux sociaux. La faible interopérabilité de ces systèmes, couplée à la mise en place de modélisations spécifiques au sein de chaque établissement, conduit à une complexité im-

1. Réunion des Opéras de France

portante dès lors que les données doivent être échangées ou croisées avec d'autres acteurs.
Par exemple, une requête permettant d'obtenir la programmation lyrique ou chorégraphique des maisons d'opéra pour la saison 2023/2024 reste aujourd'hui sans réponse satisfaisante. Ceci est dû au fait que les données, et surtout les modélisations des données, ne sont pas uniformisées.
Dans ce contexte, la Réunion des Opéras de France (ROF), réseau national des maisons d'opéra, scènes et compagnies lyriques, développe au sein de la mission ressources et développement numérique, le projet CapData Opéra. Porté en réponse aux enjeux et besoins identifiés auprès de ses membres et politiques culturelles, ce projet vise au développement d'une solution mutualisée et hautement réplicable afin de favoriser l'interopérabilité, l'échange de données et leur valorisation auprès des publics.
Dans cet article, nous aborderons le sujet du partage de données et de l'intéropérabilité en mettant en lumière les différentes facettes du projet CapData Opéra. Nous débuterons par une présentation du projet et son contexte, en expliquant les objectifs visés, le choix de l'architecture et les solutions déployées pour répondre aux besoins détectés auprès des maisons d'opéra. Nous détaillerons ensuite les outils spécifiquement développés pour faciliter la réalisation de ce projet, en soulignant le rôle essentiel de l'ontologie dans le processus de création et de mise en œuvre. Enfin, nous conclurons en évoquant les perspectives futures du projet, dont l'importance d'une approche partagée, transversale et multi-échelles.

2 Contexte et besoins

Chaque maison d'opéra produit et diffuse sur son site web, ses réseaux sociaux et auprès de la presse et partenaires, sa programmation artistique et des ressources médias (vidéos, photographies, audios et textes) à destination des publics. En leur sein, les services, dont ceux de production ou de communication, produisent des données et des métadonnées de programmation et médias qui peuvent être similaires, mais qui sont souvent saisies à de multiples reprises et stockées dans des bases de données ne permettant que trop faiblement l'échange d'information (bases silotées et faiblement interopérables).
L'absence d'une stratégie généralisée et commune de standardisation, d'identification des données et de mise en place de référentiel au sein des maisons d'opéra et plus largement des lieux de programmation du spectacle vivant, représente un véritable frein au développement de la découvrabilité et à la diffusion des créations artistiques et des contenus auprès des publics.
Il n'existe en effet pas d'identifiant normalisé pour la gestion des productions de spectacle vivant, alors que le secteur du livre utilise l'ISBN [2] au niveau international et l'industrie musicale dispose de l'ISRC [3] utilisé par exemple pour identifier les morceaux diffusés sur les plateformes en ligne[9]. La généralisation de l'usage d'identifiants publics, comme l'ISNI [4] pour les artistes, permettrait par exemple de simplifier la diffusion des mentions et la gestion des droits lors d'une diffusion, ainsi que de développer la transparence lors de la diffusion des médias auxquels ils participent, comme l'ont illustré les expériences de diffusion des données de programmation des maisons d'opéra sur les espaces « #Culture chez nous », ou au sein de l'application du Pass Culture.
La diffusion des représentations programmées par les maisons d'opéra passe aujourd'hui par une succession de saisies manuelles : les services sont invités à saisir pour chaque réutilisation leurs offres d'événement, site web, billetteries, agendas culturels, applications. Malgré l'existence d'une API pour le Pass Culture, la faible structuration des données et interopérabilité au sein des systèmes conduit les services une nouvelle fois à une saisie majoritairement manuelle. L'enquête préliminaire a fait ressortir une moyenne de six doublons de saisie pour les données de programmations, traduisant un fort besoin, pour les équipes et pour la visibilité des contenus, de mise en place d'une solution efficace et puissante pour l'exposition des données. La faible exposition des données des maisons d'opéra sur les services innovants de diffusion musicale représente également un frein considérable à l'émergence de nouvelles expériences susceptibles de répondre aux besoins et usages du public.
Le développement de données structurées et leur exposition depuis les établissements culturels apparaissent donc comme un levier extrêmement puissant pour accroître la visibilité des contenus lyriques et chorégraphiques au sein des nouveaux modes de diffusion, dont les plateformes de *streaming*, enceintes connectées et services innovants. Ce même constat apparaît dans les « freins structurels technologiques à dépasser » de la « Mission exploratoire sur les métavers »[1] commandée par le ministère de l'Économie, des Finances et de la Relance, le ministère de la Culture et le secrétariat d'État chargé de la Transition numérique et des Communications électroniques.
Outre les problématiques d'échanges transversaux, de doublons de saisies des données, l'analyse comparative des services existants a permis d'observer trois freins supplémentaires à l'exposition et la diffusion des données culturelles :

1. Faute de donnée standardisée et exposée depuis les maisons d'opéra, une majorité de services numériques externes est contrainte de demander aux services des maisons une saisie manuelle supplémentaire ou un export spécifique non standardisé des données. Ce fonctionnement entraîne une surcharge importante de travail pour les équipes et limite le rayonnement des données culturelles.
2. Dans le cas minoritaire où le service externe propose une API pour collecter les données produites par les maisons, l'absence de standardisation des données impose des développements spécifiques, faiblement réplicables et un coût financier pour chaque structure. Ce constat fait ressortir un coût important pour les finances publiques sans bénéfices

2. Voir https://www.isbn-international.org/
3. Voir https://isrc.ifpi.org/
4. Voir https://isni.org/

de réplicabilité et de ruissellement à l'ensemble des établissements culturels.

3. Pour combler ces problématiques d'absence de standardisation et d'exposition des données, plusieurs agrégateurs et services utilisent, la technique du *scraping* pour collecter les données de programmation. Cette technique, à l'impact environnemental négatif, n'est pas satisfaisante et implique des développements spécifiques non pérennes pour chaque établissement culturel.
4. Enfin, la protection de la souveraineté des établissements culturels sur les données qu'ils produisent apparaît comme prioritaire. Celle-ci passe par le développement de leur capacité à disposer et à exposer leurs données en toute autonomie.

La mise en place d'une solution mutualisée a pour objectif d'optimiser les coûts d'investissement et de fonctionnement, tout en favorisant son déploiement au sein des maisons d'opéra et potentiellement des établissements intéressés du secteur du spectacle vivant.

3 Projet CapData Opéra

Initié en 2022, le projet CapData Opéra est porté par la ROF en partenariat avec l'Opéra National de Bordeaux, le groupe de travail numérique de la ROF, et le réseau TMNlab. Une première expérimentation a été réalisée dans le cadre de l'appel à projets "Découvrabilité en ligne des contenus culturels francophones"[7] en 2023. Forte de cette première phase, la ROF s'est associée à 6 maisons d'opéras et au réseau TMNlab pour lancer le projet CapData Opéra - France 2030[12], qui s'est inscrit dans le Programme d'investissements d'avenir (PIA4) - *"Expériences augmentées du spectacle vivant"*, une opération soutenue par l'Etat et opérée par la Caisse des Dépôts. Ce nouveau chantier vise à déployer et à industrialiser la solution à plus grande échelle. Ces deux projets font appel à plusieurs prestataires et expertises techniques afin d'assurer la mise en place de la solution de valorisation des données auprès des publics. Ils visent à proposer des solutions déployables au sein des maisons d'opéra participantes, qui peuvent déverrouiller les principaux freins à l'échange et la réutilisation des données. Chaque étape de la chaîne de circulation des données, de leur structuration à leur exposition et réutilisation, fait l'objet d'un travail spécifique et de la mise en place de solutions hautement réplicables, incluant des outils techniques, un accompagnement des partenaires et des prestataires, ainsi qu'une documentation appropriée.

L'exemple des maisons d'opéra illustre les défis rencontrés dans la gestion des données inter et intra sectorielle. Historiquement, chaque maison d'opéra a élaboré son propre système d'information, caractérisé par des schémas et des formats de données hétérogènes.

Depuis 2004, les technologies du Web Sémantique, promues par le World Wide Web Consortium (W3C)[5], offrent une voie vers une interopérabilité accrue grâce aux standards de la famille RDF (Resource Description Framework)[11]. L'adoption d'un format tel que le RDF, et l'utilisation d'une ontologie, ou modèle de données commun, facilitent l'intégration des données en permettant à chaque contribution de s'aligner sur cette ontologie unifiée, assurant ainsi l'interopérabilité et la circulation des données.

Outre la facilitation de l'échange de données, cette approche soutient la souveraineté de chaque maison d'opéra sur ses propres données, en lui permettant de contrôler la manière dont elles sont partagées. Dans ce sens, le choix de l'architecture sélectionnée vise à répartir de manière équilibrée et dans le respect de la souveraineté de chaque partenaire l'enjeu de responsabilité de publication. Les maisons participantes publient elles-mêmes leurs données au format RDF, se rapportant spécifiquement à leurs besoins et stratégies de diffusion. Ces données sont ensuite collectées au sein d'un entrepôt SPARQL. La gestion mutualisée de cet entrepôt favorise l'interrogation des informations issues de l'ensemble des maisons d'opéra participantes au projet.

Cette démarche met en lumière l'importance du Web Sémantique dans le renforcement de l'interopérabilité entre systèmes d'information hétérogènes. Elle souligne également le rôle crucial de standards ouverts et partagés, comme le RDF, dans la construction d'un écosystème de données cohérent et efficace, bénéfique à l'ensemble du secteur culturel.

3.1 L'ontologie

Le développement d'une ontologie commune a pour objectif de favoriser le partage, les réutilisations et la découvrabilité des contenus culturels auprès des publics. Préalables à l'étape d'exposition des données et d'élaboration de connecteurs pour la diffusion des données depuis les systèmes d'informations respectifs des maisons participantes, les travaux ontologiques permettent également de définir le périmètre de connaissances partagées entre les maisons d'opéra et plus largement avec le secteur culturel. Sa mise en place permet de simplifier les échanges, tant au sein des services des maisons d'opéra qu'auprès des collectivités publiques et des industries culturelles et créatives (ICC).

La réalisation d'un état de l'art sur les ontologies existantes et en capacité de répondre aux besoins des maisons d'opéra et des politiques culturelles a permis de détecter deux ontologies candidates.

La première est l'ontologie schema.org[6], qui présente de nombreux avantages. Tout d'abord, son approche englobante via l'adoption de définition large des concepts, apparaît comme très efficace pour répondre à de nombreuses situations, tout particulièrement pour le partage et la description des événements (schema.org/Event). Elle semble donc particulièrement adaptée pour la diffusion et la découvrabilité des dates de représentations. De plus sa documentation et son utilisation par les principaux moteurs de recherche pour l'indexation des contenus du web l'ont rendu très populaire auprès des équipes de communication et prestataires en charge des sites web.

5. Voir `https://www.w3.org/`

6. `https://schema.org`

Trois écueils nous ont conduits à poursuivre la phase de recherche ontologique. Premièrement, bien que cette ontologie décrive, de manière détaillée, un événement, une organisation ou même une œuvre, il apparaît que les étapes préliminaires et nécessaires à l'élaboration d'un spectacle ne sont que partiellement décrites. La notion de production, essentielle au sein du spectacle vivant, car englobant l'ensemble des actions menant à la représentation, telles que la conception des décors et des costumes ou la gestion des distributions, sont absentes de schema.org. Deuxièmement, le concept de producteur apparaît trop large pour une description fine et essentielle du rôle et de l'implication juridique de chaque partie prenante dans l'élaboration d'une production. Troisièmement, l'absence de traduction se révèle être un frein dans la capacité de représenter finement la vision développée par les politiques culturelles et de la représentation de la diversité.

La seconde ontologie que nous avons considérée est IFLA-LRM[15], qui est une référence en termes de gestion de connaissances dans le monde de la Culture de manière générale. Les concepts représentés sont bien plus proches de ce que nous souhaitons représenter pour le projet CapData Opéra. Si la notion d'œuvre y apparaît comme centrale, l'ontologie est néanmoins élaborée pour répondre aux objectifs de la gestion des ressources bibliographiques, ce qui ne nous a pas semblé parfaitement adapté aux besoins et à la description du spectacle vivant.

Les spectacles vivants, incluant le théâtre, la danse, la musique live, et d'autres formes d'art performance, nécessitent en effet des informations et des métadonnées détaillées et intrinsèques à la description d'une production artistique ou d'un spectacle, par exemple de ses décors, costumes, montages, aux effectifs et compositions des formations, voir également des publics.

Les recherches sur l'ontologie ont également mis en lumière l'existence d'initiatives et travaux de recherches similaires à l'international, dont ceux du groupe de "Performing Arts Information Representation Community Group", néanmoins celui-ci semble inactif depuis quelques années. Cet intérêt à l'échelle internationale et visible lors de la journée "Rendez-vous France-Québec sur la découvrabilité des contenus culturels francophones" [4] de l'édition 2023 du MTL Connecte, a permis d'entamer une réflexion sur les enjeux et la pertinence d'une action coordonnée, voire mutualisée, du chantier ontologique.

Nous avons fait le choix de développer l'ontologie CapData Opéra afin de proposer une description des connaissances qui répond aux besoins détectés auprès des maisons d'opéra et plus largement du secteur du spectacle vivant. Nous l'avons voulue complémentaire des autres modèles et avons mis en place directement dans l'ontologie des alignements vers l'IFLA-LRM et le schema.org. Nous permettons ainsi une représentation fidèle aux besoins du domaine, tout en rendant aisée l'utilisation de l'ontologie de référence dans le domaine culturel et de l'ontologie de référence pour l'indexation par les moteurs de recherche sur le Web.

Pour faciliter la réutilisation de cette ontologie par les maisons d'opéra, nous la documentons et la publions à l'URL `https://ontologie.capdataopera.fr`. Cette page permet de représenter les différentes versions de l'ontologie (actuellement la version 1.7) et la date de publication. Il y a, pour chaque version, une documentation générée par Widoco[5] et une représentation graphique générée avec WebVOWL[10]. De plus, les URI utilisées dans l'ontologie pointent sur ce site, ce qui favorise la négociation de contenu et la récupération des formats HTML ou RDF suivant la requête HTTP.

En parallèle de cette étape, les référentiels et vocabulaires contrôlés, utilisés au sein de l'ontologie, ont fait l'objet d'une exposition sur l'entrepôt SPARQL dédié au projet. Cette action participe à accroître l'interopérabilité et la découvrabilité via l'usage de définitions et de vocabulaires partagés au sein du réseau et plus largement du secteur culturel.

Enfin, des règles SHACL de vérification sont présentes sur le même site, avec une documentation adaptée [7], pour permettre aux maisons d'opéra de valider leurs données.

L'un des avantages du processus de modélisation et des outils mis en place autour de la publication de l'ontologie est sa grande agilité. Nous avons pu, en effet, confronter la modélisation de manière concrète aux besoins des équipes des maisons participantes, prestataires, services externes et utilisateurs. Une modification de l'ontologie était rapidement intégrée dans un outil d'export de données en RDF grâce à la documentation disponible et aux outils de vérification des données. La réalisation manuelle des alignements a permis d'améliorer de manière itérative le modèle. Cette approche nous a permis de nous rendre compte très rapidement des écueils et de pouvoir corriger l'ontologie de manière agile afin d'obtenir une version stabilisée et intégrable au sein des connecteurs et des applications développées en parallèle.

3.2 Le suivi de production

Nous considérons, comme présenté en introduction de cette section, que la maison d'opéra gère la publication de ses propres données RDF. Cette publication prend la forme d'un fichier RDF disponible à une URL donnée contenant l'intégralité des données. Ce fichier est mis à jour régulièrement par un export régulier de la part des maisons d'opéra. Pour permettre l'interrogation de toutes ces données, nous souhaitons les récupérer pour les publier dans un entrepôt SPARQL dédié. Pour cela, un script s'exécute tous les jours pour récupérer ces fichiers mis à disposition derrière les URL de chaque maison. Un ensemble de traitements de nettoyage de données sont appliqués, comme l'effacement des espaces avant et après les valeurs littérales, la détection de l'utilisation d'une valeur à la place d'un identifiant de référentiels (par exemple le code pays), etc. Ces traitements sont appliqués systématiquement à chaque récolte des données. De plus, certaines données sont alignées sur les référentiels fournis par la ROF afin de faciliter l'interopérabilité. Enfin, les données sont validées en utilisant les règles SHACL liées à l'ontologie. De cette manière nous obtenons quatre graphes différents pour chaque maison d'opéra lors

7. générée grâce à `https://shacl-play.sparna.fr/play/doc`

d'une récolte de données :

— un graphe contenant les données initiales ;
— un graphe contenant les données nettoyées ;
— un graphe contenant les alignements avec les données de la ROF ;
— un graphe contenant les triplets du rapport de validation SHACL [8].

Ces graphes sont ensuite envoyés dans un entrepôt SPARQL mutualisé pour toutes les maisons d'opéra et administré par la ROF (`https://sparql.capdataopera.fr/`).

De cette manière, toutes les données sont mises au même endroit et différents graphes nommés permettent de récupérer les données qui nous intéressent. De plus, il devient trivial de faire des requêtes entre plusieurs maisons d'opéra à partir du moment où toutes les données sont sur le même entrepôt. Le choix de cette architecture répond également à une analyse approfondie des coûts de maintenance à moyen et long terme. L'étude de faisabilité du projet a en effet mis en lumière le coût non soutenable que représentait la mise en place d'API au sein de chaque maison d'opéra pour la gestion de la diffusion et la récupération des données. Une approche non mutualisée engendrait une démultiplication des coûts financiers pour les établissements et les collectivités publiques.

Afin de mettre en place une boucle rétroactive bénéfique pour les maisons d'opéra, nous avons écrit des requêtes SPARQL pour qu'elles récupèrent leurs données, nettoyées, alignées et validées. Il leur est donc possible de réinsérer ces données dans les systèmes d'information d'origine pour augmentant la qualité de leurs données. Par exemple en ajoutant pour une personne les identifiants ROF, ISNI ou ARK issus de l'alignement.

4 Les outils

Au cours du projet, nous avons poursuivi une démarche visant à industrialiser toute la chaîne de production et nous avons constaté des manques parmi les outils disponibles sous une licence de logiciel libre. Nous avons essayé de cartographier cette situation à travers l'approche SemGraph [9] qui, pour chaque étape de la chaîne, suggère des outils possibles. Nous avons développé ou fait évoluer certains outils quand nous l'avons jugé utile et que cela était dans nos moyens.

4.1 Publication de l'ontologie

Lorsque nous avons souhaité publier l'ontologie pour la rendre disponible sur le Web, nous avons identifié des portails comme OntoPortal[8] pour héberger nos modèles et avons envisagé de simplement les publier derrière un serveur Web standard. Néanmoins, nous avons souhaité pouvoir gérer plus finement les versions, avoir une documentation et de la négociation de contenu qui permettent d'accéder à la documentation directement, tout cela intégré à nos

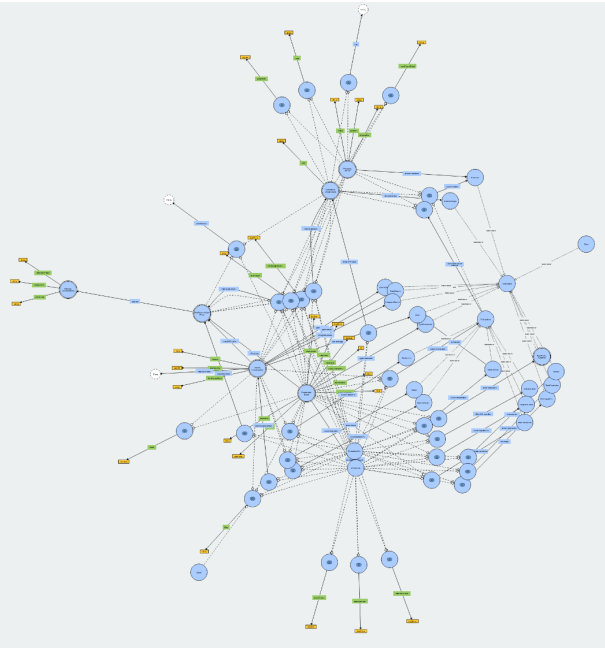


FIGURE 1 – L'ontologie CapData Opéra représentée avec WebVOWL

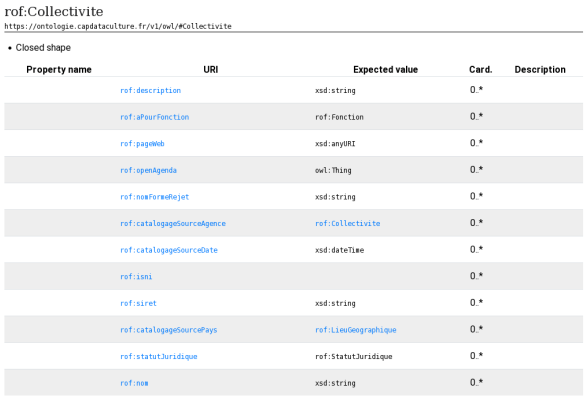
outils de déploiement continu habituels. Nous avons développé un script qui permet, lorsque l'on met à jour un entrepôt de gestion de version de code qui contient une ontologie, de générer la documentation, les règles SHACL, la documentation des règles SHACL et qui publie tout cela derrière un serveur Web automatiquement. Le déploiement continu permet de mettre à jour le site `https://ontologie.capdataopera.fr/`.

Le script utilise différents outils pour permettre la génération de tous ces éléments. Tout d'abord, nous utilisons Widoco[5] pour la génération de la documentation à partir des métadonnées de l'ontologie et de sa modélisation. Le rendu graphique est très lisible et proche de ce qui existe déjà dans d'autres projets, ce qui en fait une documentation simple à prendre en main. Cet outil utilise WebVOWL[10] pour avoir une représentation graphique de l'ontologie. Comme on peut le voir sur la figure 1, cette représentation permet d'avoir un aperçu global de ce qui est présent dans l'ontologie, et donc elle est particulièrement utile pour la découverte de l'ontologie, mais il est difficile d'en détecter les détails.

Nous avons utilisé la génération de documentation SHACL en utilisant l'outil proposé dans SHACL-play [10]. La documentation générée (comme nous pouvons l'observer sur la figure 2) permet de se rendre compte très facilement de ce qui est attendu et avoir un rapport valide lors de la validation SHACL des données.

Tous ces outils sont utilisés dans notre processus d'intégration continue proposé dans GitLab [11] et le résultat est dé-

8. `https://www.w3.org/TR/shacl/#validation-report`
9. Voir `https://semgraph.logilab.fr`
10. `https://shacl-play.sparna.fr/play/doc`
11. `https://docs.gitlab.com/ee/ci/`


rof:Collectivite
https://ontologie.capdataculture.fr/v1/owl/#Collectivite

• Closed shape

Property name	URI	Expected value	Card.	Description
	rof:description	xsd:string	0..*	
	rof:aPourFonction	rof:Fonction	0..*	
	rof:pageWeb	xsd:anyURI	0..*	
	rof:openAgenda	owl:Thing	0..*	
	rof:nomFormeRejet	xsd:string	0..*	
	rof:catalogageSourceAgence	rof:Collectivite	0..*	
	rof:catalogageSourceDate	xsd:dateTime	0..*	
	rof:isni		0..*	
	rof:siret	xsd:string	0..*	
	rof:catalogageSourcePays	rof:LieuGeographique	0..*	
	rof:statutJuridique	rof:StatutJuridique	0..*	
	rof:nom	xsd:string	0..*	

FIGURE 2 – Documentation SHACL

ployé en utilisant les GitLab Pages [12]. Chaque fois qu'une modification dans l'ontologie est effectuée, tout le processus est automatiquement relancé grâce à l'intégration continue et le résultat est accessible grâce au serveur Web proposé dans les GitLab Pages.

Ce processus d'intégration et de déploiement continu pour la publication d'ontologie est un réel atout qui peut être utilisé dans d'autres projets dès lors qu'une ontologie doit être maintenue.

Un besoin qui a été prégnant tout le long du projet a été de pouvoir vérifier ce qui a été exporté dans l'entrepôt SPARQL. Nous avons commencé à explorer les données exportées par l'intermédiaire d'un certain nombre de requêtes SPARQL pour voir le résultat. Cette solution a vite montré ses limites, car il n'a pas été simple d'écrire les requêtes SPARQL permettant de tout voir facilement et rapidement. Nous avons alors utilisé l'outil SparqlExplorer [13] qui permet de parcourir l'ensemble des données d'un entrepôt SPARQL pour découvrir les données qui y sont présentes.

La figure 3 présente la page d'accueil du SparqlExplorer une fois que l'on a spécifié l'entrepôt SPARQL à explorer. Nous pouvons observer la liste des classes et le nombre d'instances associés à chacune des classes, et un champ de recherche, qui permet de chercher parmi les littéraux.

Lorsque nous cliquons sur une URI, nous affichons la vue présentée sur la figure 4. Cette vue permet de lister l'ensemble des triplets concernant cette URI et de pouvoir filtrer ces triplets (ici un filtre a été appliqué avec la valeur "nom"). De cette manière, il est possible de parcourir les triplets pour observer ce qui a vraiment été exporté et donc de s'assurer que le résultat correspond bien à ce qui est attendu.

Nous avons intégré l'outil YASGUI[13] pour interroger l'entrepôt SPARQL grâce à une interface plus pratique à utiliser que l'interface proposée par Virtuoso. Cette interface, visible sur la figure 5, comporte une option pour partager un lien vers une requête SPARQL. Ce lien a beau-

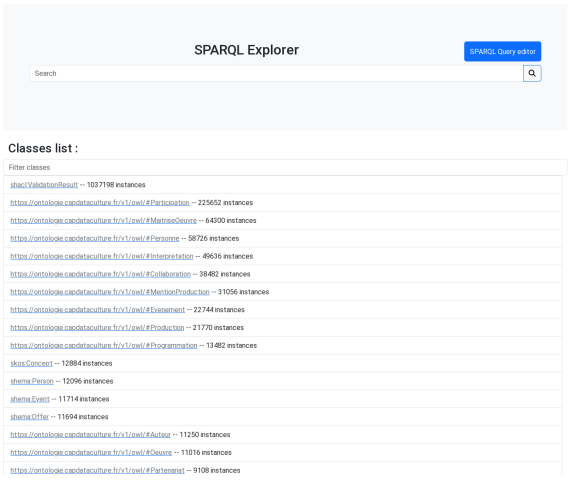
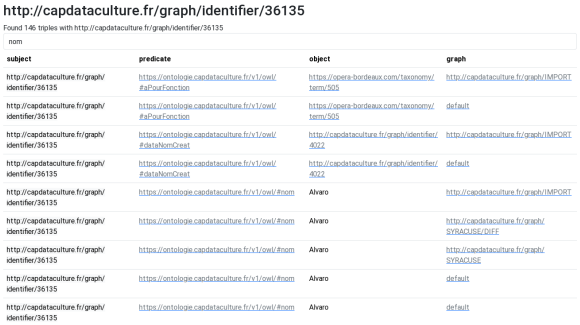


FIGURE 3 – Page d'accueil du SparqlExplorer


http://capdataculture.fr/graph/identifier/36135

Found 146 triples with http://capdataculture.fr/graph/identifier/36135

nom

subject	predicate	object	graph
http://capdataculture.fr/graph/identifier/36135	https://ontologie.capdataculture.fr/v1/owl/#aPourFonction	https://opera-bordeaux.com/taxonomy/term/505	http://capdataculture.fr/graph/IMPORT
http://capdataculture.fr/graph/identifier/36135	https://ontologie.capdataculture.fr/v1/owl/#aPourFonction	https://opera-bordeaux.com/taxonomy/term/505	default
http://capdataculture.fr/graph/identifier/36135	https://ontologie.capdataculture.fr/v1/owl/#dataNomCreat	http://capdataculture.fr/graph/identifier/4022	http://capdataculture.fr/graph/IMPORT
http://capdataculture.fr/graph/identifier/36135	https://ontologie.capdataculture.fr/v1/owl/#dataNomCreat	http://capdataculture.fr/graph/identifier/4022	default
http://capdataculture.fr/graph/identifier/36135	https://ontologie.capdataculture.fr/v1/owl/#nom	Alvaro	http://capdataculture.fr/graph/IMPORT
http://capdataculture.fr/graph/identifier/36135	https://ontologie.capdataculture.fr/v1/owl/#nom	Alvaro	http://capdataculture.fr/graph/SYRACUSE/DIFF
http://capdataculture.fr/graph/identifier/36135	https://ontologie.capdataculture.fr/v1/owl/#nom	Alvaro	http://capdataculture.fr/graph/SYRACUSE
http://capdataculture.fr/graph/identifier/36135	https://ontologie.capdataculture.fr/v1/owl/#nom	Alvaro	default
http://capdataculture.fr/graph/identifier/36135	https://ontologie.capdataculture.fr/v1/owl/#nom	Alvaro	default

FIGURE 4 – Liste de triplets dans le SparqlExplorer

12. https://docs.gitlab.com/ee/user/project/pages/
13. https://sparqlexplorer.app/

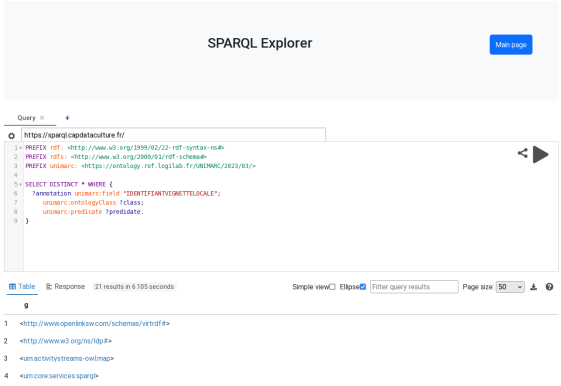


FIGURE 5 – YASGUI dans le SparqlExplorer

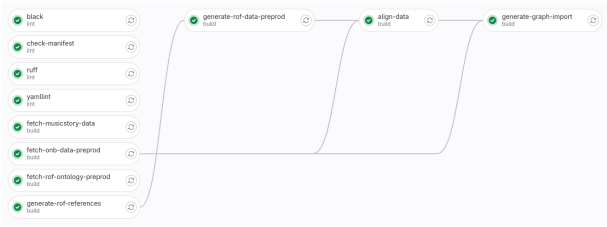


FIGURE 6 – Exécution des différentes étapes de récupération des données

coup été utilisé durant le projet et a grandement simplifié les échanges et les collaborations pour communiquer entre partenaires la présence au l'absence de certaines informations dans le graphe.

4.2 Assemblage et publication du graphe

Une fois que chaque maison d'opéra a publié ses données à une adresse URL de son choix, il faut les récupérer pour les agréger dans le même entrepôt SPARQL. Cette récupération s'effectue quotidiennement pour s'assurer d'avoir des données à jour dans l'entrepôt. Pour cela, nous avons mis en place une tâche récurrente avec le mécanisme d'intégration continue de notre forge logicielle [14]. Nous pouvons ensuite suivre l'exécution de chaque tâche et regarder le résultat obtenu à chaque étape.

La figure 6 permet de voir les différentes étapes d'une mise à jour de l'intégralité des données. Sur cet exemple, seules trois sources de données sont présentes : Les données de l'Opéra National de Bordeaux, les données de Music Story [15] et les données de la Réunion des Opéras de France (ROF). Cette dernière source requiert plusieurs étapes car nous interrogeons directement une API pour récupérer les données que nous transformons ensuite en RDF. Nous sommes en train d'étudier le transfert de cette responsabilité vers l'équipe qui gère les données de la ROF.

La solution trouvée ici permet de mettre en lumière l'importance de l'adoption de la solution. Dès qu'un fournisseur de données n'adopte pas les technologies préconisées

14. GitLab `https://docs.gitlab.com/ee/ci/`

15. Pour la valorisation notamment sur les plateformes de streaming `https://music-story.com/fr/`

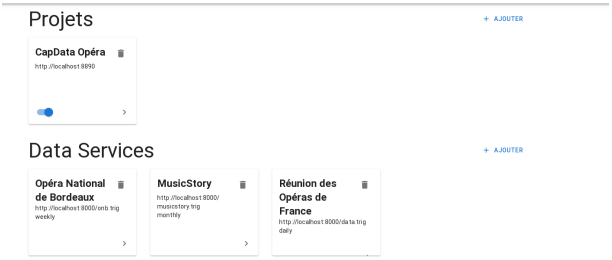


FIGURE 7 – Exemple de l'application de suivi de production

par le projet, cela demande un traitement dédié et spécifique pour cette source. Par exemple, ici, nous avions accès à une API, que nous avons utilisée pour générer le graphe RDF contenant l'intégralité des données. Ce traitement n'a pas été simple à mettre en place à cause de la complexité de l'API [16]. Ce travail a nécessité beaucoup d'échanges avec l'équipe en charge de la gestion des données. Il est apparu que les développements sont beaucoup plus fluides lorsque chaque gestionnaire de données gère la transformation en RDF de ses propres données.

Afin de faciliter les développements et de permettre aux fournisseurs de données d'être les plus autonomes possible, nous avons initié le développement d'une application de suivi de production qui permettra de suivre l'état de chaque récolte spécifiquement et d'avoir accès aux journaux d'import pour savoir comment il s'est déroulé. L'application enverra les données dans l'entrepôt SPARQL si la récolte s'est correctement déroulée. De cette manière, les maisons d'opéra seront autonomes dans la publication de leurs données. Elles pourront ajouter elle-même la source de données dans l'application de suivi de production et corriger les erreurs qui seront remontées dans les journaux suite à la vérification de la conformité des données en utilisant les règles SHACL. Cette application donnera une vision claire de ce qui a été importé dans l'entrepôt SPARQL. Elle constitue une étape importante dans la phase d'industrialisation du projet CapData Opéra.

Comme nous pouvons le voir dans la figure7, nous pouvons définir un projet, ici "CapData Opéra" et différentes sources ("Opéra National de Bordeaux", "Music Story" et "Réunion des Opéras de France").

La figure 8 montre un exemple de l'ajout d'une recette dans l'application de suivi de production. Cette recette permet d'identifier une source de données à importer pour le projet, l'URI du graphe dans lequel nous souhaitons envoyer les données et le processus à appliquer sur les données. Ce processus pourra être modifié dans le code pour permettre des traitements particuliers, comme transformer du CSV, ou utiliser une API, etc. Des erreurs d'import pour une source ne bloqueront pas l'import des autres sources, ce qui permet une plus grande flexibilité. Il sera alors possible d'intégrer

16. Basée sur une modélisation UNIMARC[2]

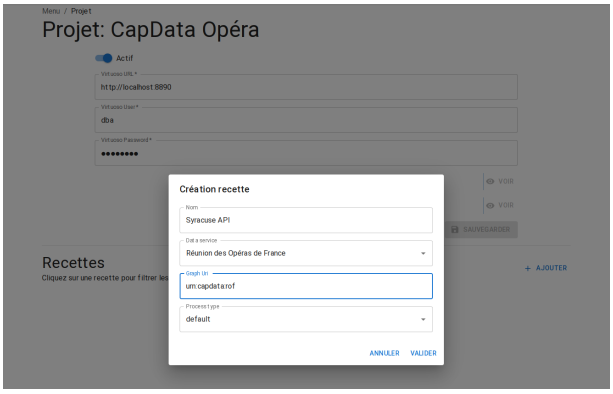


FIGURE 8 – Ajout d'une recette dans l'application de suivi de production

un plus grand nombre de maisons d'opéras plus facilement.

Le graphe final est publié dans un entrepôt RDF interrogeable en SPARQL grâce au logiciel Virtuoso [17] et parcouru avec un navigateur web en utilisant SparqlExplorer, comme présenté dans le chapitre précédent.

5 Responsabilités et périmètres

Nous avons présenté le processus de publication des données des différentes maisons d'opéra pour permettre l'interopérabilité entre ces données. L'architecture et les outils présentés facilitent l'autonomisation du processus, tout en assurant la pleine souveraineté des maisons d'opéra et structures participantes sur leurs données. La mise en place d'une solution mutualisée et la coordination des développements des outils développés dans le cadre du projet permet de répartir les responsabilités.

Tout d'abord dans le processus de publication des données, la mission numérique de la ROF en lien avec les besoins détectés auprès des groupes de travail dont celui des référents numériques et des échanges et des réflexions avec les partenaires, assure la maintenance de l'ontologie. Cette maintenance nécessite de considérer les besoins en modélisation des différentes maisons d'opéra, de les transcrire dans l'ontologie et de faire un suivi de versions de l'ontologie pour communiquer autour des changements. C'est pour cela que nous avons adopté pour la gestion rapide de l'ontologie un déploiement continu et documenté. De plus, une gestion de version de l'ontologie est intégrée directement dans l'URI de celle-ci, ce qui permet de pouvoir communiquer chaque fois qu'une modification importante dans l'ontologie a lieu en annonçant la publication d'une nouvelle version.

Un certain nombre de référentiels sont publiés par la ROF directement dans l'entrepôt SPARQL du projet en utilisant le vocabulaire SKOS (comme la liste des fonctions lors d'une participation d'une personne à une production). Il est important de pouvoir proposer des référentiels uniques pour tout le projet afin d'assurer une interopérabilité importante entre les données des différentes maisons d'opéra.

17. https://virtuoso.openlinksw.com/

Enfin, il est également nécessaire de proposer une infrastructure permettant d'héberger les outils proposés, dont principalement l'ontologie, l'application de suivi de production et l'entrepôt SPARQL. La coordination menée par le réseau ROF permet d'assurer une continuité et la stabilité des services tout en adoptant une approche mutualisée.

Au-delà de l'aspect technique, l'accompagnement et la constitution d'un socle commun de connaissances auprès des maisons participantes et des prestataires se sont révélés être une clé majeure dans la conduite et la réussite du projet. Les simples publications de l'ontologie et de l'entrepôt SPARQL seraient complètement insuffisantes pour assurer leur intégration et déploiement.

La constitution d'une documentation adaptée de l'ontologie et la communication s'avère stratégique pour le déploiement de la solution. Plusieurs présentations publiques ont eu lieu au cours de l'année 2023. Dans ce sens, nous avons souhaité développer une application de suivi de production la plus simple possible. Nous organisons aussi des réunions, ateliers et sessions de travail pour faciliter au maximum l'appropriation des concepts et des technologies que nous avons mis en œuvre ici. Les technologies du Web Sémantique restent méconnues par de nombreux prestataires du domaine culturel, la mise en place de salons de messagerie instantanée favorisant les échanges et questions, se révèlent être des leviers particulièrement puissants. Près de 5 salons dédiés à la gestion et coordination des chantiers ont été mis en place, générant plus de 10 000 interactions sur l'année 2023.

Tout en favorisant la mutualisation, le choix d'une architecture permet une grande agilité, assure le respect de la souveraineté et la pleine responsabilité des maisons, partenaires et services externes participant au projet. Ils conservent une complète autonomie dans le choix d'ouverture, de diffusion, de réutilisation et de l'enrichissement de leurs données. L'Opéra National de Bordeaux a ainsi la possibilité d'exposer ses données de programmation tout en enrichissant sa base avec des données identifiées et exposées au sein de l'entrepôt par un ou plusieurs membres des structures participantes.

6 Conclusion

Élaboré sur la base des standards du web sémantique, le projet CapData Opéra a déployé une solution mutualisée fondée sur une architecture et des outils hautement réplicables. L'expérimentation et les premiers résultats liés confirment la pertinence de l'approche choisie en vue de simplifier l'échange, la gestion et la découvrabilité des données des maisons d'opéra et des autres structures participantes.

En complète adéquation avec les politiques culturelles, l'interopérabilité des données répond à de multiples besoins détectés auprès des maisons d'opéra et enjeux du spectacle vivant, dont le développement de la découvrabilité des œuvres, des artistes et plus largement des arts lyriques et chorégraphiques auprès des publics. D'autres acteurs culturels nous semblent en effet être sur la même ligne que la

nôtre. Nous pouvons citer le monde des marionnettistes avec qui nous sommes en contact, le monde du théâtre qui porte des initiatives comme la publication des données des Registres de la Comédie Française[6], le ministère de la Culture qui conduit des ateliers dans le cadre de la deuxième génération de la feuille de route "Politique données et contenus culturels"[14] ou encore les réflexions du groupe de travail "Ouverture des données"[3] animé par le réseau du TMNlab.

La modélisation d'une ontologie et son adoption au sein de systèmes d'information hétérogènes sont des actions complexes. L'expérimentation souligne les rôles essentiels de la coordination et de l'accompagnement des établissements partenaires et prestataires. Si la question de l'héritage des logiciels et des processus humains existants est aujourd'hui bien connue, elle nécessite une attention toute particulière pour l'intégration de nouveaux modèles.

Le projet a été l'occasion d'éprouver un certain nombre d'outils du Web Sémantique et d'identifier les fonctionnalités manquantes ou les besoins pour lesquels les outils restent à concevoir. Nous avons mis en place la génération de la documentation de l'ontologie et son déploiement continu par l'intermédiaire d'un entrepôt de code, une application de suivi de production et un outil de navigation dans le graphe final. Nous prévoyons d'améliorer ces outils, mais surtout de les rendre plus génériques pour qu'ils puissent être utilisés dans d'autres projets.

L'expérimentation menée dans un premier temps avec l'Opéra National de Bordeaux est en cours d'industrialisation et de déploiement auprès de six maisons d'opéra dont l'Opéra National de Bordeaux, Théâtre du Châtelet, l'Opéra de Rennes, l'Opéra Comique, l'Opéra national Capitole Toulouse et l'Opéra de Limoges. Cette approche permet d'affiner progressivement les différents chantiers et outils : ontologie, connecteurs, applications, documentations et services dédiés à la valorisation.

Nous avons commencé à étudier la possibilité de valoriser ces données aux travers de services de valorisation dédiés, par exemple via un prestataire permettant de faire le lien avec les plateformes de streaming, ou encore un système de gestion d'agenda partagé pour la publication automatique des événements.

L'expérimentation, le constat d'un besoin présent partagé par un grand nombre d'établissements du spectacle vivant et plus largement du secteur culturel et l'émergence d'initiatives similaires à l'internationale, soulignent le besoin et la pertinence d'une démarche coordonnée des recherches et actions.

L'adoption d'une approche coordonnée met en lumière le chantier essentiel de la gouvernance. Les modèles de fonctionnement internationaux de l'IFLA ou du schema.org sont ainsi précieux d'enseignement. Les opportunités offertes par une telle approche sont nombreuses, tant pour la mutualisation des coûts financiers, le développement d'outils partagés en capacité de simplifier et d'assurer de manière pérenne leurs adoptions et déploiement, tout en répondant aux besoins et enjeux transversaux du secteur culturel.

Références

[1] Adrien Basdevant, Camille François, and Rémi Ronfard. *Rapport de la mission sur le développement des métavers*. PhD thesis, Ministère de la Culture (France), 2022.

[2] Permanent UNIMARC Committee et al. Unimarc authorities format manual. 2023.

[3] Groupe de travail TMNlab. Living lab - Ouverture des données, 2023.

[4] Mission franco-québécoise sur la découvrabilité en ligne des contenus culturels francophones. Table ronde "Normaliser la diversité des données culturelles : est-ce possible ? Rendez-vous France-Québec", Montréal Connecte. `https://www.youtube.com/watch?v=3HbgAUUNUiw`, 2023.

[5] Daniel Garijo. Widoco : a wizard for documenting ontologies. In *The Semantic Web–ISWC 2017 : 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part II 16*, pages 94–102. Springer, 2017.

[6] Charline Granger and Fabien Amarger. Les registres de la comédie-française sur le web de données liées : de l'hétérogénéité de données vers des données quantitatives en rdf. In *Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle APIA@ PFIA2023*, number 2023, pages 63–71. AFIA-Association Française pour l'Intelligence Artificielle, 2023.

[7] Direction générale des médias et des industries culturelles (DGMIC). Découvrabilité en ligne des contenus culturels francophones, 2022.

[8] Clement Jonquet, John Graybeal, Syphax Bouazzouni, Michael Dorf, Nicola Fiore, Xeni Kechagioglou, Timothy Redmond, Ilaria Rosati, Alex Skrenchuk, Jennifer L Vendetti, et al. Ontology repositories and semantic artefact catalogues with the ontoportal technology. In *International Semantic Web Conference*, pages 38–58. Springer, 2023.

[9] Julie Knibbe. Les données dans la musique : Enjeux et stratégies d'investissement. 2023.

[10] Steffen Lohmann, Vincent Link, Eduard Marbach, and Stefan Negru. Webvowl : Web-based visualization of ontologies. In *Knowledge Engineering and Knowledge Management : EKAW 2014 Satellite Events, VISUAL, EKM1, and ARCOE-Logic, Linköping, Sweden, November 24-28, 2014. Revised Selected Papers. 19*, pages 154–158. Springer, 2015.

[11] Frank Manola, Eric Miller, Brian McBride, et al. Rdf primer. *W3C recommendation*, 10(1-107) :6, 2004.

[12] Eudes-Emmanuel Peyre and Groupe de travail numérique ROF. Capdata Opéra - France 2030. `https://www.rof.fr/rof/capdata-opera.aspx`, 2022.

[13] Laurens Rietveld and Rinke Hoekstra. The yasgui family of sparql clients 1. *Semantic Web*, 8(3) :373–383, 2017.

[14] Ministère de la Culture (SNUM) Service du numérique. Ateliers préliminaires à la deuxième génération de la feuille de route "Politique des données et contenus culturels", 2024.

[15] Maja Žumer. Ifla library reference model (ifla lrm)—harmonisation of the frbr family. *KO Knowledge Organization*, 45(4) :310–318, 2018.

Vers une approche floue pour le design de plan expérimental

Olivier Rousselle, Jean-Philippe Poli, Nadia Ben Abdallah

Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France

{prenom.nom}@cea.fr

Résumé

Nous présentons dans cet article une approche floue interprétable du design expérimental sous contraintes qui peut être utilisée avec peu de données. L'objectif est de fournir aux expérimentateurs un algorithme leur permettant d'échantillonner de manière optimale. Nous détaillons les différentes étapes de notre algorithme qui consiste à recommander la prochaine expérience à réaliser et à construire une base de règles floues de Sugeno. Nous présentons ensuite quelques résultats de notre algorithme, que nous comparons avec l'approche bayésienne.

Mots-clés

Apprentissage actif, design expérimental, flou, règles, optimisation, interprétabilité.

Abstract

In this paper we present an interpretable fuzzy approach to constrained experimental design that can be used with little data. The objective is to provide experimenters with an algorithm allowing them to sample optimally. We detail the different steps of our algorithm which consists of recommending the next experiment to be carried out and constructing a Sugeno fuzzy rule base. We will then present some results of our algorithm, which we compare with the Bayesian approach.

Keywords

Active learning, experimental design, fuzzy, rules, optimization, interpretability.

1 Introduction

Les sciences expérimentales nécessitent d'explorer un espace de possibilités souvent très vaste pour s'approcher d'un optimum. Classiquement, l'approche par essais et erreurs est utilisée dans ce domaine pour collecter des données expérimentales, dont la production peut être coûteuse et longue. Le processus est répété jusqu'à ce qu'une propriété ou une performance désirée soit atteinte.

Différentes méthodes d'échantillonnage sont utilisées dans la recherche expérimentale, telles que l'échantillonnage aléatoire, l'échantillonnage factoriel, la méthode des surfaces de réponse, l'optimisation bayésienne [1], l'algorithme de couverture optimale [2], etc. Les réseaux de neurones ont également été utilisés pour la recherche expérimentale [3], mais ont l'inconvénient d'être une boîte noire.

Sans perte de généralité, nous prenons comme exemple la découverte des matériaux, qui vise à produire des matériaux performants pour un usage ciblé. Ces matériaux sont généralement produits à partir d'un mélange de composés initiaux soumis à un certain procédé de fabrication. Ces dernières années, l'intelligence artificielle a accéléré l'innovation dans ces domaines [4, 5, 6].

Dans ce contexte, l'objectif de nos travaux est de développer une méthode basée sur la logique floue appliquée au plan expérimental. Nous définissons notre problème comme tester différents ensembles de paramètres (c'est-à-dire la composition d'un matériau et les paramètres du procédé) pour maximiser une propriété donnée (par exemple la robustesse de ce matériau). En particulier, nous souhaitons trouver une méthode automatique pour échantillonner de manière itérative et optimale les paramètres expérimentaux.

Notre motivation est de fournir un outil pour aider les expérimentateurs à déterminer quels sont les prochains ensembles de paramètres à tester et qui fonctionne avec peu de données. Nous visons à réduire le nombre d'expérimentations pour atteindre une performance cible, à la fois pour réduire le gaspillage de matières premières et converger plus rapidement vers un matériau innovant. Pour garder l'expert humain dans la boucle, nous prêtons attention à l'interprétabilité, en donnant des indices à l'utilisateur sur le choix de la prochaine configuration expérimentale. En effet, l'explicabilité/interprétabilité vise à rendre le fonctionnement et les résultats du modèle plus intelligibles et transparents pour les humains, afin de renforcer la confiance dans la prise de décision et ainsi son acceptabilité.

Le document est structuré comme suit. La section suivante donne un aperçu de l'approche. La section 3 décrit la méthode de régression qui se rapproche de la fonction objectif. La section 4 explique le processus derrière la sélection de la prochaine expérience à réaliser. Nous montrons les résultats et la comparaison avec l'optimisation bayésienne dans la Section 6. Notre approche étant dédiée aux experts humains, la Section 7 présente la manière dont l'utilisateur final est considéré. Enfin, nous tirons quelques conclusions et perspectives.

2 Vue d'ensemble de l'approche

Pour satisfaire les besoins des expérimentateurs, nous avons conçu une approche basée les principes suivants :

— Elle doit mettre en œuvre un échantillonnage adap-

tatif [7], c'est-à-dire un processus séquentiel qui décide de l'emplacement du point suivant en équilibrant les critères d'exploration et d'exploitation ;
— Elle doit être capable de combiner apprentissage (à partir de quelques données expérimentales) et modélisation de connaissances expertes ;
— Robustesse : de petits changements dans les points initiaux ne doivent pas entraîner de changements importants dans les résultats et les prédictions ;
— Interprétabilité : les étapes et les résultats du modèle doivent être intelligibles pour les humains, pour renforcer la confiance dans la prise de décision ;
— Notre approche doit pouvoir travailler sur des problèmes de mélanges de grande dimension.

Pour répondre à ces prérequis, nous avons développé une approche dont les différentes étapes sont détaillées dans la Fig. 1.

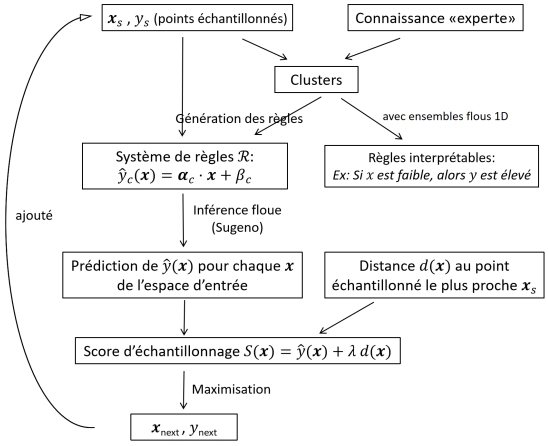


FIGURE 1 – Différentes étapes de l'approche proposée.

Le point expérimental sélectionné est celui qui maximise le score d'échantillonnage (section 4). Ce processus est répété sur plusieurs itérations, chaque itération correspondant à une expérience.

Dans ce travail, nous considérons que les propriétés des matériaux sont données sous forme de valeurs réelles (souvent bornées). Nous devons donc construire une base de règles floues pour une régression (voir section 3). Nous avons choisi d'utiliser un système d'inférence flou de type Sugeno pour son efficacité et pour le fait qu'il fonctionne avec les techniques d'optimisation adaptatives.

Avec ce choix, nous avons privilégié les performances de la prédiction plutôt que l'interprétabilité. Pour compenser, nous proposons une méthode pour extraire un substitut plus interprétable du modèle (section 5).

Remarque : les notations utilisées dans cet article sont détaillées en annexe.

3 Algorithme de régression basé sur le clustering flou

Considérons le problème de la prédiction de la propriété d'un matériau, $\hat{y}$, pour chaque point de l'espace d'entrée (c'est-à-dire tout mélange de matières premières et de paramètres de processus). Nous avons basé notre approche sur plusieurs travaux antérieurs [8, 9, 10] qui partagent l'utilisation d'une méthode de clustering comme première étape dans l'induction de règles. Notre méthode diffère légèrement dans le sens où elle est destinée à des problèmes éventuellement de grande dimension. De plus, nous avons amélioré la manière dont les fonctions d'appartenance sont apprises et le calcul des coefficients de régression pour les conclusions de la règle de Sugeno.

3.1 Clustering

Tout d'abord, nous effectuons un clustering flou multidimensionnel des points déjà échantillonnés $\boldsymbol{x}_s$ pour mettre en évidence différents groupes. Rappelons que la particularité du clustering flou est qu'un point peut appartenir à plusieurs clusters, avec éventuellement des degrés d'appartenance différents.

Le clustering flou s'applique à la fois aux variables d'entrée et de sortie, et chaque cluster c comprend un centre noté $\boldsymbol{m}_c$. Nous notons C l'ensemble des clusters. Le nombre de clusters n_c est souvent un hyperparamètre, dont la valeur doit être définie en respectant les considérations suivantes :

— Trop de clusters peuvent conduire à un surapprentissage. Le modèle s'est bien adapté aux données d'entraînement (points déjà échantillonnés), mais il peut avoir du mal à se généraliser à de nouvelles données. De plus, avoir trop de clusters implique trop de paramètres, et l'optimisation décrite plus loin dans cet article ne sera pas possible.
— Un faible nombre de clusters peut faciliter l'interprétation du modèle et réduire la complexité des calculs (moins de paramètres à optimiser).

Le nombre de clusters peut rester constant au cours des différentes itérations de l'expérience ou peut augmenter régulièrement par paliers en fonction du nombre de points déjà échantillonnés.

Nous nous intéressons maintenant à mesurer le degré d'appartenance d'un point donné $\boldsymbol{x}$ dans l'espace d'entrée à chaque cluster, noté $\mu_c(\boldsymbol{x})$. Nous avons choisi de construire une fonction d'appartenance qui dépend de plusieurs variables d'entrée. L'avantage est de prendre en compte l'interaction entre les variables et d'obtenir une partition forte multidimensionnelle :

$$\forall \boldsymbol{x}\ ,\ \sum_{c \in C} \mu_c(\boldsymbol{x}) = 1. \tag{1}$$

Les degrés d'appartenance peuvent être calculés en s'inspirant de l'algorithme FCM (Fuzzy Clustering Means) [11], c'est-à-dire en minimisant la fonction objectif [12] :

$$\sum_{\boldsymbol{x}_s} \sum_{\boldsymbol{m}_c} \mu_c(\boldsymbol{x}_s) ||\boldsymbol{x}_s - \boldsymbol{m}_c||^2 \tag{2}$$

avec $\boldsymbol{x}_s$ les points considérés, m_c les centres de chaque cluster, et $\mu_c(\boldsymbol{x}_s)$ le coefficient d'appartenance du point $\boldsymbol{x}_s$ au cluster c. Les degrés d'appartenance obtenus sont :

$$\mu_c(\boldsymbol{x}_s) = \sum_{c'} \left(\frac{||\boldsymbol{x}_s - \boldsymbol{m}_c||^2}{||\boldsymbol{x}_s - \boldsymbol{m}_{c'}||^2} \right)^{-\frac{2}{m-1}} \tag{3}$$

avec m le paramètre "fuzzifier" qui influence le flou de la partition. Il va de 1 (partition nette) à $+\infty$. Généralement, m est choisi égal à 2. Chaque point $\boldsymbol{x}$ se voit ensuite attribuer un degré d'appartenance à chaque cluster. Des exemples de degrés d'appartenance sont illustrés dans le diagramme ternaire sur la Fig. 2 avec 3 clusters et avec 3 variables d'entrée x_1, x_2, x_3. En guise de lecture du diagramme ternaire, la croix rouge indique par exemple le point $(x_1, x_2, x_3) = (0.55, 0.2, 0.25)$.

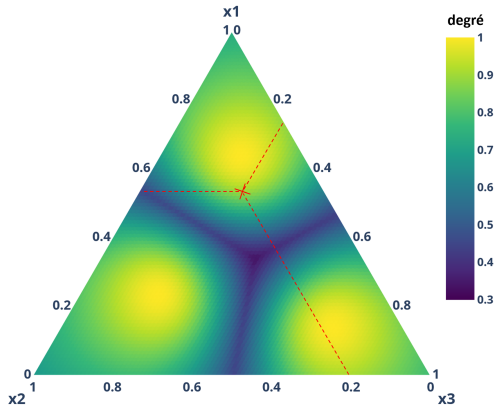


FIGURE 2 – Degrés d'appartenance - étude de cas avec 3 variables de composition et 3 clusters.

La méthode ne vise pas à caractériser les clusters, mais à les décrire dans leur ensemble. La distribution associée correspond donc à des degrés d'appartenance qui indiquent dans quelle mesure un point appartient à chaque cluster [13].
Remarque : on peut également construire une fonction d'appartenance qui dépend uniquement de la distance au centre du cluster. Il peut prendre une forme triangulaire, gaussienne [8, 14] ou Cauchy [10]. Cependant, nous n'avons pas choisi une telle fonction d'appartenance car elle ne considère que les similarités internes au sein d'un cluster (et non les dissemblances externes avec d'autres clusters), et parce que la partition floue n'est pas forte (la somme des degrés d'appartenance n'est pas égale à 1).

3.2 Génération de règles floues

Chaque cluster/région c est à l'origine d'une règle R. Chaque règle R est caractérisée par sa région antécédente multidimensionnelle et ses coefficients de régression $(\boldsymbol{\alpha}_c, \beta_c)$, sous la forme :

$$\hat{y}_c(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_{c_i} x_i + \beta_c = \boldsymbol{\alpha}_c \cdot \boldsymbol{x} + \beta_c. \tag{4}$$

Les coefficients $(\boldsymbol{\alpha}_c, \beta_c)$ de chaque règle sont déterminés par le processus d'optimisation décrit plus loin dans cet article.

Règle experte Une règle peut également provenir d'une expertise humaine ou être issue de la littérature. Par exemple, un expert peut indiquer une autre région d'intérêt en ajoutant un autre centre $\boldsymbol{m}_c$. Ce cluster générera également une règle caractérisée par sa région antécédente et ses coefficients de régression $(\boldsymbol{\alpha}_c, \beta_c)$. Cette règle sera ensuite ajoutée au système de règles déjà généré à partir des données.

Cas des variables discrètes Dans le cas de variables discrètes ou catégorielles, il est nécessaire de transformer ces données pour pouvoir intégrer ces variables dans la régression. Pour ce faire, nous utilisons l'encodage One-Hot, qui consiste à transformer la variable en plusieurs variables binaires, où chaque variable binaire représente une catégorie unique de la variable d'origine.

Processus récursif pour le partitionnement flou A chaque nouveau point échantillonné, nous évaluons à quel cluster il appartient, c'est-à-dire le cluster ayant le plus haut degré d'appartenance. Le centre $\boldsymbol{m}_c$ du cluster c est ensuite modifié en calculant la moyenne pondérée suivante (en notant $\boldsymbol{x}_{last}$ le dernier point testé) [15] :

$$\boldsymbol{m}'_c = \frac{\boldsymbol{m}_c + \mu_c(\boldsymbol{x}_{last})\, \boldsymbol{x}_{last}}{\mu_c(\boldsymbol{x}_{last})}. \tag{5}$$

3.3 Optimisation des coefficients de régression

Le résultat de notre modèle est généré en combinant les fonctions linéaires des règles, comme dans les systèmes de Sugeno :

$$\hat{y}(\boldsymbol{x}) = \sum_c \mu_c(\boldsymbol{x})\, \hat{y}_c(\boldsymbol{x}) \tag{6}$$

avec $\hat{y}_c(\boldsymbol{x}) = \boldsymbol{\alpha}_c \cdot \boldsymbol{x} + \beta_c$.
Nous déterminons les coefficients de régression optimaux $(\boldsymbol{\alpha}_c, \beta_c)$ en minimisant l'écart au carré entre les valeurs prédites des points déjà échantillonnés $\hat{y}$ et leurs valeurs réelles y . Chaque cluster contient $N + 1$ coefficients à optimiser. Au total, il y a donc $n_c(N + 1)$ paramètres à optimiser, il nous faut donc au moins $n_c(N + 1)$ points pour réaliser cette optimisation.
Pour chaque point déjà échantillonné, nous prédisons la valeur de sortie à l'aide de la formule d'inférence, que nous comparons à la valeur de sortie réelle. L'objectif est de minimiser la différence entre les valeurs prédites $\hat{y}$ et les valeurs réelles y. On cherche donc à minimiser la fonction de perte :

$$\begin{aligned} L(\boldsymbol{\alpha}, \beta) &= \sum_{\boldsymbol{x}_s} \left(\hat{y}(\boldsymbol{x}_s) - y(\boldsymbol{x}_s)\right)^2 \\ &= \sum_{\boldsymbol{x}_s} \left(\sum_c \mu_c(\boldsymbol{x}_s)\, (\boldsymbol{\alpha}_c \cdot \boldsymbol{x}_s + \beta_c) - y(\boldsymbol{x}_s) \right)^2. \end{aligned} \tag{7}$$

3.4 Prédiction de la valeur de sortie pour chaque point d'entrée

Pour chaque point d'entrée $\boldsymbol{x}$, nous prédisons la valeur de sortie $\hat{y}$. La figure 3 montre un exemple de valeurs prédites $\hat{y}$ obtenues avec notre algorithme pour un cas avec 3 variables d'entrée.
Nous pouvons évaluer la précision du modèle d'inférence en utilisant les points déjà échantillonnés. Pour chaque

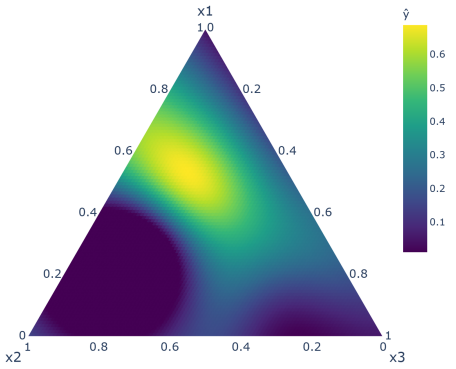


FIGURE 3 – Diagramme ternaire illustrant un exemple de valeurs prédites $\hat{y}$ pour un cas avec 3 variables d'entrée.

point échantillonné, nous prédisons la valeur de sortie $\hat{y}$ à l'aide du système d'inférence, que nous comparons à la valeur réelle y. On peut alors calculer l'écart au carré moyen RMSE et le coefficient de détermination R^2 pour juger de la qualité du modèle. L'objectif est d'avoir un modèle qui maximise R^2 et minimise le RMSE. Les résultats sont présentés dans la partie 6.
Dans [10], les auteurs ont montré que l'algorithme de régression basé sur le clustering flou avec descente de gradient est plus efficace que les réseaux de neurones pour un problème de régression (prédiction de la forme d'une fonction) pour le cas 1D, et avec beaucoup moins d'hyperparamètres.

4 Sélection de la prochaine expérience

La méthode proposée choisit également le prochain point à échantillonner de manière déterministe, comme un compromis entre exploitation et exploration. L'exploitation signifie privilégier les régions à fort potentiel, c'est-à-dire avec des valeurs de sortie élevées prédites $\hat{y}$, tandis que l'exploration signifie explorer les régions encore non testées.
Nous devons introduire une variable qui capture cette partie exploration. Une variable naturelle pour cela est la distance euclidienne au point échantillonné le plus proche, notée d. Le calcul des distances entre chaque point de l'espace d'entrée et le point déjà échantillonné le plus proche est effectué de manière optimisée à l'aide de l'algorithme KD-tree [16]. Ensuite, pour chaque point d'entrée $\boldsymbol{x}$, $\hat{y}$ et d sont calculés ; $\hat{y}$ code l'exploitation, et d code l'exploration. Pour trouver un compromis entre exploitation et exploration, nous introduisons une nouvelle variable appelée "score d'échantillonnage" et notée S :

$$S(\boldsymbol{x}) = \hat{y}(\boldsymbol{x}) + \lambda\, d(\boldsymbol{x}) \tag{8}$$

où $\hat{y}$ et d sont ici normalisés, et où λ est un hyperparamètre. λ peut être choisi constant ou il peut changer en fonction du nombre d'expériences réalisées. Augmenter λ favorisera l'exploration par rapport à l'exploitation.

Une illustration de S est présentée dans la Fig. 4 pour un exemple avec 3 variables d'entrée. Les zones en bleu sont celles autour des points déjà échantillonnés, et les zones en jaune sont les régions d'intérêt. Une structure en réseau peut être observée, due au compromis entre exploitation et exploration. Ainsi, par exemple, pour un problème à trois variables x_1, x_2, x_3, l'algorithme utilise le critère S pour calculer une région de la forme :
Le prochain point proposé par l'algorithme sera celui qui maximise $S(\boldsymbol{x})$. Par exemple, le prochain point à tester pourrait être

$$\{x_1 = 0.35\ ,\ x_2 = 0.5\ ,\ x_3 = 0.15\}\,. \tag{9}$$

Alternativement, l'algorithme peut également proposer une région d'intérêt à l'expérimentateur. Cette région comprend tous les points ayant un score d'échantillonnage S supérieur à un seuil donné (par exemple 0.9). Par exemple, une région d'intérêt pourrait être :

$$\begin{cases} 0.3 \leq x_1 \leq 0.425 \\ 0.46 \leq x_2 \leq 0.535 \\ 0.11 \leq x_3 \leq 0.21. \end{cases} \tag{10}$$

puis l'expérimentateur choisira un point dans cette région.
Une explication peut également être donnée pour justifier le point/région proposé. Par exemple : "l'algorithme propose ce point/région à explorer à côté d'une zone déjà exploitée et qui a donné de bons résultats."
La proportion exploitation/exploration du prochain point testé peut également être fournie à l'expérimentateur :

$$p_{\text{exploitation}} = \frac{\hat{y}(\boldsymbol{x}_{next})}{\hat{y}(\boldsymbol{x}_{next}) + \lambda\, d(\boldsymbol{x}_{next})} \tag{11}$$

$$p_{\text{exploration}} = 1 - p_{\text{exploitation}}. \tag{12}$$

Enfin, des contraintes peuvent être prises en compte pour restreindre l'espace d'entrée ; par exemple $x_2 < 0.5$. Expérimentalement, ces contraintes pourraient être imposées par le dispositif expérimental, par exemple du fait des limites de la machine expérimentale ou du fait de lois physico-chimiques (comme la loi de miscibilité) dans le cas d'un mélange de matériaux.

5 Interprétabilité

Les avantages d'avoir des règles multidimensionnelles par rapport aux règles basées sur des ensembles flous 1D sont la possibilité de contrôler le nombre de règles souhaité (correspondant au nombre de clusters), d'éviter l'explosion du nombre de règles par rapport au nombre de dimensions, et de capturer interactions entre variables. Cependant, les règles multidimensionnelles sont moins interprétables que le cas unidimensionnel. Nous pouvons rendre le système multidimensionnel plus interprétable en établissant différentes catégories linguistiques par variable (par exemple faible/moyen/élevé) et en projetant les centres de chaque cluster sur chaque axe de variable [14].
Notre algorithme d'interprétabilité suit les étapes suivantes :

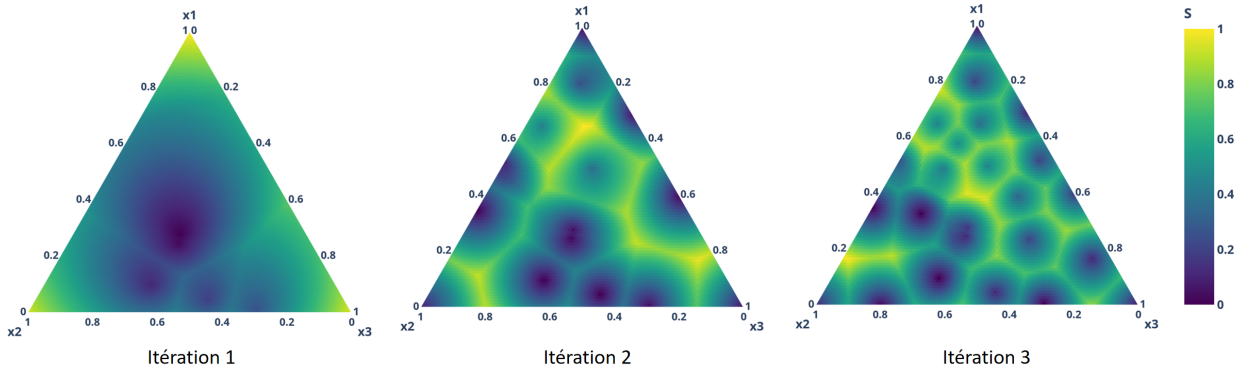


FIGURE 4 – Diagrammes ternaires illustrant l'évolution du score d'échantillonnage S à différentes itérations, pour un cas avec 3 variables d'entrée.

— Nous divisons chaque variable d'entrée et de sortie en différents ensembles flous triangulaires/trapézoïdaux f, dont les sommets correspondent aux centres des clusters projetés sur chaque axe. Si deux ensembles flous sont trop proches (par exemple distance $<$ distance seuil), nous les fusionnons ;
— Pour une variable x_i $(i \in [1; N+1])$, en notant m_{c_i} la i^{ieme} composante du centre $\boldsymbol{m}_c$, alors le sous-ensemble associé au cluster R est celui avec le degré d'appartenance le plus élevé $\mu^f(m_{c_i})$.

Un cluster sera associé à $N+1$ sous-ensembles flous (un par variable).

La figure 5 illustre le système de règles interprétables obtenu pour un cas avec 1 variable d'entrée x et 3 clusters. Les règles 1D sont utilisées comme substituts des règles multidimensionnelles pour aider l'utilisateur à comprendre ce modèle.

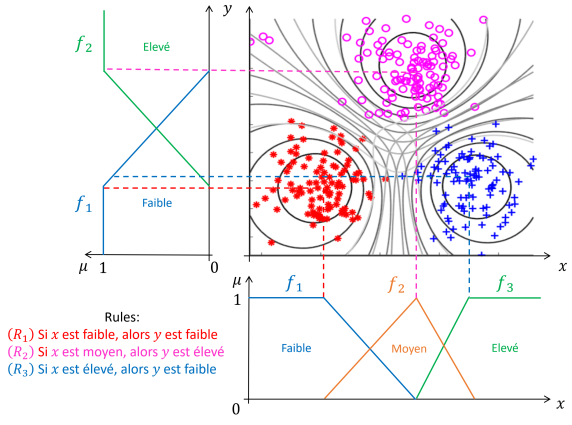


FIGURE 5 – Illustration de 3 clusters avec des degrés d'appartenance représentés par des lignes de niveau noires (tirées de [13]) et des ensembles flous f_i des variables x et y. L'association de chaque centre de cluster à ces ensembles flous 1D rend les règles plus interprétables.

6 Résultats expérimentaux

Dans cette section, nous présentons les résultats de différents tests pour caractériser notre algorithme. Par souci de reproductibilité, nous donnons d'abord quelques détails sur la mise en œuvre.

6.1 Considérations d'implémentation

Pour effectuer le clustering multidimensionnel, nous avons hybridé deux approches : nous utilisons la méthode de clustering hiérarchique [17] pour obtenir les centres des clusters et nous utilisons la fin de l'approche c-means floue [11] pour déterminer les fonctions d'appartenance. En effet, le clustering hiérarchique a l'avantage d'être totalement déterministe. Nous avons déterminé empiriquement le nombre de clusters pour chaque ensemble de données.

L'optimisation de la fonction de coût Eq. 7 peut être implémentée par la méthode Trust Region Reflective (TRF) [18], l'algorithme de Levenberg-Marquardt [9], l'algorithme des moindres carrés récursifs [8] ou descente de gradient [19, 10]. Nous avons utilisé l'algorithme TRF car il s'agit d'une méthode robuste (peu sensible au choix du point de départ), bien adaptée aux problèmes complexes avec des résidus non linéaires, adaptée aux grands problèmes clairsemés avec des bornes, et qui ne nécessite pas d'hyperparamètres supplémentaires.

6.2 Jeux de données jouet

Nous avons d'abord évalué notre méthode sur un jeu de données jouet que nous avons généré à partir d'une fonction sinus avec éventuellement plusieurs entrées. Un petit nombre d'entrées nous aide à visualiser les résultats pour les qualifier, tandis qu'un grand nombre permet de valider notre algorithme.

Fonction sinus à 1 variable d'entrée Nous avons d'abord testé notre algorithme de régression décrit dans la partie 3 avec le cas simple de la fonction $f(x) = \sin(2\pi x)$, avec 6 points initiaux et 2 clusters flous. Le résultat est tracé sur la Fig. 6, avec les valeurs de sortie prédites et les valeurs de sortie réelles.

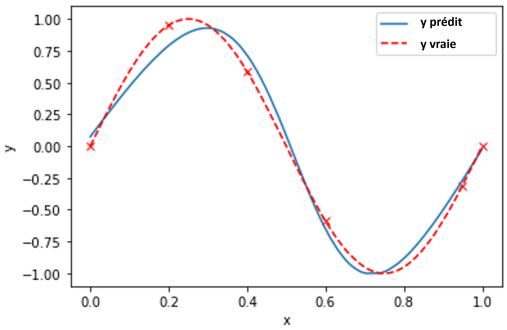


FIGURE 6 – Illustration de la sortie prédite $\hat{y}$ à l'aide de l'algorithme de régression basé sur le clustering flou.

Fonction sinus à 3 variables d'entrée Nous avons ensuite testé notre algorithme pour la fonction objectif sinus suivante avec 3 variables d'entrée :

$$f(\boldsymbol{x}) = \left| \prod_{k=1}^{3} \sin(k\pi x_k) \right|. \tag{13}$$

Cette fonction objectif a été choisie car elle présente une forme non triviale avec plusieurs maxima et un maximum global, avec des valeurs de sortie comprises entre 0 et 1, et parce qu'elle peut être tracée dans un diagramme ternaire (voir Fig. 7). Cela nous aide également à caractériser l'approche puisque nous n'aurons jamais un plan expérimental complet à partir d'une application réelle.

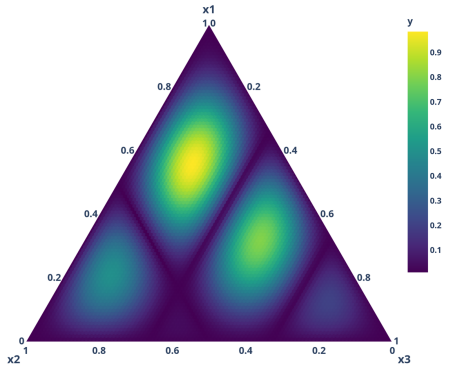


FIGURE 7 – Diagramme ternaire de la fonction objectif f donnée par l'équation (13).

Chaque axe x_i contient 100 valeurs de 0 à 1, donc au total nous avons 5151 points dans l'espace d'entrée. Notre objectif est de converger le plus rapidement possible vers la valeur optimale.
On choisit initialement 5 points aléatoires et on effectue 50 itérations (avec un point testé par itération). L'efficacité de notre algorithme est mesurée avec le critère du nombre d'itérations M nécessaires pour atteindre 80% de la valeur optimale de y (qui est de 0,984 dans notre cas d'étude). La figure 8 montre la meilleure valeur y obtenue parmi les points testés jusqu'à une itération donnée ; 80% de la valeur optimale est atteinte à l'itération $M = 24$.

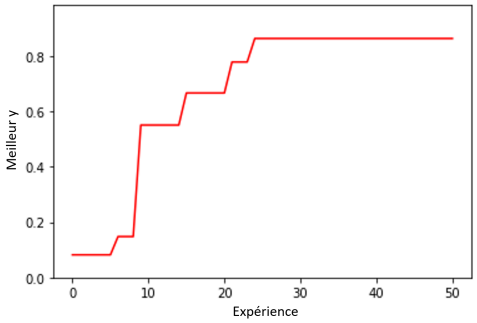


FIGURE 8 – Meilleure valeur de y obtenue pour chaque itération.

La proportion d'exploitation/exploration du point testé à chaque itération est tracée dans l'histogramme Fig. 9.

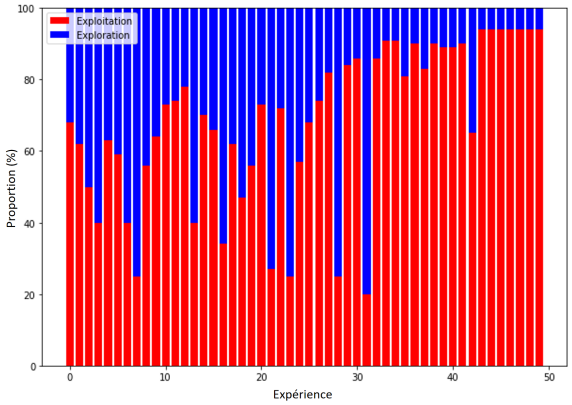


FIGURE 9 – Proportion exploitation/exploration pour chaque point testé.

On observe, comme prévu, que la proportion d'exploration diminue à mesure que le nombre de points échantillonnés augmente. A noter qu'il y a encore quelques explorations même après un nombre élevé d'expériences ; en effet, il vaut mieux continuer à explorer pour éviter de tomber sur un maximum local.
Nous avons ensuite testé la sensibilité de notre approche à son initialisation. En effet, le nombre d'expériences M nécessaires pour atteindre 80% de la valeur optimale dépend de la pertinence des points aléatoires initiaux. Nous avons répété la simulation 100 fois où à chaque simulation nous avons 5 points initiaux aléatoires avec $y < 0,1$ (c'est-à-dire que nous avons choisi des points non pertinents). Nous avons évalué que le nombre d'itérations nécessaires pour atteindre 80% de la valeur optimale est $M = 20 \pm 12.5$. Cette variabilité est due au fait que l'algorithme est sensible aux points initiaux, d'autant plus quand le nombre de points initiaux est faible.

Fonction sinus avec $N > 3$ variables d'entrée Pour nous rapprocher d'un problème du monde réel, nous avons testé notre approche avec plus de variables. Pour cela, nous considérons la fonction objectif suivante avec des $N \geq 3$

variables d'entrée :

$$f_N(\boldsymbol{x}) = \left| \prod_{k=1}^{N} \sin(i\pi x_k) \right| \tag{14}$$

avec $\boldsymbol{x} = (x_1, x_2..., x_N)$ variables continues, discrètes ou catégorielles. Les valeurs de y sont comprises entre 0 et 1. Plus précisément, nous étudions le cas de variables discrètes d'entrée $3 \leq N \leq 10$ (avec 5 valeurs différentes chacune). Ceci simule une expérience d'optimisation de composition dont les valeurs sont très contraintes.
Nous utilisons ce dernier jeu de données jouet pour comparer notre approche avec l'optimisation bayésienne (OB). L'OB diffère de notre algorithme flou pour les étapes suivantes [1, 20] :

— Pour l'OB, nous évaluons la fonction de substitution à l'aide du processus gaussien (GP) ou de l'algorithme de Parzen structuré arborescent (TPE), pour modéliser la fonction objectif avec une valeur moyenne m et une dispersion σ. Cet algorithme rend l'OB non déterministe. Pour notre algorithme flou, la fonction surrogate est obtenue à partir de l'algorithme de régression basé sur le clustering flou décrit en partie 3.

— Pour l'OB, le compromis entre exploitation et exploration est modélisé en utilisant la fonction d'acquisition "Expected Improvement" : $EI(\boldsymbol{x}) = (m(\boldsymbol{x}) - f(\boldsymbol{x}^*))\phi\left(\frac{m(\boldsymbol{x})-f(\boldsymbol{x}^*)}{\sigma(\boldsymbol{x})}\right) + \sigma\Phi\left(\frac{m(\boldsymbol{x})-f(\boldsymbol{x}^*)}{\sigma(\boldsymbol{x})}\right)$, avec ϕ/Φ la densité de probabilité/fonction de partition de la distribution normale et $f(\boldsymbol{x}^*)$ la meilleure valeur y obtenue jusqu'à présent. Pour notre algorithme flou, ce compromis exploitation/exploration est modélisé à travers le score d'échantillonnage $S(\boldsymbol{x}) = \hat{y}(\boldsymbol{x}) + \lambda\, d(\boldsymbol{x})$.

De manière analogue à notre algorithme, l'OB est répétée sur un certain nombre d'itérations. Chaque boucle fournit des informations supplémentaires jusqu'à atteindre une valeur optimale. L'algorithme TPE est efficace en termes de calculs et bien adapté aux problèmes d'optimisation de grande dimension avec une fonction objectif coûteuse [21]. De plus, il est bien adapté aux problèmes d'optimisation impliquant des variables discrètes et catégorielles, ainsi que des variables continues [22].
Pour les deux algorithmes, nous évaluons le nombre d'itérations M nécessaires pour atteindre 80% de la valeur optimale pour un nombre N donné de variables d'entrée et pour 5 points initiaux donnés (voir Fig. 10). Dans le cas de l'OB, en raison du caractère aléatoire du calcul de la fonction surrogate, nous avons dû répéter la simulation plusieurs fois pour obtenir une valeur moyenne. Nous observons que globalement notre algorithme flou donne un meilleur résultat que l'OB, il converge avec moins d'itérations :

$$\forall N, M_{fuzzy} < M_{BO}. \tag{15}$$

De plus, l'OB est une méthode non déterministe et on observe que la disparité de convergence est assez importante (voir les barres d'écart type élevées en bleu sur la figure).

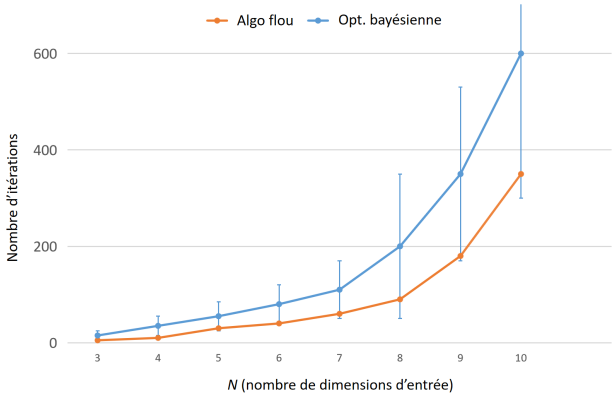


FIGURE 10 – Nombre d'itérations nécessaires pour atteindre 80% de la valeur optimale - Cas avec N variables discrètes.

6.3 Jeu de données réel

Nous avons ensuite testé notre approche sur un ensemble de données réelles du UC Irvine Machine Learning Repository appelé "Concrete Compressive Strength" [1].
Il comprend 8 variables d'entrée et le but est de maximiser la variable de sortie normalisée "Concrete compressive strength (MPa)". Cet ensemble de données comprend 1030 instances. En utilisant un processus de validation croisée (avec 80% de l'ensemble d'entraînement et 20% de l'ensemble de tests), nos simulations sur cet ensemble de données ont montré que la fonction de substitution de notre algorithme flou conduit à une meilleure prédiction de régression que le processus gaussien d'optimisation bayésienne : $RMSE = 8,7 \pm 2,9$ pour notre algorithme flou, et $RMSE = 15,1 \pm 3,6$ pour le processus gaussien. En utilisant notre substitut décrit dans la partie 5, nous obtenons $RMSE = 10.7 \pm 4.9$; ce résultat est moins bon que notre algorithme, soulignant la nécessité d'utiliser les règles N-dimensionnelles décrites dans la partie 3.2.
Pour déterminer le nombre d'expériences pour atteindre un point optimal, nous avons effectué le processus suivant : nous avons choisi 5 points aléatoires parmi les 1030 instances avec une mauvaise valeur de sortie $y < 0,1$; une expérience consiste ici à prendre un point parmi les instances choisies par l'algorithme et on souhaite converger rapidement vers un point optimal. La simulation complète est répétée plusieurs fois pour obtenir une valeur moyenne et un écart type. Le nombre d'expériences nécessaires pour atteindre 80% de la valeur optimale de y est $M = 9 \pm 6,7$. Qualitativement, notre algorithme est particulièrement utile lorsqu'il s'agit de données expérimentales. En effet, l'explicabilité est nécessaire pour comprendre pourquoi un point/une région précis est proposé par l'algorithme. Le déterminisme est également très important puisque l'expérimentateur ne souhaite pas se voir proposer un point différent à chaque fois qu'il exécute l'algorithme. De plus, des contraintes peuvent être prises en compte sur la base

1. https://archive.ics.uci.edu/dataset/165/concrete+compressive+strengt

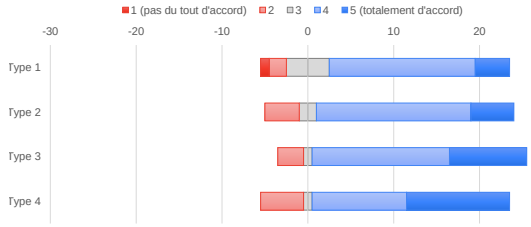


FIGURE 11 – Réponses à la question “Êtes-vous satisfait de la prochaine expérience proposée par l’algorithme ?”

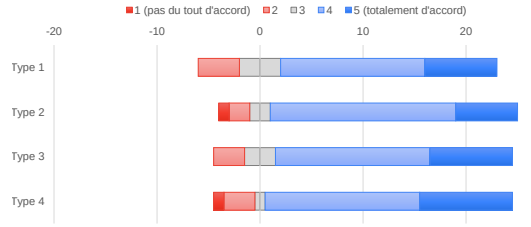


FIGURE 12 – Réponses à la question “Faites-vous confiance au choix de l’algorithme ?”

de connaissances expérimentales et théoriques.

7 Considération de l’utilisateur final

Nous avons évalué l’interaction avec les utilisateurs finaux à l’aide d’un questionnaire (évaluation basée sur l’humain). Le panel est constitué de 29 personnes travaillant dans le domaine de la science des matériaux, allant des chercheurs académiques aux chercheurs industriels, âgées de 22 à 62 ans. Pour recueillir leurs avis, nous avons utilisé une échelle de Likert en 5 points, allant de totalement en désaccord à totalement d’accord.

Le questionnaire décrit une situation basée sur un mélange de 3 composés (x_1, x_2, x_3) et une propriété (y) qui va de 0 à 1. Sur un diagramme ternaire, 17 expériences précédentes sont représentées avec les valeurs respectives de la propriété. Nous avons demandé au panel de comparer 4 résultats différents :

- Type 1 : l’algorithme donne exactement les valeurs suivantes pour x_1, x_2, x_3, comme dans l’équation. 9 ;
- Type 2 : idem Type 1 avec une explication (ex : “exploiter une zone déjà explorée et qui a donné de bons résultats”) ;
- Type 3 : l’algorithme donne une région d’intérêt comme dans l’équation 10 ;
- Type 4 : identique au Type 3 et avec la même explication que dans le Type 2.

La figure 11 montre les réponses à la question : “êtes-vous satisfait de la ou des prochaines expériences proposées par l’algorithme ?”. Les résultats sont majoritairement positifs, mais le troisième type de production semble avoir les meilleures notes (c’est-à-dire une région sans explication). De plus, la Figure 12 montre les réponses à la question : “faites-vous confiance au choix de l’algorithme ?”. Les deux types qui n’apportent aucune explication ont des résultats moins positifs. Cependant, les deux types qui fournissent une explication aux utilisateurs ont des résultats plus positifs, mais ils ont également un panéliste qui est totalement en désaccord. Nous avons regardé de plus près ses réponses et elles sont parfois contradictoires : cela peut être considéré comme une valeur aberrante.

Nous avons demandé aux panélistes quelle est leur sortie préférée : 12 d’entre eux répondent au Type 4, 8 répondent au Type 2, 5 pour le type 3 et enfin 4 pour le Type 1. Ainsi, les deux sorties préférées sont accompagnées d’explications.

Il est important de mentionner que nous avons demandé aux panélistes s’ils avaient peur de l’intelligence artificielle dans leur travail : 7 panélistes sont tout à fait d’accord et 10 d’entre eux sont d’accord. 9 d’entre eux sont restés neutres, ce qui signifie que seuls 3 d’entre eux n’ont pas peur.

Les résultats confirment ce que nous attendions : les utilisateurs finaux préfèrent avoir un choix et une explication, ce qui signifie que la méthode finale doit fournir la région de la prochaine expérience et une explication de la recommandation. Dans un commentaire, un panéliste a écrit qu’il préférait avoir plusieurs expériences possibles plutôt qu’une seule, mais puisque l’algorithme a choisi une expérience centrale, son choix serait le même. Il souligne l’importance de prendre en compte les préférences humaines dans de tels outils pour accroître leur acceptabilité.

8 Conclusion et perspectives

En conclusion, nous résumons les avantages de notre approche en fonction des résultats obtenus. Notre algorithme est transparent en conséquence directe du caractère interprétable de ses paramètres, de la prédominance d’un cluster dans chaque région de l’espace d’entrée-sortie, de la simplicité et de la nature linguistique de ses règles floues ; il est déterministe, c’est-à-dire qu’il converge vers les mêmes valeurs à chaque exécution ; il est nettement plus rapide que les approches méta-heuristiques telles que les algorithmes évolutionnaires ; il est robuste au surentraînement et résilient au bruit grâce à la contribution fusionnée des clusters ; enfin, des contraintes issues de la littérature peuvent être appliquées pour réduire l’espace de recherche d’entrée. Cette approche a le mérite d’être interprétable, intuitive, et peut être d’une réelle aide aux expérimentateurs.

L’originalité de notre algorithme inclut la transformation de clusters flous multidimensionnels en ensembles flous 1D interprétables, et la définition d’une fonction de score d’acquisition/échantillonnage à partir des variables $\hat{y}$ (valeur de sortie prédite) et d (distance à le point échantillonné le plus proche).

Les possibilités d’amélioration incluent la vitesse de calcul dans le cas de grande dimension, une meilleure détection des extrema locaux et une optimisation multi-objectifs.

Remerciements

Ce travail est financé par le Programme Transversal de Compétences en Matériaux et Procédés du CEA.

Références

[1] S. GREENHILL et al. “Bayesian optimization for adaptive experimental design : A review”. In : *IEEE access* 8 (2020), p. 13937-13948.

[2] D.S. BEM et al. “Combinatorial experimental design using the optimal-coverage algorithm”. In : *Experimental Design for Combinatorial and High Throughput Materials Development* (2003).

[3] S. TAKAMOTO et al. “Towards universal neural network potential for material discovery applicable to arbitrary combination of 45 elements”. In : *Nature Communications* 13.2991 (2022).

[4] A. PAGLIARO et P. SANGIORGI. “AI in Experiments : Present Status and Future Prospects”. In : *Applied Sciences* 13.10415 (2023).

[5] B. CAO et al. “How To Optimize Materials and Devices via Design of Experiments and Machine Learning : Demonstration Using Organic Photovoltaics”. In : *ACS Nano* 12.8 (2018), p. 7434-7444.

[6] F. REN et al. “Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments”. In : *Science Advances* 4.4 (2018), eaaq1566.

[7] D. XUE et al. “Accelerated search for materials with targeted properties by adaptive design”. In : *Nature communications* 7.1 (2016), p. 1-9.

[8] S.L. CHIU. “Fuzzy model identification based on cluster estimation”. In : *Journal of Intelligent & fuzzy systems* 2.3 (1994), p. 267-278.

[9] K. WIKTOROWICZ. “RFIS : regression-based fuzzy inference system”. In : *Neural Computing and Applications* 34 (juill. 2022).

[10] J. VIAÑA et al. “Explainable fuzzy cluster-based regression algorithm with gradient descent learning”. In : *Complex Engineering Systems* (2022).

[11] J.C. BEZDEK, R. EHRLICH et W. FULL. “FCM : The fuzzy c-means clustering algorithm”. In : *Computers & Geosciences* 10.2 (1984), p. 191-203. ISSN : 0098-3004.

[12] I.B. TÜRKŞEN. “A review of developments from fuzzy rule bases to fuzzy functions”. In : *2012 Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS)*. 2012, p. 1-5.

[13] M.-J.e LESOT, M. RIFQI et B. BOUCHON-MEUNIER. “Fuzzy Prototypes : From a Cognitive View to a Machine Learning Principle”. In : *Fuzzy Sets and Their Extensions : Representation, Aggregation and Models*. Berlin, Heidelberg : Springer Berlin Heidelberg, 2008, p. 431-452. ISBN : 978-3-540-73723-0.

[14] G. TSEKOURAS et al. “A hierarchical fuzzy-clustering approach to fuzzy modeling”. In : *Fuzzy sets and systems* 150.2 (2005), p. 245-266.

[15] S. GUILLAUME et B. CHARNOMORDIC. “Generating an interpretable family of fuzzy partitions from data”. In : *IEEE Transactions on Fuzzy Systems* 12.3 (2004), p. 324-335.

[16] S. MANEEWONGVATANA et D.M. MOUNT. “Data Structures, Near Neighbor Searches, and Methodology”. In : American Mathematical Society, 2002. Chap. Analysis of approximate nearest neighbor searching with clustered point sets.

[17] L.-J. LI et Y.-L. LIANG. “A Hierarchical Fuzzy Clustering Algorithm”. In : *International Conference on Computer Application and System Modeling* (2010).

[18] A.R. CONN, N.I.M. GOULD et P.L. TOINT. “Trust-Region Methods”. In : *MPS-SIAM Series on Optimization 1. SIAM and MPS, Philadelphia* (2000).

[19] G.E. TSEKOURAS et al. “A fuzzy clustering-based algorithm for fuzzy modeling.” In : *WSEAS Transactions on Systems* 3.5 (2004), p. 1958-1963.

[20] S. WATANABE. “Tree-structured Parzen estimator : Understanding its algorithm components and their roles for better empirical performance”. In : *arXiv preprint arXiv :2304.11127* (2023).

[21] J. BERGSTRA, D. YAMINS et D.D. COX. “Making a Science of Model Search : Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures”. In : *Proc. of the 30th International Conference on Machine Learning* (2013).

[22] T. AKIBA et al. “Optuna : A next-generation hyperparameter optimization framework”. In : *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019, p. 2623-2631.

Notations

- N nombre de dimensions d’entrée
- $\boldsymbol{x}_s$ points déjà échantillonnés, $\boldsymbol{x}$ point de l’espace d’entrée, $\boldsymbol{x}_{next}$ prochain point testé
- f ensemble flou
- c cluster, C ensemble des clusters
- $\boldsymbol{m}_c$ centre du cluster c, n_c nombre de clusters, μ_c degré d’appartenance au cluster c
- R règle floue
- $\boldsymbol{\alpha}_c, \beta_c$ coefficients de régression relatifs au cluster c
- $\hat{y}_c$ sortie prévue pour une entrée $\boldsymbol{x}$, par rapport au cluster c
- $\hat{y}$ sortie globale prédite pour une entrée $\boldsymbol{x}$
- M nombre d’itérations nécessaires pour atteindre un point optimal

Validation temporelle explicable de faits par la découverte de contraintes temporelles complexes dans les graphes de connaissances

Thibaut Soulard, Joe Raad, Fatiha Saïs
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique
91190 Gif-sur-Yvette, France

prénom.nom@lisn.fr

Résumé

La question de la détection de fausses nouvelles a pris de l'importance avec la diffusion croissante de flux d'informations non vérifiées sur le Web. Pour relever ce défi, les graphes de connaissances (GC) sont un moyen efficace pour vérifier le contenu des informations diffusées grâce aux faits structurés qu'ils contiennent. Toutefois, la validité de certains faits est dépendante d'un certain contexte temporel. Cette validation temporelle est une question qui n'a pas encore reçu beaucoup d'attention dans la littérature. Notre travail introduit une nouvelle approche interprétable et explicable qui exploite la puissance des graphes de connaissances pour classer les faits en évaluant leur validité ou leur réfutation dans un intervalle temporel. Nous avons dévelopé une approche symbolique fondée sur les relations d'Allen entre intervalles temporels et qui étend ces relations aux séquences temporelles. Nous avons évalué notre approche sur l'un des plus grands graphes de connaissances disponibles publiquement et analysé les résultats en fonction de multiples hyper-paramètres. Nous avons procédé de même pour une variante neuro-symbolique que nous avons aussi proposée.

Mots-clés

Graphe de connaissances, Véracité, Séquence temporelle, IA Hybride

Abstract

The issue of fake news has gained significance in light of the escalating flow of unverified information among users. To face this challenge, Knowledge Graphs (KGs) are a means for verifying news content from statements in the KG. However, an aspect that has yet to receive attention is the temporal validation of these statements (i.e. verifying whether a statement is true in a given time interval). Our work introduces a new interpretable and explainable approach that leverages the power of KGs to classify user-inputted facts by assessing their temporal validity or refutation. We developed a symbolic approach that is based on Allen's relations between temporal intervals and extends these relations to time sequences. We test our symbolic framework on one of the largest publicly available KGs and compare its performance across multiple hyper-parameters with the neuro-symbolic extension that we also developed.

Keywords

Knowledge Graph, Veracity, Temporal Aequence, Hybrid AI

1 Introduction

Les graphes de connaissances sont devenus des sources de données cruciales, en particulier pour le développement de nombreuses applications liées à l'intelligence artificielle. En s'appuyant sur un langage simple et standardisé tel que RDF[1], les graphes de connaissances peuvent représenter des faits et des connaissances complexes du monde réél, sous la forme de triplets : `<sujet, propriété, objet>`.

Cependant, parmi le grand nombre de graphes de connaissances publiés ces dernières années, seuls quelques-uns associent explicitement une composante temporelle à leurs faits. Cette composante temporelle, incluse par exemple dans Wikidata[2] -l'une des plus grandes bases de données publiques, est essentielle pour la qualité des bases de données, car de nombreux faits ne sont valables que pendant un intervalle de temps spécifique ou à un moment précis. Par exemple, `<Obama, headOfState, USA>` n'est valable que dans l'intervalle de temps commençant à `2009/01/20` et se terminant à `2017/01/20`.

Ces graphes de connaissances étant de plus en plus utilisés pour des applications de vérification de faits, il devient nécessaire de s'assurer de leur validité temporelle. Dans ce travail, nous proposons une approche explicable pour vérifier si un fait est valide dans un intervalle de temps donné. Cette approche peut être appliquée à tout graphe de connaissances associant un intervalle temporel à ses faits.

Nous commençons par présenter les travaux connexes (section 2) et par introduire les notations nécessaires (section 3). Nous présentons ensuite notre approche pour découvrir les contraintes d'ordre temporel dans les données (section 4), telles que la contrainte selon laquelle un député doit

1. Resource Description Framework : https://www.w3.org/RDF/
2. https://www.wikidata.org/

être élu avant d'occuper sa fonction. Ensuite, nous étendons et combinons les contraintes découvertes pour prendre une décision sur la validité temporelle d'un fait donné (section 5). Nous explorons différentes méthodes de combinaison, telles que le traitement démocratique de toutes les contraintes en utilisant un système de vote pour évaluer la validité d'un fait, ou l'entraînement d'un modèle d'apprentissage automatique pour apprendre quelles contraintes temporelles sont les plus appropriées pour valider un fait. Enfin, nous montrons les premiers résultats d'évaluation de notre approche sur différentes classes, contenant des millions de faits temporels, dans le graphe de connaissance Wikidata (section 6).

2 État de l'art

Bien que les questions liées à la spécification du temps de validité des faits ou de leur portée temporelle, soient bien connues [9] dans la communauté des bases de données relationnelles, ce n'est que ces dernières années que ce sujet a pris une importance particulière dans le développement des graphes de connaissances [3, 13]. Lorsque ces graphes prennent en compte la dynamicité des prédicats des faits dont la vérité est une fonction du temps, un fait est alors considéré comme vrai dans un laps de temps donné. Ce type d'informations sur la dynamicité des prédicats, lorsqu'elles sont disponibles, peut améliorer les résultats des approches de complétion de graphes de connaissances [11] qui peuvent tirer parti de la portée temporelle pour désambiguïser les différents candidats possibles. Cependant, ces approches dépendent à la fois de la disponibilité et de la validité de ces informations temporelles dans le graphe de connaissances et de leur validité. C'est pourquoi certaines approches ont été conçues pour générer de nouvelles informations temporelles, comme décrit dans [11]. Ces approches peuvent s'intéresser soit à la tâche d'interpolation, où l'on peut compléter des faits qui se sont déroulés dans le passé, soit à une tâche d'extrapolation où l'on peut chercher à prédire l'apparition de nouveaux faits.

Comme présenté dans l'étude [11], les approches de complétion GC existantes qui s'intéressent aux informations temporelles utilisent diverses techniques telles que celles fondées sur l'apprentissage profond ou les approches neuro-symboliques. Les approches fondées sur l'apprentissage profond peuvent utiliser *la traduction dans un espace vectoriel*, *la décomposition du tenseur*, *GCN*, *LSTM*, *GRU* pour apprendre un modèle de prédiction. Les approches neuro-symboliques telles que [14] qui sont à la fois basées sur la connaissance du domaine et l'apprentissage automatique utilisent des contraintes temporelles pour injecter la connaissance du domaine dans le modèle de prédiction. Cette approche améliore TTransE [7] et TA-TransE [6] grâce à l'introduction d'un *ordre temporel* (par exemple, `wasBornIn` $\leq$ `diedIn`) pour les prédicats de la même entité et *disjonction temporelle*. Par exemple, dans les pays monogames, une personne ne peut pas être mariée à deux individus différents au cours du même intervalle temporel. Dans [4], les auteurs exploitent l'interaction entre les intervalles temporels des prédicats pour la même entité mais s'appuient sur *Markov Logic Networks* et *Probalistic Soft Logics* pour résoudre la tâche finale.

En ce qui concerne la découverte de contraintes temporelles, il existe dans les bases de données relationnelles, et plus particulièrement dans le domaine du profilage des données [1], certaines approches qui découvrent des relations d'ordre entre les attributs, telles que [5, 12]. Mais les contraintes découvertes n'impliquent que des opérateurs arithmétiques (c'est-à-dire $\leq, <$) et n'utilisent pas d'opérateurs dédiés aux intervalles. A notre connaissance, il n'existe aucune approche permettant de découvrir des contraintes temporelles avec l'expressivité présentée dans ce travail, et il n'existe aucune approche traitant du problème de la validation de l'information temporelle en combinant et en exploitant des contraintes temporelles aussi expressives que celles utilisées le travail que nous présentons dans cet article.

3 Préliminaires

Notre approche de validation de faits temporels est conçue pour être appliquée aux graphes de connaissances temporels et repose sur l'algèbre d'Allen [2] pour la comparaison des intervalles de temps. Dans cette section, nous présentons les notions préliminaires et introduisons les notations utilisées dans la suite de l'article.

3.1 Algèbre d'Allen

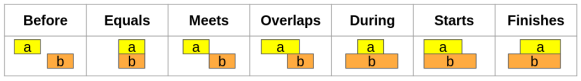


FIGURE 1 – Les 7 relations atomiques d'Allen

L'algèbre d'intervalles d'Allen est une référence pour la représentation des relations entre les intervalles de temps. Elle est composée de treize relations élémentaires qui sont distinctes, c'est-à-dire, qu'au plus une relation peut être énoncée pour une paire d'intervalles. Ces relations sont exhaustives car chaque paire d'intervalles peut être décrite par l'une des treize relations et qualitatives, dans le sens où qu'aucune durée numérique n'est prise en compte. Dans cet article nous utilisons seulement sept relations atomiques présentées dans la figure 1.

Definition 1 (L'ensemble des axiomes d'Allen)
L'ensemble des axiomes d'Allen peut être divisé en deux groupes différents : Axiomes disjoints (DA) représentant l'ensemble des axiomes où les intervalles sont disjoints, et Axiomes avec intersection (IA) représentant les axiomes où les intervalles impliquent une intersection temporelle :
– $DA = \{Before, Meets\}$,
– $IA = \{Equals, Overlaps, During, Starts, Finishes\}$.

Pour comparer deux intervalles de temps $I1$ et $I2$, nous pouvons soit spécifier la relation atomique entre eux, soit la généraliser en spécifiant s'ils sont disjoints ou s'ils ont une intersection. Par exemple, *before(I1, I2)* peut être généralisée en *DA(I1, I2)*.

3.2 Graphe de connaissances temporels

Dans ce travail, nous nous concentrons sur les graphes de connaissances (GC) temporels, c'est-à-dire les GC qui associent certains de leurs faits à une information temporelle pour exprimer l'intervalle de temps pendant lequel un fait est valide. Nous définissons tout d'abord les graphes de connaissances RDF, puis les graphes de connaissances temporels.

Definition 2 (Graphe de connaissance RDF) *Nous considérons un graphe de connaissances défini par une paire $(\mathcal{O}, \mathcal{G})$, où :*

- *$\mathcal{O} = (\mathcal{C}, \mathcal{P})$ est une ontologie représentée en OWL*[3] *et composée d'un ensemble de classes $\mathcal{C}$ et de propriétés $\mathcal{P}$.*
- *$\mathcal{G}$: est un graphe de données RDF, composé d'un ensemble de faits représentés par des triplets de la forme $\{(s, p, o) \mid s \in \mathcal{I}, p \in \mathcal{P}, o \in \mathcal{I} \cup \mathcal{L}\}$, où $\mathcal{I}$ est l'ensemble des entités (IRIs), $\mathcal{P}$ est l'ensemble des propriétés, et $\mathcal{L}$ est l'ensemble de littéraux (tels que des nombres ou des chaînes de caractères).*

Dans un graphe de connaissances composé d'un ensemble de triplets <s, p, o>, on peut distinguer trois types de faits :

- **Faits temporels concrets :** dont l'objet est de type `xsd:date` et dont la validité est illimitée dans le temps, comme : `<Mozart, dateNaissance, "1756/01/27">`.
- **Faits tautologiques :** vrais pendant toute la durée de vie d'une entité et dont le prédicat n'est pas sensible au temps, comme : `<Mozart, lieuNaissance, Salzbourg>`.
- **Faits dépendant du temps :** dont la validité est limitée à un intervalle de temps et dont le prédicat est sensible au temps, comme : `<Obama, présidentDe, USA>`.

Dans ce travail, nous nous concentrons sur *faits dépendant du temps* qui sont associés à une composante temporelle, en plus de *faits temporels concrets* qui aident à générer des contraintes temporelles.

Definition 3 (Graphe de connaissances temporels) *Nous définissons un graphe de connaissances temporel $\mathcal{TKG}$ comme un ensemble de quadruplets sous la forme de (s, p, o, t), qui étend les triples du graphe de données RDF en ajoutant la composante temporelle t exprimant la validité temporelle du fait qui peut être un un instant dans le temps ou un intervalle de temps.*

Nous considérons que information temporelle t peut être représenté par un intervalle de temps $[t; t + \epsilon]$, ϵ étant une durée insignifiante qui peut être déterminée en fonction de la granularité temporelle considérée (par exemple, des siècles, des années, des jours, des minutes). Par conséquent, dans le reste de l'article, nous nous référons à l'information temporelle en tant qu'intervalle de temps. Nous désignons par $\mathcal{T}$ l'ensemble de tous les intervalles utilisés dans TKG, chaque intervalle $I \in \mathcal{T}$ ayant une date de début et une date de fin, notées respectivement $I.s$ et $I.e$.

3. https://www.w3.org/OWL/

4 Découverte de contraintes temporelles

Pour valider un fait dans un graphe de connaissances temporel TKG, notre approche consiste à vérifier si l'information temporelle encodée pour ce fait est cohérente par rapport à une liste de contraintes temporelles. Ces contraintes peuvent exprimer soit une disjonction, soit une intersection entre les intervalles temporels des faits (voir la définition 1). Par exemple, une contrainte temporelle exprimant la disjonction peut indiquer qu'un président américain doit être élu *avant* d'occuper sa fonction, sous la forme `Before(elected, headOfState)`. Cependant, de telles contraintes temporelles sont rarement incluses dans le TKG et sont difficiles à collecter manuellement. C'est pourquoi, dans une première étape de notre approche, nous introduisons une nouvelle méthode pour découvrir ce type de contraintes temporelles à partir du TKG. Notre approche consiste d'abord à découvrir toutes les contraintes temporelles pour une seule entité, soit des contraintes simples qui peuvent être exprimées en utilisant les axiomes d'Allen (section 4.2), soit des contraintes complexes (section 4.3). Ensuite, les contraintes découvertes sont évaluées et généralisées à toutes les entités du TKG du même type (section 4.4).

4.1 Définition et comparaison des séquences temporelles

Pour chaque entité du TKG, nous pouvons construire une séquence temporelle pour chaque propriété dépendant du temps et décrivant l'entité dans le graphe (voir la définition 4). Par abus de langage, nous utiliserons parfois la formulation *un quadruplet se produit dans un intervalle de temps* pour faire référence au fait que le fait est valide dans l'intervalle de temps.

Definition 4 (Séquence temporelle) *La séquence temporelle d'une entité x pour une propriété p est l'ensemble ordonné des intervalles S des quadruplets $\{q_1, \ldots, q_n\}$, sous la forme de $< x, p, y_k, I_k >$ avec I_1 ayant la date de début la plus ancienne et I_n ayant la date de début la plus tardive dans la séquence temporelle.*

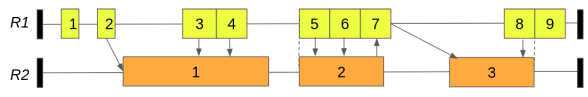


FIGURE 2 – Exemple d'une paire de deux séquences temporelles contenant respectivement 9 et 3 intervalles de temps. Les flèches renvoient à des comparaisons entre intervalles dans les séquences.

La figure 2 présente les séquences temporelles de deux propriétés $R1$ et $R2$ pour une seule entité, chaque élément de la séquence représentant un intervalle de temps. Afin de pouvoir comparer les intervalles de temps au sein d'une même

séquence temporelle et entre différentes séquences temporelles, nous nous limitons dans ce travail aux propriétés qui sont temporellement fonctionnelles (voir la définition 5).

Definition 5 (Propriété temporellement fonctionnelle) *Une propriété p est temporellement fonctionnelle si, pour chaque entité, il n'existe pas une paire d'intervalles se chevauchant dans la séquence temporelle correspondante. Nous notons $\mathcal{FP}$ l'ensemble des propriétés temporellement fonctionnelles. Une propriété p est dans $\mathcal{FP}$ ssi :*

$$\forall x \in \mathcal{I},\ \forall y_1, y_2 \in \mathcal{I} \cup \mathcal{L},\ \forall I_1, I_2 \in \mathcal{T},$$

$$< x, p, y_1, I_1 > \wedge < x, p, y_2, I_2 > \wedge DA(I_1, I_2)$$

Par exemple, `headOfState` est temporellement fonctionnelle, puisqu'un président ne peut pas être `headOfState` de deux pays différents dans des intervalles de temps qui se chevauchent.

Dans notre approche, pour une entité donnée, nous générons les séquences temporelles pour toutes ses propriétés temporellement fonctionnelles dans TKG. L'objectif est de générer des contraintes temporelles en comparant les intervalles au sein de la séquence temporelle (intra-séquence) et entre ses séquences temporelles (inter-séquence). Pour éviter la génération de contraintes bruitées et inutiles, nous limitons les comparaisons aux intervalles pertinents qui sont proches dans les séquences, sur la base des définitions suivantes :

Definition 6 (Comparaisons intra-séquence pertinentes) *Pour une séquence temporelle donnée S, les comparaisons intra-séquence pertinentes sont l'ensemble des paires d'intervalles consécutifs.*

Par exemple, dans la figure 2, il y a quatre comparaisons intra-séquence pertinentes : `Meets(3,4)`, `Meets(5,6)`, `Meets(6,7)`, et `Meets(8,9)`. Ces informations sont ensuite stocké dans la matrice $M_{\triangleleft}$.

Definition 7 (Comparaisons pertinentes entre séquences) *Pour une paire donnée de séquences temporelles S et S' des propriétés temporellement fonctionnelles P et P' respectivement, deux intervalles I de S et I' de S' sont considérés comme pertinents pour la comparaison si :*

$(I \cap_t I' \neq \emptyset)$
$\vee\ (I.s < I'.e)$
$\wedge(\nexists I'' \in S \setminus \{I\}, (I''.s \geq I.e \wedge I''.s \leq I'.s)$
$\wedge(\nexists I'' \in S' \setminus \{I'\}, (I''.e \leq I'.e \wedge I''.e \geq I.s)$
$\vee\ (I.s > I'.e)$
$\wedge(\nexists I'' \in S \setminus \{I\}, (I''.e \geq I'.s \wedge I''.e \leq I.s)$
$\wedge(\nexists I'' \in S' \setminus \{I'\}, (I''.s \leq I.s \wedge I''.s \geq I'.e),$

Nous désignons l'ensemble des inter-comparaisons pertinentes entre S et S' par $\Omega(S, S')$.

Dans l'exemple de la figure 2, les comparaisons inter-séquences pertinentes sont illustrées par des flèches. Elles peuvent également être représentées dans une matrice $M_{\triangleright}$

Axiom	o($R_1.I$, $R_2.I$)	o($R_2.I$, $R_1.I$)
Before	2	0
Equals	0	0
Meets	0	0
Overlaps	0	1
During	3	0
Starts	1	0
Finishes	1	0

TABLE 1 – Matrice $M_{\triangleright}$ pour les séquences S_1 et S_2

qui peut être utilisée pour indiquer le nombre de comparaisons inter-séquences pertinentes qui remplissent chaque axiome, comme présenté dans le tableau 1. Par exemple, les comparaisons `Starts(5,2)` et `Finishes(8,3)` représentent respectivement les seules comparaisons de début et de fin dans $M_{\triangleright}$.

4.2 Découverte de contraintes temporelles simples

Sur la base du nombre d'intervalles dans chaque séquence temporelle, l'algorithme fournit les axiomes d'Allen o tels que l'expression : $\forall I$ d'une séquence S_1, $\exists\ I'$ de la séquence S_2, tel que o($S_1.I$, $S_2.I'$) est satisfaite et que $(I, I') \in \Omega(S_1, S_2)$. Nous notons que dans notre méthode, l'axiome $equal$ est assoupli en une contrainte non symétrique que nous notons $subsumes$. Dans l'exemple du tableau 1, il n'y a pas d'axiome d'Allen généralisable pour les deux séquences S_1 et S_2.

4.3 Découverte de contraintes temporelles complexes

Afin d'obtenir des contraintes temporelles plus expressives, notre algorithme procède en combinant plusieurs axiomes d'Allen. L'algorithme obtient un ensemble de contraintes temporelles complexes de différents types. D'abord les contraintes complexes qui sont fondées sur les axiomes d'Allen exprimant la disjonction temporelle (DA) et celles qui sont fondées sur les axiomes d'Allen exprimant une certaine intersection (IA). En outre, les contraintes obtenues peuvent être représentées dans un arbre d'ordonnancement, dans lequel les contraintes sont organisées grâce à la relation `isMorePrecise`. Comme dans les axiomes d'Allen, certaines de nos contraintes complexes sont symétriques, comme `NAND`, alors que `Sequence Meets` ne l'est pas. La figure 3 présente l'arbre d'ordonnancement de toutes les contraintes complexes que nous considérons.
Dans ce qui suit, nous définissons formellement les contraintes qui peuvent être découvertes par notre algorithme.

4.3.1 Contrainte NAND.

Pour deux séquences temporelles S et S', et leur ensemble correspondant d'inter-comparaisons pertinentes $\Omega(S, S')$, une contrainte `NAND` exprime que pour chaque inter-comparaison pertinente (i, i') il n'y a pas d'axiome d'intersection qui soit satisfait.

Definition 8 (Contrainte NAND) *Considérons IA l'ensemble des axiomes d'intersection (Définition 1), les deux*

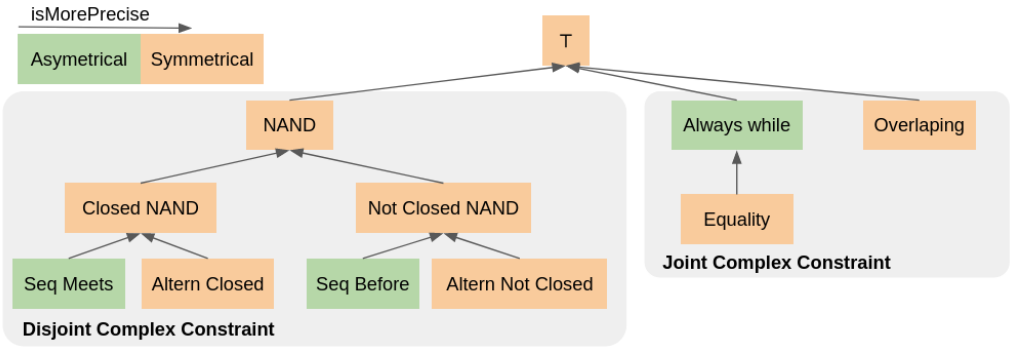


FIGURE 3 – Schema des différentes relation complexes.

séquences temporelles S et S' des propriétés P et P' respectivement, et la matrice d'inter-comparaisons $M_{\triangleleft}$ de S et S' remplit la contrainte NAND si :

$$(\sum_{a \in IA} M_{\triangleright}[a][o(P,P')]) = 0$$

4.3.2 Contrainte NAND fermée.

Pour deux séquences temporelles S et S', et leur ensemble correspondant de comparaisons pertinentes $\Omega(S,S')$, une contrainte NAND fermée exprime qu'aucune interruption n'apparaît entre le premier et le dernier quadruplet, quelle que soit la séquence temporelle.

Definition 9 *(Contrainte NAND fermée.) Considérons les deux séquences temporelles S et S' des propriétés P et P' respectivement, la matrice des inter-comparaisons $M_{\triangleright}$ de S et S', et la matrice des intra-comparaisons $M_{\triangleleft}$ de S et S' remplissent la contrainte NAND fermée si :*

$$M_{\triangleright}[Meets][o(P,P')] + M_{\triangleright}[Meets][o(P',P)] +$$
$$M_{\triangleleft}[Meets][P] + M_{\triangleleft}[Meets][P'] = |S| + |S'| - 1$$

4.3.3 Contrainte d'alternance fermée.

Pour deux séquences temporelles S et S', et leur ensemble correspondant de comparaisons pertinentes $\Omega(S,S')$, une contrainte d'*alternance fermée* exprime qu'après l'apparition d'un quadruplet d'une séquence temporelle, un quadruplet de l'autre séquence temporelle se produira.

Definition 10 *(Contrainte d'alternance fermée.) Considérons les deux séquences temporelles S et S' des propriétés P et P' respectivement, et la matrice d'inter-comparaisons $M_{\triangleright}$ de S et S' satisfait la contrainte d'*alternance fermée *si :*

$$M_{\triangleright}[Meets][o(P,P')] + M_{\triangleright}[Meets][o(P',P)]$$
$$= |S| + |S'| - 1$$

4.3.4 Contrainte de séquence coïncidente.

Pour deux séquences temporelles S et S', et leur ensemble correspondant de comparaisons pertinentes $\Omega(S,S')$, une contrainte de séquence coïncidente appelée *Sequence Meets* exprime que le dernier quadruplet de S rencontre le premier quadruplet de S'.

Definition 11 *(Contrainte de séquence coïncidente.) Considérons AA l'ensemble des axiomes conjoints (Définition 1), les deux séquences temporelles S et S' des propriétés P et P' respectivement, et la matrice d'inter-comparaisons $M_{\triangleright}$ de S et S' remplit la contrainte de séquence coïncidente si :*

$$M_{\triangleright}[meets][o(P,P')] = 1 \wedge$$
$$(\sum_{a \in AA} M_{\triangleright}[a][o(P,P')] + M_{\triangleright}[a][o(P',P)]) = 1$$

4.3.5 Contrainte NAND non fermée.

Pour deux séquences temporelles S et S', et leur ensemble correspondant de comparaisons pertinentes $\Omega(S,S')$, une contrainte *NAND non fermée* exprime qu'un espace apparaît toujours entre les intervalles (inter ou intra séquence temporelle).

Definition 12 *(Contrainte NAND non fermée). Considérons IA l'ensemble des axiomes intersectés (Définition 1), les deux séquences temporelles S et S' des propriétés P et P' respectivement, la matrice des inter-comparaisons $M_{\triangleright}$ de S et S', et la matrice des intra-comparaisons $M_{\triangleleft}$ de S et S' remplit la contrainte NAND non fermée si :*

$$M_{\triangleleft}[meets][P] + M_{\triangleleft}[meets][P'] = 0 \wedge$$
$$(\sum_{a \in AA/\{before\}} M_{\triangleright}[a][o(P,P')] + M_{\triangleright}[a][o(P',P)]) = 0$$

4.3.6 Contrainte d'alternance non fermée.

Pour deux séquences temporelles S et S', et leur ensemble correspondant de comparaisons pertinentes $\Omega(S,S')$, une contrainte *Alternance non fermée* exprime qu'après l'apparition d'un quadruplet d'une séquence temporelle, un quadruplet de l'autre séquence temporelle se produira après un intervalle.

Definition 13 *(Contrainte d'alternance non fermée.) Considérons les deux séquences temporelles S et S' des propriétés P et P' respectivement, et la matrice d'inter-comparaisons $M_{\triangleright}$ de S et S' satisfait la contrainte d'alternance non fermée si :*

$$M_{\triangleright}[before][o(P,P')] +$$
$$M_{\triangleright}[before][o(P',P)] = |S| + |S'| - 1$$

4.3.7 Contrainte de séquence-amont.

Pour deux séquences temporelles S et S', et leur ensemble correspondant de comparaisons pertinentes $\Omega(S, S')$, une contrainte séquence-amont appelée*sequence-before* exprime que le dernier quadruplet de S se produit avant tous les autres quadruplets de S'.

Definition 14 *(Contrainte de sequence-before.) Considérons IA l'ensemble des axiomes d'intersection (Définition 1), les deux séquences temporelles S et S' des propriétés P et P' respectivement, et la matrice d'inter-comparaisons $M_{\triangleright}$ de S et S' remplit la contrainte* sequence-before *ssi :*

$$M_{\triangleright}[before][o(P, P')] = 1 \wedge$$

$$(\sum_{a \in AA} M_{\triangleright}[a][o(P, P')] + M_{\triangleright}[a][o(P', P)]) = 1$$

4.3.8 Contrainte d'apparition simultanée.

Pour deux séquences temporelles S et S', et leur ensemble correspondant de comparaisons pertinentes $\Omega(S, S')$, une contrainte *d'apparition simultanée* exprime que tous les quadruplets d'une séquence temporelle partagent une intersection avec un autre quadruplet de l'autre séquence temporelle qui est égale à son intervalle temporel (c'est à dire que $q.I \cap_T q'.I = q.I$).

Definition 15 *(Contrainte d'apparition simultanée.) Étant donné la paire de séquences temporelles S et S' des propriétés P et P' respectivement, et la matrice d'inter-comparaisons $M_{\triangleright}$ de S et S' remplit la contrainte d'apparition simultanée si :*

$$M_{\triangleright}[equals][o(P, P')] + M_{\triangleright}[during][o(P, P')] +$$

$$M_{\triangleright}[starts][o(P, P')] + M_{\triangleright}[finishes][o(P, P')] = |S|$$

4.3.9 Contrainte d'égalité.

Pour deux séquences temporelles S et S', et leur ensemble correspondant de comparaisons pertinentes $\Omega(S, S')$, une contrainte d'*égalité* exprime que chaque quadruplet d'une séquence temporelle a un quadruplet dans l'autre séquence temporelle qui a le même intervalle.

Definition 16 *(Contrainte d'égalité.) Considérons les deux séquences temporelles S et S' des propriétés P et P' respectivement, et la matrice d'inter-comparaisons $M_{\triangleright}$ de S et S' satisfait la contrainte d'égalité si :*

$$M_{\triangleright}[equality][o(P, P')] +$$

$$M_{\triangleright}[equality][o(P', P)] = |S| + |S'|$$

4.3.10 Contrainte de chevauchement.

Pour deux séquences temporelles S et S', et leur ensemble correspondant de comparaisons pertinentes $\Omega(S, S')$, une contrainte *chevauchement* exprime que chaque quadruplet chevauche un quadruplet de l'autre séquence temporelle (à l'exception du quadruplet qui commence le plus tard).

Definition 17 *(Contrainte de chevauchement.) Considérons IA l'ensemble des axiomes d'inersection (Définition 1), les deux séquences temporelles S et S' des propriétés P et P' respectivement, et la matrice d'inter-comparaisons $M_{\triangleright}$ de S et S' remplit la contrainte de chevauchement si :*

$$M_{\triangleright}[overlapping][o(P, P')] +$$

$$M_{\triangleright}[overlapping][o(P', P)] = |S| + |S'| - 1$$

4.4 Généralisation des contraintes temporelles

Dans les sections précédentes, nous avons décrit comment des contraintes temporelles simples ou complexes peuvent être découvertes à partir de deux séquences temporelles pour une seule entité $e \in C$. Dans cette section, nous présentons notre approche pour généraliser les contraintes découvertes parmi l'ensemble d'entités de la classe C.
Pour évaluer si une contrainte peut être généralisée pour une classe C, nous introduisons deux mesures : le *taux d'erreur* et le *taux de généralisation*. La première permet de prendre en compte l'imperfection des données dans un graphe de connaissances temporel lorsqu'une entité est décrite avec des informations temporelles erronées, ou de surmonter la présence d'entités aberrantes qui ne suivent pas un comportement temporel similaire à celui d'autres entités dans C (voir la définition 18). Cette dernière méthode permet de filtrer les contraintes qui ne sont pas partagées par un pourcentage minimum d'entités dans C (voir la définition 19). Ensuite, étant donné un seuil d'erreur err et un seuil de généralisation gen, nous sélectionnons toutes les contraintes qui peuvent être généralisées, c'est-à-dire toutes les contraintes ayant un taux d'erreur supérieur à err et un taux de généralisation inférieur à gen (voir la définition 20). Enfin, parmi les contraintes généralisées restantes, nous ne conservons que celles dont la relation complexe la plus précise (comme décrit précédemment dans la section 4.3) afin de filtrer les contraintes temporelles redondantes.

Definition 18 (Taux d'erreur) *Étant donné la contrainte temporelle TC entre les propriétés P et P', l'ensemble des entités $E_{P,P'}$ de la classe C qui sont décrites par les deux propriétés, et le sous-ensemble d'entités $X_{P,P'} \subseteq E_{P,P'}$ où TC a été réfuté. Nous définissons le taux d'erreur comme suit :*

$$ErrorRate(TC) = \frac{|X_{P,P'}|}{|E_{P,P'}|}$$

Definition 19 (Taux de généralisation) *Étant donné la contrainte temporelle TC entre les propriétés P et P', l'ensemble des entités E de la classe C, et l'ensemble des entités $E_{P,P'} \subseteq E$ qui sont décrites par les deux propriétés. Nous définissons le taux de généralisation comme suit :*

$$GeneRate(TC) = \frac{|E_{P,P'}|}{|E|}$$

Definition 20 (Contraintes temporelles généralisées) *Étant donné un seuil d'erreur $err \in [0, 1]$ et*

un seuil de généralisation $gen \in [0,1]$, *une contrainte de temps* TC *peut être généralisée ssi :* $ErrorRate(TC) \leq err \wedge GeneRate(TC) \geq gen$.

4.5 Extension aux propriétés non fonctionnelles temporellement

Restreindre l'approche aux propriétés du graphe de connaissances temporel qui sont temporellement fonctionnelles peut conduire à ne pas prendre en compte des contraintes pertinentes, en particulier pour certaines propriétés ayant un large éventail de types de valeurs. Par exemple, une personne peut être liée par la propriété `memberOf` à différentes valeurs (par exemple, un groupe musical ou une certain congrès), chacune désignant un type d'assertion différent dans lequel les intervalles de temps sont susceptibles de se croiser dans la séquence temporelle. Ainsi, la probabilité de découvrir des contraintes temporelles pertinentes pouvant être généralisées à toutes les entités de la classe est plus faible. Par exemple, la propriété `memberOf` peut être spécialisée par valeur en `memberOf-USACongress` afin de découvrir des contraintes plus pertinentes entre la séquence temporelle spécialisée par valeur pour cette propriété et d'autres séquences temporelles (voir la définition 21).
Par conséquent, l'extension proposée dans cette section vise à améliorer la portée et la précision de notre approche, en spécialisant toutes les propriétés qui ont une entité comme valeur. Ces propriétés spécialisées sont ensuite utilisées pour générer les contraintes de la même manière que les propriétés normales. Pour les grands graphes de connaissances temporels, le processus de spécialisation des valeurs peut être limité aux propriétés qui ont comme valeurs des entités qui sont communément partagées entre plusieurs entités.

Definition 21 (Valeur de séquence temporelle spécialisée) *La séquence temporelle spécialisée d'une entité* x *d'une propriété* p *et pour une valeur* v $dans\mathcal{I}$ *est l'ensemble ordonné* S *d'un ensemble de quadruplets* $\{q_1, \ldots, q_n\}$, *sous la forme de* $< x, p, v, I_k >$ *tel que* $I_1.start$ *est le plus tôt* $I_n.start$ *est le plus tard.*

5 Validation des faits temporels basée sur des contraintes

Après avoir décrit notre approche de découverte et de généralisation des contraintes temporelles, nous présentons dans cette section comment une contrainte peut être appliquée pour valider ou réfuter un fait (section 5.1). Nous décrivons ensuite comment l'ensemble des contraintes peuvent être combinées et utilisées pour valider ou réfuter un fait (section 5.2).

5.1 Application d'une seule contrainte

Pour vérifier la validité d'un fait (s, p, o, t) dans un graphe de connaissances temporel, nous recherchons toutes les contraintes temporelles qui sont pertinentes pour ce fait. Pour qu'une contrainte $tc = o(P_1, P_2)$ soit pertinente pour

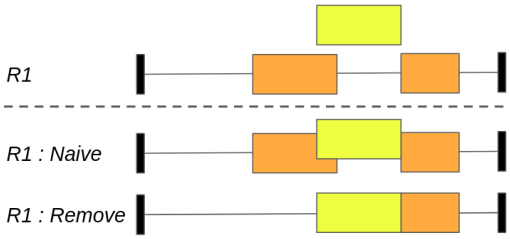


FIGURE 4 – Stratégies d'insertion dans la séquence temporelle $R1$ pour l'intervalle t du fait que nous cherchons à valider (représenté en jaune).

valider ou réfuter un fait, l'une des propriétés incluses dans une contrainte doit être liée à p : soit $P_k = p$, soit P_k représente la spécification de la valeur de p, avec $P_k \in P_1, P_2$.
Lorsqu'une contrainte temporelle est pertinente pour un fait, nous récupérons les séquences temporelles de l'entité s pour les relations P_1 et P_2. Ensuite, nous insérons le fait que nous voulons valider dans les séquences temporelles de s pour vérifier son comportement. Nous proposons deux stratégies d'insertion, illustrées dans la figure 4 : une **insertion naïve** et une **suppression de l'intersection**. Dans la première stratégie, nous ajoutons l'intervalle t du fait à la séquence temporelle de s, sans tenir compte du fait que cette insertion rompt la fonctionnalité temporelle de la propriété. Dans la stratégie de suppression de l'intersection, nous supprimons tous les intervalles de la séquence qui partagent une partie de leur ligne temporelle avec l'intervalle t.
Enfin, après l'insertion de l'intervalle du fait en utilisant l'une des deux stratégies, nous vérifions si les deux séquences temporelles sont toujours temporellement fonctionnelles et non vides. Si les deux conditions sont remplies, nous vérifions si les séquences temporelles respectent toujours la contrainte temporelle tc (ce qui indique que le fait est temporellement valide), ou si tc est maintenant violé (ce qui indique que le fait est réfuté).

5.2 Application globale des contraintes

Dans la section précédente, nous avons décrit comment la validité temporelle d'un fait peut être vérifiée par rapport à une seule contrainte temporelle. Cette section décrit comment un ensemble de contraintes temporelles peut être combiné et utilisé pour valider ou réfuter un fait. Nous proposons deux stratégies de combinaison : une *approche symbolique* basée sur un système de vote, et une *approche neuro-symbolique* qui tire parti des techniques d'apprentissage automatique pour apprendre quelles contraintes temporelles sont les plus appropriées pour valider temporellement un fait.

5.2.1 Approche symbolique par système de vote

Dans cette stratégie de combinaison, nous vérifions la validité temporelle d'un fait par rapport à toutes ses contraintes temporelles pertinentes. En utilisant un système de vote simple, nous comptons le nombre de contraintes que le fait respecte et le nombre de contraintes que le fait viole. Si le pourcentage de contraintes respectées est supérieur à

un seuil minimum, alors notre approche considère ce fait comme temporellement valide. Cette stratégie a l'avantage de rendre la décision résultante complètement explicable, c'est-à-dire que l'on peut retrouver toutes les contraintes temporelles qui ont été utilisées pour justifier la validation ou la réfutation d'un fait.

5.2.2 Approche Neuro-Symbolique

Dans cette stratégie de combinaison, nous vérifions la validité temporelle d'un fait par rapport à toutes les contraintes temporelles disponibles. Ensuite, pour chaque contrainte temporelle tc, nous associons un nombre pour représenter tous les comportements possibles : 0 si la tc est respectée par le fait ; 1 si la tc est violée ; et différentes valeurs pour indiquer si la tc n'est pas pertinente pour le fait, si la fonctionnalité temporelle de la séquence de la propriété est rompue, ou si l'une des séquences temporelles est vide. Il en résulte une matrice $n * m$, où n est la taille de l'ensemble des contraintes temporelles et m le nombre de faits à vérifier, ainsi qu'un vecteur de vérité terrain de dimension n qui peut être utilisé pour former et tester le modèle d'apprentissage automatique.

6 Évaluation expérimentale

Dans cette section, nous présentons l'évaluation expérimentale de notre approche sur une série de jeux de données. Toutes les expériences sont réalisées sur un processeur "Intel® Xeon® E5-2630 v4" avec 10 cœurs et 128 Go de RAM. Le code source et les jeux de données sont disponibles sur ce dépot github. [4]

6.1 Jeux de données

Nous évaluons notre approche sur trois jeux de données extraits extraits de de Wikidata [10], représentant tous les faits liés aux entités de type Pays (Q6256), Groupe musical (Q215380) et Homme politique (Q82955). Le tableau 2 indique le nombre d'entités pour chaque classe et le nombre total de faits temporels (quadruplets) les décrivant.

Les jeux de données ont été divisés selon un ratio de 80%, 10%, 10% pour l'ensemble d'apprentissage, de validation et de test. L'échantillonnage négatif a été effectué en changeant de manière aléatoire la partie temporelle de chaque fait autour de la durée de vie de l'entité. Ainsi, pour une entité ayant existé entre 1900 et 2000, la valeur aléatoire ne prend sa valeur qu'entre 1850 et 2050 afin de créer des faits raisonnablement faux. L'ensemble d'entraînement échantillonné non négatif est utilisé pour découvrir la contrainte temporelle, tandis que l'ensemble augmenté sert à l'entraînement de l'algorithme d'apprentissage automatique.

6.2 Réglage étape par étape

Pour évaluer les différents hyper-paramètres de notre approche, nous avons procédé étape par étape en réglant d'abord la stratégie pour la décision finale. Ensuite, le type de contrainte temporelle autorisé (le type par défaut est uniquement avec des relations), suivi de la stratégie d'insertion

4. https ://github.com/SoulardThibaut/TemporalConstraints

Classe	# Entités	# Quadruplets
Pays (Q6256)	205	183 249
Groupe de musique (Q215380)	55 507	131 476
Politicien (Q82955)	658 445	2 085 232

TABLE 2 – Description des trois ensembles de données

Classe	Deci. Type	Acc.	Cov.	RT
Country	Symbolic	79.5	9.4	**2m 30s**
	Neuro-Symb.	**80.4**	9.4	52m
Groupe de musique	Symb.	64.0	37.5	**2m 50s**
	Neuro-Symb.	**64.3**	37.5	5m 50s
Politicien	Symb.	61.6	44.0	**44m**
	Neuro-Symb.	**62.3**	44.0	1h 30m

TABLE 3 – Comparaison des stratégies de combinaison de contraintes

(la stratégie par défaut est la stratégie naïve). Puis la stratégie d'insertion (la stratégie par défaut est la stratégie naïve), et enfin le réglage du seuil d'erreur en même temps que celui de la généralisation (la valeur par défaut est $ET = 5\%$ et $GT = 5\%$). Chaque expérience est ensuite évaluée selon deux métriques : **Précision** (Acc.) et la **Couverture** (cov.) réalisée. Nous notons également le temps d'exécution (RT), de la découverte des contraintes à l'application, de chaque hyperparamètre.

6.2.1 Stratégies de combinaison des contraintes

La première expérience consiste à évaluer quelle stratégie, utilisée pour combiner toutes les contraintes temporelles de validation ou de réfutation d'un fait, permet d'obtenir de meilleures performances. Le tableau 3 présente les résultats de cette expérience sur les trois classes. Il montre que la stratégie neuro-symbolique peut valider des faits avec une précision légèrement supérieure à celle de la stratégie symbolique pour toutes les classes testées. Par conséquent, nous choisirons la stratégie de combinaison neuro-symbolique pour le reste des expériences, malgré l'augmentation significative du temps d'exécution.

6.2.2 Contraintes avec spécialisation des valeurs

La deuxième expérience consiste à évaluer si l'ajout de séquences temporelles spécialisées (RxV) permet d'obtenir de meilleures performances par rapport à l'utilisation de séquences temporelles normales (R). Le tableau 4 montre les avantages de l'extension de notre approche pour la classe *Politicien* (Q6256), où le pourcentage de couverture qui peut être fait est augmenté de près de 50% (de 9,4 à 14,1) et le nombre d'évaluations exactes est augmenté de 13% (de 80,4 à 90,8). Par conséquent, nous appliquons cette extension pour les expériences restantes, malgré l'absence d'amélioration pour les deux classes restantes, car elle ne présente pas d'inconvénients.

6.2.3 Stratégie d'insertion

Cette expérience consiste à comparer les deux stratégies d'insertion présentées à la section 5.1. Le tableau 5 montre que la stratégie de suppression d'insertion réduit légère-

Classe	Const. Type	Acc.	Cov.	RT
Pays	R	80.4	9.4	**52m**
	R & RxV	**90.8**	**14.1**	2h 35m
Groupe de musique	R	64.2	37.5	5m 50s
	R & RxV	64.2	37.5	5m 50s
Politicien	R	61.6	44.0	1h 30m
	R & RxV	61.6	44.0	1h 30m

TABLE 4 – Comparaison des spécialisations

Classe	Insert Type	Acc.	Cov.	RT
Country	Naive	**90.8**	14.1	**2h 35m**
	Remove	88.5	**14.4**	2h 54m
Groupe de musique	Naive	64.2	37.5	5m 50s
	Remove	64.2	37.5	5m 50s
Politicien	Naive	**62.3**	44.0	**1h 30m**
	Remove	62.1	**46.9**	1h 32m

TABLE 5 – Comparaison des stratégies d'insertions

ment la précision de notre approche, mais qu'en contrepartie, elle augmente légèrement le pourcentage de couverture possible. Pour les expériences suivantes, nous avons utilisé la stratégie de suppression d'insertion car elle permet d'évaluer le plus grand nombre de faits.

6.2.4 Seuils d'erreur et de généralisation

Le tableau 6 présente les résultats pour différents seuils d'erreur err et de généralisation gen. Nous pouvons constater qu'une valeur de err plus élevée et une valeur de gen plus faible permettent à notre approche d'évaluer un plus grand nombre de faits, sans impact significatif sur la précision. Cependant, cela est fait au détriment d'un temps d'exécution plus élevé car l'approche dispose d'un plus grand nombre de contraintes temporelles à utiliser pour valider ou réfuter un fait. Le nombre élevé de contraintes peut entraîner des problèmes d'évolutivité en termes de mémoire, par exemple, la première ligne où la présence d'environ 24.000 contraintes temporelles crée un problème pour le classifier basé sur les arbres des décison.

Classe	gen	err	Acc.	Cov.	RT
Country	2	10	-	-	-
	2	5	**88.6**	16	8h 50m
	5	10	87.9	**17.2**	4h 30m
	5	5	88.5	14.4	**2h 54m**
Groupe de musique	2	10	**64.6**	**38.2**	6m 6s
	2	5	**64.6**	**38.2**	5m 54s
	5	10	64.2	37.5	**5m 51s**
	5	5	64.2	37.5	**5m 51s**
Politicien	2	10	63.4	**51.9**	1h 35m
	2	5	62.1	49.9	**1h 32m**
	5	10	**63.5**	48.9	1h 34m
	5	5	62.1	46.9	**1h 32m**

TABLE 6 – Comparaison des seuils d'erreur et de généralisation

7 Conclusion

Dans cet article, nous avons présenté une nouvelle approche pour évaluer la validité temporelle des faits dans un graphe de connaissances. L'approche utilise et étend l'algèbre d'intervalles d'Allen pour découvrir les contraintes temporelles du graphe de connaissances. Pour l'évaluation, nous avons procédé à une mise au point étape par étape afin d'évaluer et d'expliquer l'impact de chaque stratégie proposée dans ce travail. Grâce à ces expériences, nous avons montré que notre approche peut valider ou réfuter un fait temporel avec une grande précision (jusqu'à 90.8%), malgré une couverture relativement faible (couverture maximale de 51.9%).
À l'avenir, nous nous efforcerons de résoudre les différents problèmes soulevés par les expériences. Nous prévoyons tout d'abord de réduire le nombre de caractéristiques qui ont un impact important sur les performances de l'approche neuro-symbolique, en éliminant les contraintes temporelles qui sont moins importantes. Ensuite, comme nous n'avons considéré que des ensembles de données extraits d'une seule source (Wikidata), nous n'avons pas pu évaluer si l'ensemble des contraintes temporelles découvertes dans un graphe peut être transféré et utilisé pour valider ou réfuter des faits temporels dans un autre graphe avec une précision et une couverture élevées. C'est pourquoi nous souhaitons tester la transférabilité de ces contraintes temporelles sur plusieurs autres graphes de connaissances temporels, tels que YAGO [8].

Références

[1] Ziawasch Abedjan, Lukasz Golab, Felix Naumann, and Thorsten Papenbrock. *Data Profiling*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2018.

[2] James F Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11) :832–843, 1983.

[3] Borui Cai, Yong Xiang, Longxiang Gao, Heng Zhang, Yunfeng Li, and Jianxin Li. Temporal knowledge graph completion : A survey. In *International Joint Conference on Artificial Intelligence*, 2022.

[4] Melisachew Chekol, Giuseppe Pirrò, Joerg Schoenfisch, and Heiner Stuckenschmidt. Marrying uncertainty and time in knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[5] Cristian Consonni, Paolo Sottovia, Alberto Montresor, and Yannis Velegrakis. Discovering order dependencies through order compatibility. In Melanie Herschel, Helena Galhardas, Berthold Reinwald, Irini Fundulaki, Carsten Binnig, and Zoi Kaoudi, editors, *Advances in Database Technology - 22nd International Conference on Extending Database Technology, EDBT 2019, Lisbon, Portugal, March 26-29, 2019*, pages 409–420. OpenProceedings.org, 2019.

[6] Alberto García-Durán, Sebastijan Dumančić, and Mathias Niepert. Learning sequence encoders for tem-

poral knowledge graph completion. *arXiv preprint arXiv :1809.03202*, 2018.

[7] Julien Leblay and Melisachew Wudage Chekol. Deriving validity time in knowledge graph. In *Companion Proceedings of the The Web Conference 2018*, pages 1771–1776, 2018.

[8] Farzaneh Mahdisoltani, Joanna Biega, and Fabian M Suchanek. Yago3 : A knowledge base from multilingual wikipedias. In *CIDR*, 2013.

[9] Vangipuram Radhakrishna, P. V. Kumar, and V. Janaki. A survey on temporal databases and data mining. In *Proceedings of the The International Conference on Engineering & MIS 2015*, ICEMIS '15, New York, NY, USA, 2015. Association for Computing Machinery.

[10] Denny Vrandečić and Markus Krötzsch. Wikidata : A free collaborative knowledgebase. *Commun. ACM*, 57(10) :78–85, September 2014.

[11] Jiapu Wang, Boyue Wang, Meikang Qiu, Shirui Pan, Bo Xiong, Heng Liu, Linhao Luo, Tengfei Liu, Yongli Hu, Baocai Yin, and Wen Gao. A survey on temporal knowledge graph completion : Taxonomy, progress, and prospects, 2023.

[12] Renjie Xiao, Zijing Tan, Haojin Wang, and Shuai Ma. Fast approximate denial constraint discovery. *Proc. VLDB Endow.*, 16(2) :269–281, oct 2022.

[13] Jiasheng Zhang, Shuang Liang, Yongpan Sheng, and Jie Shao. Temporal knowledge graph representation learning with local and global evolutions. *Knowledge-Based Systems*, 251 :109234, 2022.

[14] Jiasheng Zhang, Yongpan Sheng, Zheng Wang, and Jie Shao. Tkgframe : a two-phase framework for temporal-aware knowledge graph completion. In *Web and Big Data : 4th International Joint Conference, APWeb-WAIM 2020, Tianjin, China, September 18-20, 2020, Proceedings, Part I 4*, pages 196–211. Springer, 2020.

Graphaméléon : apprentissage des relations et détection d'anomalies sur les traces de navigation Web capturées sous forme de graphes de connaissances

L. Tailhardat[1,3], B. Stach[2], Y. Chabot[1], R. Troncy[3]

[1] Orange, France
[2] UTBM, Belfort, France
[3] EURECOM, Sophia-Antipolis, France

lionel.tailhardat@orange.com ; benjaminstach.pro@gmail.com ; yoan.chabot@orange.com ; raphael.troncy@eurecom.fr

Résumé

Les modèles comportementaux sont essentiels pour la détection d'anomalies ou d'actes malveillants sur des systèmes de télécommunication à travers le Web. Cependant, les données nécessaires ne sont pas toujours disponibles et une connaissance complète de la topologie des systèmes est nécessaire pour exploiter pleinement les inférences faites par ces modèles. Pour résoudre ce problème, nous proposons l'extension Web Graphaméléon et une représentation des traces de navigation sous forme de graphe de connaissances RDF en utilisant les ontologies UCO et NORIA-O.

Mots-clés

Traces de navigation Web, Analyse du comportement des utilisateurs et des entités (UEBA), Analyse des processus, Graphe de connaissances.

Abstract

Behavioral models are essential for detecting anomalies or malicious activities on telecommunications systems occuring through the Web. However, the necessary data is not always available, and a complete understanding of the system's topology is required to fully exploit the inferences made by these models. To address this issue, we propose the Graphameleon Web extension and a representation of navigation traces in the form of an RDF knowledge graph using the UCO and NORIA-O ontologies.

Keywords

Web Browsing Traces, User and Entity Behavior Analytics (UEBA), Process Mining, Conformance Checking, Knowledge Graph.

1 Introduction

En même temps que les technologies de l'information et de la communication évoluent et posent de nouveaux défis, la cybercriminalité n'a cessée d'augmenter durant la dernière décennie. Détecter et diagnostiquer rapidement des anomalies sur les réseaux et systèmes d'information sont de fait devenus une préoccupation majeure pour de nombreuses entreprises, notamment pour les gestionnaires de réseaux critiques et de grande envergure (téléphonie fixe et mobile, fourniture d'accès Internet, réseaux nationaux et internationaux d'échange de données). En cybersécurité, l'analyse du comportement des utilisateurs et des entités [1] correspond à un ensemble de techniques pour identifier et atténuer les menaces au niveau des éléments structurants des réseaux (p.ex. routeurs, serveurs, applications) à partir de données d'usage. Cela consiste typiquement à découvrir des motifs comportementaux nominaux (ou standards), tant au niveau des interactions entre les utilisateurs et les systèmes techniques qu'entre les éléments structurants eux-mêmes, et de s'en servir comme références pour alerter sur une utilisation potentiellement malveillante.

Une part importante des interactions utilisateurs-applications se fait désormais via une interface Web. Prenons l'exemple d'un scénario simple d'exploitation d'une vulnérabilité d'une application accessible via l'Internet [2] : après une phase de reconnaissance du système ciblé, l'attaquant accède directement à la page d'accueil de la plateforme de services, utilise une technique d'injection SQL [3] pour tromper le système d'authentification, exporte des données privées, puis quitte le service en naviguant directement vers une autre page Web. Analyser les interactions de l'utilisateur avec la plateforme, et ainsi détecter ce scénario, suppose l'analyse conjointe des journaux de l'application et du trafic réseau. Or les journaux peuvent être inaccessibles ou inutilisables en raison de problèmes de confidentialité ou de format. De même, le trafic réseau peut être chiffré ou inaccessible à la collecte. Ces deux aspects entraînent une perte des informations nécessaires pour qualifier le scénario d'attaque [13].

De nombreux outils de détection existent aujourd'hui dans le domaine de la cybersécurité, chacun se concentrant sur un type spécifique de source de données. Dans cet article,

1. "User and Entity Behavior Analytics" (UEBA) en Anglais.
2. https://attack.mitre.org/techniques/T1190/
3. https://fr.wikipedia.org/wiki/Injection_SQL

nous affirmons que la mise en œuvre simultanée de ces outils n'est pas suffisante pour une compréhension efficace des situations anormales, et qu'il est nécessaire d'utiliser un vocabulaire commun pour analyser les anomalies en associant les observables (p.ex journaux applicatifs, traces réseau, alertes des outils de détection) à la topologie du réseau. Dans cette optique, nous étendons le projet Dynagraph [23] (une approche combinant des outils de capture de traces avec une application Web pour un rendu graphique des données de navigation) afin d'apprendre des modèles d'activité interprétables sous forme de données liées : l'extension Web Graphaméléon collecte les traces d'activité de l'utilisateur (trafic réseau, interactions avec le navigateur Web) lors d'une session de navigation Web et sérialise ces données dans la syntaxe RDF selon le vocabulaire UCO [40]. Les données résultantes sont ensuite injectées dans un graphe de connaissances [1] pour interpréter les traces d'activité à un niveau sémantique et dériver des motifs, notamment sous forme de réseaux de Petri. Ces modèles d'activité peuvent ensuite être utilisés, du côté utilisateur ou du côté réseaux, pour identifier des situations analogues en les projetant sur le graphe de connaissances et, sur la base de cette projection, obtenir des informations contextuelles en parcourant le graphe.

Le reste de ce document est organisé comme suit. En Section 2, nous présentons les travaux connexes du point de vue de la cartographie du Web, de la modélisation de l'activité et de la détection des anomalies. En Section 3, nous présentons notre approche pour capturer les connaissances à partir des traces de navigation Web. Cela implique une modélisation de l'activité en trois couches (HTTP, micro-activités, macro-activités) basée sur le vocabulaire UCO. Nous décrivons également le composant de collecte Graphaméléon et l'utilisation des réseaux de Petri pour la détection des anomalies dans les traces de navigation. Nos expériences et résultats sont présentés en Section 4. Enfin, nous concluons et abordons les travaux futurs en Section 5. Le code source de Graphaméléon est disponible sur `https://github.com/Orange-OpenSource/graphameleon`, ainsi que le jeu de données expérimentales sur `https://github.com/Orange-OpenSource/graphameleon-ds`.

2 Travaux connexes

Collecte et représentation des connaissances. La cartographie du Web [12] est une thématique de recherche visant la compréhension de la structure du Web et de ses utilisateurs. Les études du domaine portent sur des sujets variés tels que les méthodes de prétraitement des données [26, 37], les techniques d'identification des utilisateurs [21], les algorithmes de reconnaissance de session [8, 31], et les méthodes de découverte de motifs [9]. Du point de vue de l'analyse de l'activité, le concept de raisonnement basé sur les traces [3] guide la conception d'outils d'interprétation sémantique des artefacts de services numériques en suggérant l'utilisation de vocabulaires contrôlés et de modèles de données liées. Concernant la représentation des événements et des activités au sein de graphes de connaissances, divers modèles de données – tantôt génériques, tantôt spécifiques à un domaine d'application – sont disponibles : modélisation de processus (BBO [4], réseaux de Petri [11, 18], HTTPinRDF [16, 28]) ; analyse causale (FARO [39]) ; cybersécurité (UCO [40], MITRE D3FEND [33]) ; opérations réseau (NORIA-O [25]) ; villes intelligentes (iCity ActivityOntology [29]).

Détection d'anomalies et analyse des processus. Pour la détection d'anomalies, diverses approches ont été proposées autour d'un principe commun d'identification des écarts par rapport aux comportements normaux, notamment par des modèles statistiques [38, 34, 32], des techniques d'apprentissage automatique [41, 30] et des méthodes basées sur les graphes [7, 10]. Le domaine de l'analyse des processus se concentre sur l'extraction de modèles de processus à partir de journaux d'événements et sur l'analyse du flux réel des activités. Ces modèles fournissent des informations sur le comportement typique et la structure sous-jacente des processus de navigation Web. Des techniques de vérification de conformité [17] ont été développées dans l'exploration de processus pour comparer le comportement observé aux modèles de processus attendus et identifier les écarts.

Positionnement. Nous étendons le concept de raisonnement basé sur les traces au domaine de la cartographie du Web en considérant l'utilisation des graphes de connaissances comme moyen de représenter les données de topologie du Web et les données d'utilisation de façon conjointe et cohérente. Nous abordons ainsi une nouvelle opportunité induite par l'émergence de modèles de données applicables dans les domaines des infrastructures réseaux (pour la description de systèmes hétérogènes) et de la cybersécurité (pour la description et la gestion des attaques et des risques). Nous supposons que, dans cette émergence, la communauté est désormais en mesure de répondre au besoin de corréler les informations d'usage du Web avec la description de la structure du Web lui-même, afin d'améliorer la compréhension et la conception de systèmes complexes tout en tenant compte du couple utilisateur-système. Pour exemple, il est évident que (en particulier dans UCO), l'analyse des attaques et des vulnérabilités repose principalement sur des indicateurs de compromission par l'énumération d'artefacts issus de situations passées. Ces indicateurs ne sont cependant jamais corrélés avec la topologie des réseaux et des services, ni même avec l'organisation temporelle des artefacts, ce qui correspond à une description statique des situations anormales et met de côté la structure propre des activités (i.e. la stratégie employée en rapport à la dynamique des événements). À cet égard, nous montrons notamment avec notre proposition comment incorporer le concept de traces de navigation dans l'ontologie UCO, ce qui permet de bénéficier simultanément des connaissances en cybersécurité et du contexte réseau enregistré par ailleurs (i.e. par les analystes en cybersécurité et les opérateurs réseau, respectivement) via l'ontologie NORIA-O, tout en garantissant une représentation nor-

malisée et homogène des données. De plus, nous étendons l'utilisation de l'analyse des processus et de la vérification de conformité aux graphes de connaissances, en capitalisant sur l'alignement des ces techniques avec les principes du raisonnement basé sur les traces.

3 Approche

L'approche proposée comporte trois parties, le but étant de réaliser une collecte de données dont le résultat permettra à la fois d'analyser les traces de navigation Web dans leur contexte réseau et d'apprendre des motifs d'activité. La première partie consiste à développer une modélisation sémantique des activités des utilisateurs sous forme de graphe de connaissances en réutilisant l'ontologie UCO (§3.1). La seconde partie concerne la conception de l'outil Graphaméléon, une extension de navigateur Web permettant de capturer les données de navigation et de les sérialiser en RDF (§3.2). La troisième partie porte sur l'intégration de Graphaméléon dans une chaîne de traitement de données (Figure 1) dont le principe est d'extraire des motifs d'activité en utilisant les outils d'analyse des processus et une représentation sous la forme de réseaux de Petri (§3.3).

3.1 Modélisation sémantique

Modélisation de l'activité des utilisateurs. Le concept d'activité manque de définition précise pour analyser la navigation sur le Web, car son interprétation repose fortement sur les données et l'échelle d'observation choisies. Il est en effet nécessaire de distinguer si l'identification d'une activité repose sur les interactions d'un utilisateur avec un site Web, ou si elle repose sur les échanges de paquets TCP entre le navigateur et le serveur portant ledit site Web. Pour commencer, supposons qu'une connexion HTTP est établie entre le navigateur Web de l'utilisateur et le serveur Web à partir d'une demande initiée par l'utilisateur. Le document demandé (p.ex. une page Web) nécessite, en règle générale, le chargement de ressources complémentaires telles que des images, des scripts ou autres. Ces dépendances impliquent un ensemble de sous-requêtes. Du point de vue de l'utilisateur, l'action consiste à naviguer vers un site Web par un clic sur un hyperlien (ou à accéder directement à la page via une URL), alors que du point de vue du navigateur Web, il s'agît d'une séquence de requêtes. De cette distinction, nous définissons deux niveaux de granularité pour discuter des traces de navigation. Celui nommé "micro-activité" correspond aux requêtes. Dans le niveau supérieur, nommé "macro-activité", nous considérons une trace comme étant un ensemble de requêtes et d'interactions. Par interaction, nous entendons toute action à l'initiative de l'utilisateur qui a une conséquence sur une page Web (p.ex. clic sur un hyperlien, renseigner un champ de formulaire, clic sur un bouton du navigateur Web).

Projection sémantique. Les graphes de connaissances permettent de gérer de façon unifiée des données hétérogènes et issues de sources variées. Le fonctionnement type des navigateurs Web repose d'ores-et-déjà sur des normes et des protocoles établis. Par rapport à cette normalisation, nous considérons que l'apport stratégique des graphes de connaissances est de faciliter l'intégration de données provenant de sources extérieures au contexte du navigateur Web. Nous remarquons que l'ontologie UCO [40] semble bien adaptée à notre objectif car elle permet la représentation des activités de navigation Web à différentes échelles, en incluant des informations concernant les cycles d'actions, les actions individuelles, les connexions réseaux, les protocoles de communication, les ressources techniques utilisées, les noms de domaines Internet et les adresses IP.
Ainsi, notre stratégie pour construire le graphe de connaissances consiste à maximiser la réutilisation des concepts/propriétés définis dans UCO, et de faire correspondre les champs et les valeurs capturés au niveau du navigateur Web avec ces concepts/propriétés chaque fois que leur sémantique s'aligne. La Figure 2 illustre cette mise en œuvre en présentant le modèle de données. Les règles de construction de graphe correspondantes en syntaxe RML [5] sont disponibles dans le dépôt de code `https://github.com/Orange-OpenSource/graphameleon`.
Dans les détails, une requête HTTP est représentée par une entité de la classe `ucobs:HTTPConnectionFacet`, et ses en-têtes sont représentées par des propriétés spécifiques telles que `ucobs:startTime` et `ucobs:endTime` pour les horodatages, et `core:tag` pour les en-têtes de type Fetch Metadata [36]. Une adresse IP ou une URL pouvant être communes à plusieurs requêtes (p.ex. un utilisateur répétant le même appel à un site Web, un site Web avec divers services hébergés sur le même serveur), ces éléments sont matérialisés par l'intermédiaire des classes `ucobs:IPAddressFacet` et `ucobs:URLFacet`, respectivement. Les références croisées entre les entités résultantes sont établies grâce à des propriétés telles que `ucobs:hasFacet` et `ucobs:host`. Pour les macro-activités, nous considérons les interactions de l'utilisateur comme des instances de la classe `ucoact:ObservableAction`, avec des relations vers les entités `ucobs:HTTPConnectionFacet` et `ucobs:URLFacet` mentionnées ci-dessus pour décrire le contexte dans lequel elles se produisent. De plus, nous utilisons les propriétés `types:threadNextItem` et `types:threadPreviousItem` de UCO pour représenter la chronologie des traces d'activité.

3.2 Collecte de données avec Graphaméléon

Le navigateur Web étant l'interface principale entre un utilisateur et le Web, nous considérons pour la suite que la collecte de données doit porter à la fois sur les requêtes HTTP et des interactions utilisateur/navigateur pour comprendre et analyser pleinement le système utilisateur-réseau-application car ces deux ensembles reflètent l'intention directe et indirecte de l'utilisateur.

Collecte des requêtes. À son activation au sein du navigateur, l'outil Graphaméléon associe des fonctions de rappel aux processus d'envoi et de réception du navigateur. Cela permet d'intercepter toutes les requêtes gérées par le navigateur pour récupérer des informations à partir

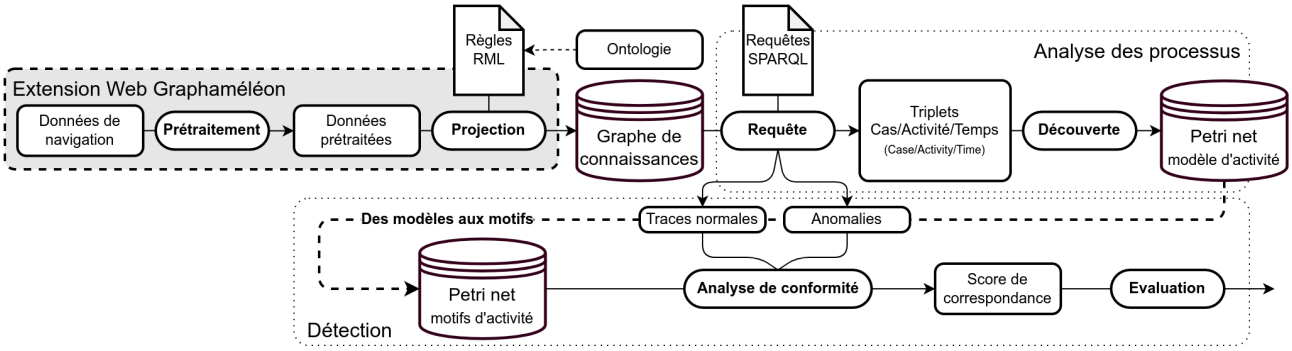


FIGURE 1 – Aperçu de la chaîne de traitement des données.
L'extension Web Graphaméléon capture et annote l'activité de l'utilisateur au niveau du navigateur Web. Un composant d'extraction des processus dérive des modèles d'activité places/transitions (Petri net) à partir du graphe de connaissances RDF résultant. Ces modèles peuvent être utilisés pour construire une bibliothèque de motifs d'activité, qui sont ensuite utilisés par un composant de vérification de conformité, côté client ou côté réseaux/serveurs, pour classer de nouvelles traces d'activité comme normales ou anormales.

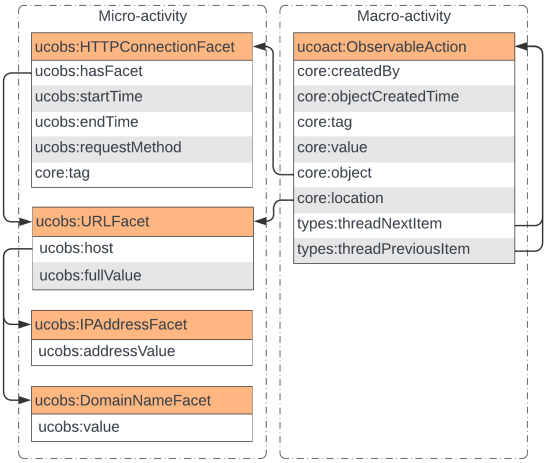


FIGURE 2 – Modèle de données.
Ce diagramme de classe définit les concepts et les propriétés utilisés pour la représentation sémantique des micro-activités (à gauche) et des macro-activités (à droite) tels que décrits dans la section 3.1. Pour les micro-activités, les classes et les propriétés présentées décrivent de manière précise une séquence de requêtes capturées au niveau du navigateur Web. Les macro-activités améliorent encore la modélisation en permettant la description des interactions. Les noms des concepts et des propriétés utilisés ici sont définis dans le vocabulaire UCO, les espaces de noms suivants s'appliquent : *core* = `https://ontology.unifiedcyberontology.org/uco/core#`, *ucobs* = `https://ontology.unifiedcyberontology.org/uco/observable#` et *types* = `https://ontology.unifiedcyberontology.org/uco/types#`.

des en-têtes de celles-ci. La Table 1 résume le type de données collectées par Graphaméléon. Ces informations incluent les URLs, les addresses IP et les noms de domaines associés, l'horodatage de la requête et les Fetch Metadata [4]. Les Fetch Metadata nous permettent de déduire des connaissances indirectes à partir des traces de navigation. Par exemple, le champ `Sec-Fetch-Site` indique la relation entre l'initiateur de la requête et sa cible, fournissant ainsi des informations sur la topologie du réseau. De même, le champ `Sec-Fetch-Mode` aide à différencier les requêtes initiées par l'utilisateur de celles correspondant à des sous-requêtes pour charger des images et autres ressources. Enfin, nous tokenisons les URLs utilisées dans les requêtes en remplaçant tous les arguments présents par les noms de leurs paramètres respectifs. Cela permet d'abstraire d'éventuelles informations de contexte définies par les sites Web, et éviter ainsi une diversité excessive dans l'interprétation des activités pour des cas similaires. Pour exemple, les URLs `https://www.shop.com/?client_id=2313` et `https://www.shop.com/?client_id=346`, indépendamment de l'utilisateur initiant ces requêtes, reflètent le même comportement. Après tokenisation, ces URLs sont représentées par `https://www.shop.com/?client_id=[client_id]`.

Portée	Paramètre ou nom de l'en-tête HTTP	Micro	Macro
Requête	Method	✓	✓
	URL	✓	✓
	IP	✓	✓
	Domain	✓	✓
	Sec-Fetch-Dest	✓	✓
	Sec-Fetch-Site	✓	✓
	Sec-Fetch-User	✓	✓
	Sec-Fetch-Mode	✓	✓
Interaction	EventType	-	✓
	Element	-	✓
	Base URL	-	✓
Les deux	User-Agent	✓	✓
	Start time	✓	✓
	End time	✓	✓

TABLE 1 – Données collectées par Graphaméléon.
Types de données collectées par l'extension Web Graphaméléon en fonction du mode de capture (micro-activité *vs* macro-activité), et regroupées selon leur portée (requête *vs* interactions *vs* les deux).

Collecte des interactions. Afin de collecter les interactions entre l'utilisateur et le navigateur, l'outil Graphaméléon lie un "script de contenu" à chaque onglet actif du navigateur. Ces scripts associent des fonctions de rappel à tous les éléments interactifs des pages Web, tels que les hyperliens, les boutons, les formulaires, etc. Cette approche minimise l'impact de la collecte sur les performances du navigateur et évite de capturer des interactions indésirables, telles que des clics erronés sur des éléments non interactifs.

Afin d'identifier les interactions, nous prenons en compte le type d'événement enregistré, l'élément avec lequel l'utilisateur a interagi, et l'URL de la ressource correspondante. Lorsqu'un élément a un attribut id, définir une référence

4. `https://www.w3.org/TR/fetch-metadata/`

vers celui-ci est évident. Ce n'est cependant pas le cas général et il est donc nécessaire de construire une référence grâce à la position absolue de l'élément dans la hiérarchie du DOM [5] ; pour exemple : `body > maindiv[2] > div > div > a`. Bien que cette méthode permette de faire référence de manière déterministe aux éléments de la page, il est important de noter que les références sous forme de chemin hiérarchique sont difficiles à interpréter sans une capture de la page Web et des interactions car ces références portent peu d'information sur la finalité de l'élément. Une alternative pour générer ces références consisterait à injecter des attributs id dans les éléments de la page Web à l'aide de fonctions de rappel, mais cela ne résout pas le problème de la stabilité des références entre chaque session de navigation pour les pages ayant un contenu dynamique.

3.3 Détection d'anomalies et réseaux de Petri

Trois familles de techniques de détection d'anomalies sont présentées dans [24] pour analyser des données de réseau représentées à l'aide d'un graphe de connaissances : *Model-Based Design*, où le graphe de connaissances contient les données nécessaires et suffisantes pour déduire les situations indésirables à l'aide de requêtes ; *Process Mining*, pour les situations liées à un modèle de décision et limitées dans le temps et l'espace, en utilisant des outils de vérification de conformité et une représentation des cas de détection sous forme de réseaux de Petri (réseaux P/T) ; *Statistical Learning* (apprentissage statistique) à l'aide de techniques de plongement de graphes [14] où les modèles d'anomalies (i.e. une généralisation du contexte pour un ensemble de situations anormales) sont dérivés de la structure du graphe de connaissances.

Dans ce travail, nous nous concentrons sur l'approche du *Process Mining* (analyse des processus), en considérant que la collecte de données à l'aide de Graphaméléon correspond à des sessions de navigation Web relativement bien définies en termes de durée et d'activités : un seul utilisateur génère une trace d'activité capturée au niveau du navigateur Web, trace qui peut être directement annotée par l'utilisateur en termes de but de l'activité à la fin de la session de navigation Web. Nous postulons que les traces d'activité sont similaires à des modèles de décision, car la séquence d'actions de l'utilisateur lors d'une session de navigation (p.ex. cliquer sur un hyperlien, utiliser le bouton de retour du navigateur Web, remplir une zone de saisie) conditionne l'atteinte d'un objectif spécifique (i.e. le but de l'activité) en fonction des informations présentées sur les pages Web. Nous supposons de même que les réseaux P/T sont une représentation adaptée pour analyser et catégoriser les traces de navigation car : *1)* ils possèdent une explicabilité intrinsèque par leur nature graphique ; *2)* les modèles de décision associés aux réseaux P/T peuvent se généraliser à différentes situations indépendamment de la représentation des connaissances sous-jacente ; *3)* les modèles de décision peuvent être facilement dérivés à partir de documents de spécifications produits par des experts métiers (p.ex. ingénieurs et techniciens réseaux), et implémentés sous forme de réseaux P/T à l'aide d'outils tels que TINA [6]. L'utilisation des réseaux P/T permet de tirer parti des techniques de détection d'anomalies couramment appelées "vérification de conformité", c'est-à-dire évaluer la pertinence d'une trace par rapport à un modèle donné (score de correspondance) ou rejouer une trace à travers un modèle pour analyser les étapes incohérentes au sein de l'activité.

Dans ce qui suit, nous définissons deux concepts pour clarifier la notion d'anomalie par rapport à ce qui est observé et à ce qui est attendu du point de vue des activités. Tout d'abord, nous définissons un "modèle d'activité" comme la traduction de toute trace d'activité (obtenue lors de la phase de collecte de données) en une représentation de type réseau P/T à l'aide d'un algorithme de découverte de processus (process mining). Selon cette définition, les collectes de données réalisées avec l'extension Web Graphaméléon (§3.2) permettent aux utilisateurs d'établir un catalogue de modèles d'activité. Ensuite, nous définissons un "motif d'activité" comme un modèle universel d'activité représenté avec des réseaux P/T. Un motif est établi en se basant soit sur une spécification du comportement attendu du couple utilisateur-système pour une situation spécifique, soit sur un comportement idéal dérivé de l'agrégation et du raffinement de plusieurs traces d'activité provenant du catalogue de modèles d'activité. Nous considérons que la gestion et la conversion des modèles d'activité en modèles relèvent de la responsabilité de l'utilisateur (analyse, sélection, raffinement), et dépasse le cadre de cet article.

La détection d'anomalies est donc définie par la comparaison d'un modèle d'activité à un motif d'activité. Ainsi, en supposant un "motif d'activité normale" (p.ex. l'authentification à une messagerie Web suivie d'une phase de consultation des e-mails), une mesure de correspondance inférieure à un seuil d'acceptation équivaut à détecter une situation anormale : $anormal \equiv correspondance_{\{alignement|rejeu\}}(\textit{modèle}, \textit{motif}) < \eta$, avec η un paramètre de seuil. Dans ce cas, nous pouvons déclencher une alerte, sans être pour autant en mesure de fournir plus de détails sur la nature de l'anomalie. En pratique, nous considérons qu'il est nécessaire de tester par rapport à un ensemble de "motifs d'activité anormaux" complémentaires dans une deuxième phase appelée phase de "qualification" afin de catégoriser l'anomalie.

4 Experimentations et résultats

Dans cette section, nous détaillons les expériences menées sur la base des approches décrites précédemment (§3), et présentons les résultats associés. Tout d'abord, nous analysons la corrélation entre le volume de triplets RDF générés par Graphaméléon et l'objet de sites Web visités (§4.1). Ensuite, nous modélisons et identifions trois scénarios de navigation Web en utilisant Graphaméléon et des réseaux P/T au sein d'un environnement contrôlé (§4.2). Les expériences sont menées à l'aide de Graphaméléon v2.1.0. Les données associées à ces expériences sont disponible sur `https://github.com/Orange-`

5. `https://developer.mozilla.org/fr/docs/Web/API/Document_Object_Model`

OpenSource/graphameleon-ds.

4.1 Trafic réseau et complexité des sites Web

Dans cette première expérience, nous cherchons à comprendre dans quelle mesure le comportement d'un site Web varie lors d'une première connexion, et de fait génère des indicateurs significatifs pour créer une signature du site utilisable ultérieurement pour la détection d'anomalies. Pour cela, nous étudions la relation entre la complexité *a priori* d'un ensemble de sites Web et les ressources téléchargées, en termes de nombre et de taille. Nous étudions cette complexité en mesurant le nombre de triplets RDF générés par Graphaméléon lors de la connexion initiale. La Table 2 présente les mesures enregistrées.

À notre connaissance, il n'existe actuellement aucune étude décrivant des groupes (clusters) de complexité de sites Web bien connus, sauf d'un point de vue marketing [35] (p.ex. secteur d'activité *vs* nombre moyen de connexions à la page d'accueil du site, poids moyen de la page en octets, indice de vitesse de chargement). De plus, avec plus d'un milliard de sites Web référencés à ce jour [19], les outils d'analyse de sites Web proposent principalement des analyses de positionnement par rapport à la concurrence [15]. Cela souligne le défi de sélectionner des exemples représentatifs pour chaque groupe. Pour cette expérience, nous proposons d'établir un corpus de sites Web organisé selon trois groupes de complexité arbitraires. L'idée sous-jacente est que la complexité est liée au volume du contenu éditorial à afficher. Pour chaque catégorie, nous sélectionnons un sous-ensemble de trois sites Web de référence sur la base d'opinions d'experts tiers :

One-Page où "Swappa Bottle" [6], "Garden Studio" [7] et "Mark My Images" [8] (MMI) sont identifiés dans [27] comme les trois meilleurs exemples de sites Web d'une seule page dont s'inspirer dans le cadre de projets de conception de sites ;

Encyclopedia où "Encyclopedia Britannica Online" [9] (EBO), "Scholarpedia" [10] et "Encyclopedia.com" [11] sont présentés dans [20] comme les trois principales alternatives à Wikipédia du point de vue de la fiabilité de l'information ;

Content-Heavy où "RTI International" [12], "PrintMag" [13] et la "International Women's Media Foundation" [14] (IWMF) sont identifiés dans [22] comme les trois principaux sites Web présentant une grande quantité de contenu tout en offrant une expérience intuitive.

Ensuite, nous réalisons la collecte et l'analyse des données de traces de navigation pour chaque page d'accueil des sites Web de la manière suivante : *1)* dans une instance de Firefox sur ordinateur (anti-pistage $\in \{stricte, standard\}$), charger Graphaméléon et activer la capture de données (mode de collecte $\in \{micro, macro\}$, type de sortie $= semantize$) ; *2)* ouvrir un onglet de navigation et la console Network Monitor [15] (mise en cache = *désactivée*) ; *3)* accéder au site Web cible en saisissant son URL dans la barre de navigation ; *4)* arrêter la capture par Graphaméléon 10 secondes après la détection de l'événement de chargement complet de la page dans la console Network Monitor pour garantir l'exécution cohérente des scripts intégrés à la page Web (i.e. l'événement `DOMContentLoaded` [16]) ; *5)* enregistrer les données dans un fichier (sérialisation $= Turtle$) ; *6)* recueillir les statistiques de collecte de données (nombre de requêtes, nombre de réponses, nombre d'interactions, nombre de sommets, nombre d'arêtes) à partir de l'interface utilisateur de Graphaméléon, ainsi que celles du graphe de connaissances résultant grâce à un ensemble de requêtes SPARQL (nombre de triplets, nombre de sujets, nombre d'instances de classe).

Site Web	CM-Trk.	TC	SC	UDN	UHC	UIP	UURL
One-Page							
Swappa Bottle	μ-Str.	n.a.	-	-	-	-	-
	μ-Std.	n.a.	-	-	-	-	-
	M-Str.	n.a.	-	-	-	-	-
Garden Studio	μ-Str.	886	163	5	84	5	69
	μ-Std.	985	189	11	89	11	78
	M-Str.	21	5	1	1	1	1
MMI	μ-Str.	427	81	3	38	3	37
	μ-Std.	423	80	3	38	3	36
	M-Str.	21	5	1	1	1	1
Encyclopedia							
EBO	μ-Str.	599	122	13	54	13	42
	μ-Std.	2195	472	71	194	70	137
	M-Str.	21	5	1	1	1	1
Scholarpedia	μ-Str.	452	111	4	55	4	48
	μ-Std.	579	143	11	64	11	57
	M-Str.	n.a.	-	-	-	-	-
Encyclopedia	μ-Str.	350	66	2	31	2	31
	μ-Std.	1483	320	44	125	144	
	M-Str.	21	5	1	1	1	1
Content-Heavy							
RTI	μ-Str.	381	76	6	33	6	31
	μ-Std.	562	118	14	48	14	42
	M-Str.	21	5	1	1	1	1
PrintMag	μ-Str.	552	111	9	47	8	47
	μ-Std.	1143	234	25	101	24	84
	M-Str.	21	5	1	1	1	1
IWMF	μ-Str.	362	72	5	31	5	31
	μ-Std.	388	78	6	33	6	6
	M-Str.	21	5	1	1	1	1

TABLE 2 – Statistiques pour l'expérimentation "Trafic réseau et complexité des sites Web".
Statistiques basées sur les modes de collecte "micro" (CM = μ) et "macro" (CM = M), et en fonction de la politique de blocage des traqueurs du navigateur Web. Abréviations : CM = mode de collecte, Trk. = politique de blocage des traqueurs (strict *vs* standard), TC = nombre de triplets, SC = nombre de sujets, UOA = nombre d'entités *ucobs :DomainNameFacet*, UDN = nombre d'entités `ucobs:DomainNameFacet`, UHC = nombre d'entités `ucobs:HTTPConnectionFacet`, UIP = nombre d'entités `ucobs:IPAddressFacet`, UURL = nombre d'entités `ucobs:URLFacet`, n.a. = non applicable.

Résultats & discussion. En utilisant cette procédure, 27 échantillons de données ont été produits (trois groupes × trois sites × trois configurations du mode de collecte), dont 23 ont permis une analyse et quatre sont inexploitables (une

6. https://swappabottle.com/
7. https://gardenestudio.com.br/
8. https://www.markmyimages.com/
9. https://www.britannica.com/
10. http://www.scholarpedia.org/
11. https://www.encyclopedia.com/
12. https://www.rti.org/
13. https://www.printmag.com/
14. https://www.iwmf.org/
15. https://firefox-source-docs.mozilla.org/devtools-user/network_monitor/
16. https://developer.mozilla.org/en-US/docs/Web/API/Document/DOMContentLoaded_event

	Stricte		Standard		Std. / Str.	
	UHC	UIP	UHC	UIP	UHC	UIP
One-Page	61.0	4.0	63.5	7.0	1.04	1.8
Encyclopedia	46.7	6.3	127.7	41.7	**2.73**	**6.6**
Content-Heavy	37.0	6.3	60.7	14.7	1.64	2.3

TABLE 3 – Moyenne du nombre d'entités en mode micro.
Comparaison de la moyenne du nombre d'entités UHC et UIP à partir de la Table 2 en fonction du niveau de complexité et de la politique anti-pistage. Seules les valeurs "Garden Studio" et "MMI" sont prises en compte pour la catégorie "One-Page". Abréviations : UHC = nombre d'entités `ucobs:HTTPConnectionFacet`, UIP = nombre d'entités `ucobs:IPAddressFacet`.

erreur d'accès `SSL_ERROR_NO_CYPHER_OVERLAP` côté serveur pour "Swappa Bottle" en mode micro et macro, et une erreur de traitement indéterminée de l'extension Web pour "Scholarpedia" en mode macro). La Table 2 présente les statistiques relatives aux triplets RDF. Pour les échantillons issus du mode macro (CM = M), nous observons que les statistiques sur les triplets RDF restent cohérentes quel que soit le site visité. Une analyse des fichiers Turtle résultants révèle également que la structure de données RDF est conforme au modèle de données de la Figure 2. En ce qui concerne le mode micro (CM = μ), les mesures présentent une variabilité significative entre chaque catégorie de complexité pour une politique d'anti-pistage donnée. La Table 3 permet de préciser ce point en présentant le nombre moyen d'entités pour les classes d'objets `ucobs:HTTPConnectionFacet` (UHC) et `ucobs:IPAddressFacet` (UIP), ce pour chaque scénario. La comparaison des valeurs moyennes du nombre d'entités en fonction de la politique d'anti-pistage (colonne "Std. / Str." dans la Table 3) révèle une augmentation du nombre moyen de connexions et de serveurs distants vers lesquels une connexion a été faite lorsque les politiques sont assouplies, et ce quel que soit le niveau de complexité. De ces mesures, nous concluons à la fois sur le bon fonctionnement de Graphaméléon et sur sa pertinence pour l'étude des primo-connexions. Bien que les groupes de complexité proposés puissent être discutés en raison de la taille limitée de l'échantillon et de la variabilité du contenu des sites Web, l'augmentation des échanges réseau en fonction des politiques d'anti-pistage fournit une base pour de futurs travaux de catégorisation par les stratégies de suivi mises en œuvre par les sites Web (tracking & analytics) et de la topologie de réseau associée.

4.2 Catégorisation de traces de navigation

Dans cette deuxième expérience, notre objectif est de catégoriser les traces de navigation Web comme comportements normaux ou anormaux. Nous analysons les trois scénarios suivants en utilisant la modélisation de macro-activité (§3.1) et les réseaux de Petri (§3.3), puis rendons compte de la capacité à identifier une anomalie.

Scénario de base (normal) : un utilisateur accède au site Web, se connecte à son compte en utilisant son nom d'utilisateur et son mot de passe, navigue vers la page "Vendre un livre", renseigne une formulaire, puis retourne à la page d'accueil où il retrouve son livre dans la liste des ventes.

Scénario alternatif (normal alternatif) : un utilisateur accède au site Web, se connecte à son compte en utilisant un système à authentification unique (SSO), navigue vers la page "Vendre un livre", renseigne un formulaire, puis retourne à la page d'accueil où il retrouve son livre.

Scénario d'attaque XSS (anormal) : un attaquant accède au site Web, se connecte à son compte en utilisant son nom d'utilisateur et son mot de passe, navigue vers la page "Vendre un livre" et effectue une injection de code dans le champ "Auteur". Enfin, il retourne à la page d'accueil où le script injecté est exécuté.

Nous utilisons une simulation de site Web de librairie en ligne afin d'être en situation d'expérience contrôlée. Cela permet une exposition intentionnelle du site à diverses vulnérabilités de sécurité (une vulnérabilité XSS dans le cas présent, une forme courante d'attaque). Cela permet de même, avant l'étude, d'étiqueter chaque élément des pages Web du site, ce qui améliore l'interprétabilité des données collectées.

Nous réalisons la collecte et l'analyse des données de traces de navigation pour chaque scénario de la manière suivante : *1)* dans une instance de Firefox sur ordinateur, charger Graphaméléon et activer la capture de données (mode de collecte $= macro$, type de sortie $= semantize$) ; *2)* ouvrir un onglet de navigation et parcourir le site Web simulé selon le scénario de navigation ; *3)* arrêter la capture par Graphaméléon et enregistrer les données dans un fichier (sérialisation $= Turtle$) ; *4)* recueillir les statistiques de collecte de données (nombre de requêtes, nombre de réponses, nombre d'interactions, nombre de sommets, nombre d'arêtes) à partir de l'interface utilisateur de Graphaméléon ; *5)* calculer le modèle d'activité à partir de la trace enregistrée en utilisant la bibliothèque PM4PY Process Mining [2] (méthode $\in \{Inductive, Alpha, LogSkeleton, Heuristic, AlphaPlus\}$) ; *6)* calculer la correspondance du modèle d'activité au motif de référence en utilisant la bibliothèque PM4PY (méthode $\in \{TokenBasedReplay, Alignment\}$). Le scénario de base, qui correspond au comportement "normal", est utilisé comme motif d'activité (i.e. le modèle d'activité du scénario de base en tant que référence).

	Base	Alternatif	Attaque XSS
Requêtes	10	13	11
Interactions	18	14	18
Nœuds	263	283	277
Arcs	404	431	426

TABLE 4 – Statistiques pour l'expérimentation "catégorisation de traces de navigation".
Statistiques en termes du nombre de requêtes réseau, des interactions de l'utilisateur avec le navigateur Web, des nœuds et des arcs du graphe de navigation résultant, tel que rapporté par l'interface utilisateur de Graphaméléon pour les scénarios de navigation définis au §4.2.

Résultats & discussion. En utilisant cette procédure, trois échantillons de données ont été produits. La Table 4 présente les statistiques relatives aux graphes de navigation résultants, et la Table 5 compare les résultats de diffé-

		Alternatif	Attaque XSS
Token-Based	Alpha	0.886	0.968
	Alpha+	0.890	0.969
	Inductive	0.923	1.000
	Heuristic	0.923	1.000
Alignement	Alpha	-	-
	Alpha+	-	-
	Inductive	0.718	0.976
	Heuristic	0.718	0.976
Log Skeleton		0.684	0.999

TABLE 5 – Scores de correspondance au motif de référence.
Comparaison des scores de correspondance au motif de référence (modèle d'activité du scénario de base) pour les modèles d'activité des scénarios "alternatif" et "attaque XSS". Différentes techniques et algorithmes de vérification de conformité sont utilisés pour calculer les scores de correspondance. Les techniques "token-based" et "alignement" nécessitent une découverte préalable du modèle d'activité ; les algorithmes "Alpha/Alpha+", "Inductive" et "Heuristic Miner" sont utilisés pour cela. La technique "Log Skeleton" fournit directement les scores de correspondance en utilisant les traces d'activité.

rentes techniques d'évaluation de la correspondance au motif d'activité. Du point de vue des statistiques des graphes de navigation, nous observons que le scénario alternatif implique moins d'interactions mais plus de transactions réseau que pour le scénario de base. Cela correspond au fait que l'utilisateur n'a qu'un seul bouton à cliquer pour l'authentification, et que l'authentification est déléguée à diverses entités externes fournissant le service d'authentification. Pour le scénario "attaque XSS", le nombre d'interactions reste le même, mais le nombre de requêtes augmente d'une unité. Cela correspond à la séquence d'authentification identique à celle du scénario de base, mais avec une requête supplémentaire causée par l'injection SQL. Toujours pour ce même scénario, nous remarquons une légère variation dans les scores de correspondance (une correspondance moyenne de 98% au motif d'activité), ce qui correspond également à la requête supplémentaire causée par l'injection SQL. Nous observons en outre que cette requête supplémentaire est facilement identifiable par l'utilisation des techniques d'alignement de séquences sur les traces sémantisées, le modèle de données proposé en §3.1 et appliqué au niveau de l'extension Graphaméléon permettant en effet de standardiser l'interprétation des traces.

Par conséquent, bien que notre approche fournisse une représentation formelle des traces de navigation, nous observons que son utilisation directe n'est pas adaptée à la détection d'anomalies lorsque des micro-changements se produisent par rapport à un motif d'activité (i.e. lorsque les éléments de différenciation pour qualifier les écarts sont relativement rares dans la séquence). De même, nous remarquons que, bien que les algorithmes de découverte utilisent généralement plusieurs échantillons de traces pour générer un modèle généralisé de l'activité, nous avons dans notre cas considéré un motif parfait déduit d'une seule réalisation de trace. Or un modèle de comportement normal en situation réelle est potentiellement plus complexe. Pour exemple, lors de la saisie d'un formulaire, l'ordre de lecture conventionnel est généralement suivi. Cependant, en raison de biais cognitifs, un utilisateur pourrait le remplir dans un ordre différent tout en restant dans les limites d'un comportement normal réel.

Enfin, en prenant du recul sur la collecte de données et le traitement sémantique, nous remarquons une faible compression lexicale des données de trace de navigation en raison d'un formatage cohérent (p.ex l'URL de la requête est toujours située dans l'en-tête "url"). Cependant, cette compression concerne d'avantage la sémantique des interactions. En effet, l'un des défis de l'alignement des modèles d'activité réside dans le manque d'une méthode fiable pour identifier les éléments HTML (surtout en l'absence d'un ID explicite) à travers les navigateurs, les sessions et les utilisateurs. Ce défi devient apparent lorsque le DOM du contenu de la page change à chaque visite du site, en particulier lorsque des insertions publicitaires dynamiques se produisent.

5 Conclusions et travaux futurs

Dans ce travail, nous avons cherché des moyens d'analyser les traces d'activités de navigation sur le Web dans le but de caractériser les activités des utilisateurs et le comportement des infrastructures réseaux. Les domaines d'application types envisagés dans cette recherche sont la gestion d'incident concernant les systèmes de télécommunication, la cybersécurité, et l'ingénierie des infrastructure réseaux. Sur les bases du projet DynaGraph [23], nous avons émis l'hypothèse que les graphes de connaissances peuvent structurer de façon adéquate les données collectées sur un navigateur Web au cours de sessions de navigation, et ce dans l'idée d'une analyse avancée des traces de navigation au travers d'une modélisation sous forme de réseaux de Petri et l'utilisation des outils associés aux techniques d'analyse des processus.

Pour tester notre approche, nous avons développé les concepts de micro-activité et de macro-activité en rapport au vocabulaire UCO [40] pour la représentation sémantique des activités. Nous avons également mis au point l'outil Graphaméléon, une extension Web en open source disponible à l'adresse `https://github.com/Orange-OpenSource/graphameleon` permettant la collecte en direct de données au niveau du navigateur Web et la sémantisation des traces de navigation. Enfin, nous avons analysé des traces d'activité collectées via Graphaméléon selon un plan expérimental en deux parties. Nous avons montré, dans l'expérience d'analyse de trafic par famille de complexité des sites Web, que l'augmentation des volumes d'échanges est fonction des politiques d'anti-pistage et fournit une base de travail intéressante pour la catégorisation des sites selon les stratégies d'analyse d'audience employées et la topologie de réseau associée. Ensuite, avec l'expérience de catégorisation des traces de navigation, nous avons montré les limites de la technique de vérification de conformité pour la détection d'anomalies lorsque des micro-changements se produisent par rapport à un motif de référence. Nous avons également remarqué le défi que représente l'harmonisation des modèles d'activité en raison de l'absence d'une mé-

thode fiable pour identifier les éléments HTML au sein des navigateurs Web, notamment par comparaison entre sessions de navigation ou entre utilisateurs.

Sur la base de ces développements et résultats, nous envisageons des travaux futurs approfondissant les aspects de la cartographie du Web, de l'analyse du comportement du couple utilisateur-système, et de la détection d'anomalies. En ce qui concerne l'outil Graphaméléon, des aspects techniques spécifiques nécessitent des développements complémentaires, tels que la génération de graphe en flux, l'annotation des activités via l'interface utilisateur et la gestion simultanée de plusieurs sessions de navigation Web. En ce qui concerne l'analyse de conformité, trois options se présentent pour réduire la sensibilité de notre approche. La première consiste à partitionner le motif de référence en sous-motifs, ce qui devrait réduire l'amplitude de variation du score de correspondance en cas de non-conformité. La seconde consiste à utiliser des motifs spécifiques pour qualifier un groupe d'actions et localiser le groupe par alignement de séquence (p.ex. un motif décrivant une injection SQL plutôt qu'une description générale du comportement normal). La troisième approche consiste à pondérer l'importance des actions dans le calcul du score de correspondance activité/motif en utilisant le graphe de connaissances pour fournir du contexte (p.ex. une adresse IP source peu fréquente lors d'une attaque par injection SQL, un saut réseau impossible, un même utilisateur connecté depuis deux pays) ; l'idée étant d'utiliser une pondération qui masquerait les variations mineures dues au "bruit" par rapport aux variations causées par des erreurs réelles. Enfin, nous envisageons d'intégrer les modèles d'activité – via des vocabulaires appropriés pour les réseaux de Petri [11, 18] – dans un graphe RDF structuré par l'ontologie NORIA-O [25], ce afin de calculer des contextes d'anomalies enrichis par un processus de décision en utilisant la technique de plongement de graphes [24]. Nous analyserons notamment en quoi les modèles d'activité renforcent l'aide à la décision (p.ex. performance de la détection, interprétabilité) dans une situation de gestion d'incident avec connaissance partielle de l'activité des utilisateurs, comme cela peut être le cas lorsque l'analyse est menée côté réseaux/serveurs.

Références

[1] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, José Emilio Labra Gayo, Sabrina Kirrane, Sebastian Neumaier, Axel Polleres, Roberto Navigli, Axel-Cyrille Ngonga Ngomo, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge Graphs, 2020.

[2] Alessandro Berti, Sebastiaan van Zelst, and Wil van der Aalst. Process Mining for Python (pm4py) : Bridging the Gap between Process-and Data Science. In *Proceedings of the ICPM Demo Track 2019, co-located with 1st International Conference on Process Mining (ICPM 2019)*, 2019.

[3] Amélie Cordier, Marie Lefevre, Pierre-Antoine Champin, Olivier Georgeon, and Alain Mille. Trace-Based Reasoning - Modeling Interaction Traces for Reasoning on Experiences. In *The 26th International FLAIRS Conference*, 2013.

[4] Amina Annane, Nathalie Aussenac-Gilles, and Mouna Kamel. BBO : BPMN 2.0 Based Ontology for Business Process Representation. In *20th European Conference on Knowledge Management (ECKM)*, Lisbon, Portugal, 2019.

[5] Anastasia Dimou, Miel Vander Sande, Pieter Colpaert, Ruben Verborgh, Erik Mannens, and Rik Van de Walle. RML : A Generic Language for Integrated RDF Mappings of Heterogeneous Data. In *Proceedings of the Workshop on Linked Data on the Web, LDOW 2014, co-located with the 23rd International World Wide Web Conference (WWW 2014)*. CEUR-WS.org, 2014.

[6] Bernard Berthomieu, Pierre-Olivier Ribet, and François Vernadat. The tool tina – construction of abstract state spaces for petri nets and time petri nets. *International Journal of Production Research*, 2004.

[7] Caleb C. Noble and Diane J. Cook. Graph-based anomaly detection. 2003.

[8] Maria Carla Calzarossa and Luisa Massari. Analysis of header usage patterns of http request messages. In *IEEE International Conference on High Performance Computing and Communication*, 2014.

[9] Giovanna Castellano, Anna M. Fanelli, and Maria A. Torsello. *Web Usage Mining : Discovering Usage Patterns for Web Applications*. Springer Berlin Heidelberg, 2013.

[10] Pierre Dagnely, Tom Ruette, Tom Tourwé, and Elena Tsiporkova. Ontology-driven multilevel sequential pattern mining : mining for gold in event logs of photovoltaic plants. In *2018 International Conference on Intelligent Systems (IS)*, 2018.

[11] Dragan Gašević and Vladan Devedžić. Petri net ontology. *Knowledge-Based Systems*, 2006.

[12] Franck Ghitalla, Dominique Boullier, and Mathieu Jacomy. *Qu'est-Ce Que La Cartographie Du Web ?: Expéditions Scientifiques Dans l'univers Des Données Numériques et Des Réseaux*. 2021.

[13] Iman Akbari, Mohammad A. Salahuddin, Leni Ven, Noura Limam, Raouf Boutaba, Bertrand Mathieu, Stephanie Moteau, and Stéphane Tuffin. Traffic classification in an increasingly encrypted web. *Communications of the ACM*, 2022.

[14] Ines Chami, Sami Abu-El-Haija, Bryan Perozzi, Christopher Ré, and Kevin Murphy. Machine Learning on Graphs : A Model and Comprehensive Taxonomy, 2020.

[15] James Parsons. Alexa.com is dead – here are 20 of the best alternatives. `https://www.contentpowered.com/blog/alexa-com-`

dead-alternatives/, 2023. Accessed : 2023-08-10.

[16] Johannes Koch, Carlos A. Velasco, and Philip Ackermann. Http vocabulary in rdf 1.0. W3c working group note, W3C, 2017.

[17] Jorge Munoz-Gama. *Conformance Checking and Diagnosis in Process Mining : Comparing Observed and Modeled Processes*. PhD thesis, Universitat Politècnica de Catalunya – BarcelonaTech, Barcelona, 2014.

[18] Juan C. Vidal, Manuel Lama, and Alberto Bugarın. A High-level Petri Net Ontology Compatible with PNML. 2006.

[19] Kathy Haan. Top website statistics for 2023. https://www.forbes.com/advisor/business/software/website-statistics/, 2023. Accessed : 2023-08-10.

[20] Kent Campbell. Seven free wikipedia alternatives. https://blog.reputationx.com/wikipedia-alternatives, 2023. Accessed : 2023-08-10.

[21] Pierre Laperdrix, Nataliia Bielova, Benoit Baudry, and Gildas Avoine. Browser fingerprinting : A survey, 2019.

[22] Laura Held. Examples of content heavy editorial website designs. https://www.newmediacampaigns.com/blog/best-examples-of-content-heavy-editorial-website-designs, 2021. Accessed : 2023-08-10.

[23] Lionel Tailhardat, Raphaël Troncy, and Yoan Chabot. Walks in cyberspace : Towards better web browsing and network activity analysis with 3d live graph rendering. Association for Computing Machinery, 2022.

[24] Lionel Tailhardat, Raphael Troncy, and Yoan Chabot. Leveraging knowledge graphs for classifying incident situations in ict systems. In *18th International Conference on Availability, Reliability and Security (ARES)*, 2023.

[25] Lionel Tailhardat, Yoan Chabot, and Raphaël Troncy. NORIA-O : an Ontology for Anomaly Detection and Incident Management in ICT Systems. In *Semantic Web – 21st International Conference, ESWC 2024, Hersonissos, Crete, Greece, May 26 - 30, 2024, Proceedings*, 2024.

[26] Vítor Santos Lopes and João Mendes-Moreira. A comparative analysis of data preprocessing techniques in web usage mining. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 2019.

[27] Madhu Murali. 11 examples of one-page websites to inspire you. https://blog.hubspot.com/website/11-examples-of-one-page-websites-for-inspiration, 2023. Accessed : 2023-08-10.

[28] Mathieu Lirzin and Béatrice Markhoff. Vers Une Ontologie Des Interactions HTTP. In *31emes Journées Francophones d'Ingénierie Des Connaissances*, Angers, France, 2020.

[29] Megan Katsumi and Mark Fox. iCity Transportation Planning Suite of Ontologies. Technical report, University of Toronto, 2020.

[30] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection. *ACM Computing Surveys*, 2021.

[31] Heeryon Park and Doo-Kwon Baik. Web log session identification based on cluster-based classification. In *7th International Conference on Advanced Information Networking and Applications (AINA)*. IEEE, 2011.

[32] Stephen Pauwels and Toon Calders. Extending dynamic bayesian networks for anomaly detection in complex logs, 2018.

[33] Peter E. Kaloroumakis and Michael J. Smith. Toward a Knowledge Graph of Cybersecurity Countermeasures. Technical report, The MITRE Corporation, 2021.

[34] Sasan Saqaeeyan, Hamid Haj Seyyed Javadi, and Hossein Amirkhani. Anomaly detection in smart homes using bayesian networks. *KSII Transactions on Internet and Information Systems*, 2020.

[35] thinkwithgoogle.com. Find out how you stack up to new industry benchmarks for mobile page speed. https://think.storage.googleapis.com/docs/mobile-page-speed-new-industry-benchmarks.pdf, 2017. Accessed : 2023-08-10.

[36] W3C. Fetch metadata request headers. Working draft, W3C, July 2021.

[37] Xindong Wu, Xingquan Zhu, Gongqing Wu, and Wei Ding. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 2004.

[38] Nong Ye. A markov chain model of temporal behavior for anomaly detection. 2000.

[39] Youssra Rebboud, Pasquale Lisena, and Raphael Troncy. Beyond Causality : Representing Event Relations in Knowledge Graphs. In *Knowledge Engineering and Knowledge Management*. Springer International, 2022.

[40] Zareen Syed, Ankur Padia, M. Lisa Mathews, Tim Finin, and Anupam Joshi. UCO : A Unified Cybersecurity Ontology. In *AAAI Workshop on Artificial Intelligence for Cyber Security*. AAAI Press, 2016.

[41] Yan Zhao, Liwei Deng, Xuanhao Chen, Chenjuan Guo, Bin Yang, Tung Kieu, Feiteng Huang, Torben Bach Pedersen, Kai Zheng, and Christian S. Jensen. A comparative study on unsupervised anomaly detection for time series : Experiments and analysis, 2022.

Activité AFIA

1er août 2022 – 31 juillet 2024

Éditeurs : Conseil d'Administration de l'AFIA – Année 2024

Tables des matières

AfIA
Association française
pour l'Intelligence Artificielle

À propos de l'AFIA

L'objet de l'AFIA, Association Loi 1901 sans but lucratif, est de promouvoir et de favoriser le développement de l'Intelligence Artificielle (IA) sous ses différentes formes, de regrouper et de faire croître la communauté française en IA et, à la hauteur des forces de ses membres, d'en assurer la visibilité.

L'AFIA anime la communauté par l'organisation de grands rendez-vous. Se tient ainsi chaque été une semaine de l'IA, la Plate-forme IA (PFIA 2022 à Saint-Étienne, PFIA 2023 à Strasbourg, PFIA 2024 à La Rochelle) au sein de laquelle se tiennent la Conférence Nationale d'Intelligence Artificielle (CNIA), les Rencontres des Jeunes Chercheurs en IA (RJCIA) et la Conférence sur les Applications Pratiques de l'IA (APIA) ainsi que des conférences/journées thématiques hébergées qui évoluent d'une année à l'autre, sans récurrence obligée.

Ainsi, PFIA 2024 a hébergé du 1 au 5 juillet 2024 à La Rochelle, outre la 27e CNIA, les 22es RJCIA et la 10e APIA : les 2 conférences IC et JIAF, 3 journées thématiques (Agents & IA, Santé & IA, Société & IA), 7 ateliers thématiques (Défense & IA, Jeux & IA, MAFTEC, SOSEM, CÉCILIA, IA en Nouvelle-Aquitaine, GDR RADIA) et 8 tutoriels hébergés.

Forte du soutien de ses 427 adhérents à jour de leur cotisation en juillet 2024, l'AFIA assure :

- le maintien d'un site Web dédié à l'IA reproduisant également les Brèves de l'IA ;
- une *journée industrielle* « Forum Industriel en IA » (FIIA 2023) ;
- une *journée recherche* « Perspectives et Défis en IA » (PDIA 2024) ;
- une *journée enseignement* « Enseignement et Formation en IA » (EFIA 2023) ;
- une « École Saisonnière en IA » (ESIA2023, édition 2025 en préparation) ;
- la remise annuelle d'un prix de thèse en IA ;
- le soutien à 8 collèges ayant leur propre activité :
  - collège Industriel (janvier 2016),
  - collège Apprentissage Artificiel (janvier 2020),
  - collège Interaction avec l'Humain (juillet 2020),
  - collège Représentation et Raisonnement (avril 2017),
  - collège Science de l'Ingénierie des Connaissances (avril 2016),
  - collège Systèmes Multi-Agents et Agents Autonomes (janvier 2017),
  - collège Technologies du Langage Humain (juillet 2019),
  - collège Création d'Événements Collaboratifs, Inclusifs et Ludiques en IA (octobre 2021) ;
- la parution trimestrielle des Bulletins de l'AFIA ;
- un lien entre ses membres et sympathisants sur les réseaux sociaux LinkedIn, Facebook et Twitter ;
- le parrainage scientifique, mais aussi éventuellement financier, d'événements en IA ;
- la diffusion mensuelle de Brèves sur les actualités de l'IA en France (abonnement ou envoi à la liste) ;
- la réponse aux consultations officielles ou officieuses (Ministères, Missions, Organismes) ;
- la réponse aux questions de la presse, écrite ou orale, également sur internet ;
- la divulgation d'offres de collaborations, de formations, d'emploi, de thèses et de stages.

L'AFIA organise aussi des journées communes avec d'autres associations. Pour 2024 : EGC & IA avec EGC et Groupe de Travail GAST; Humanité numérique & IA avec GDR MADICS et MAGIS.

Enfin, l'AFIA encourage la participation de ses membres aux grands événements de l'IA, dont PFIA. Ainsi, les membres de l'AFIA, pour leur inscription à PfIA, bénéficient d'une réduction équivalente à deux fois le coût de leur adhésion, leur permettant d'assister à PFIA 2024 sur 5 jours au tarif de 123 € TTC !

Rejoignez-nous vous aussi et adhérez à l'AFIA pour contribuer au développement de l'IA en France. L'adhésion peut être individuelle ou au titre de personne morale. Merci également de susciter de telles adhésions en diffusant ce document autour de vous !

AfIA
Association française
pour l'Intelligence Artificielle

Conseil d'Administration 2023

Président	Benoit LE BLANC	GT Bureau, Prix Thèse, International, Représentation
V.-Présidente	Domitile LOURDEAUX	GT Bureau, GT Collèges, Parrainages, Représentation
Trésorière	Isabelle SESÉ	GT Bureau, Adhésions, Partenariats
Secrétaire	Grégory BONNET	GT Rédaction, Prix Thèse
Porte-parole	Emmanuel ADAM	GT Communication, Brèves, Réseaux
Rédacteur	Dominique LONGIN	GT Rédaction, GT Communication, GT Plateforme
Webmestre	Catherine ROUSSEY	GT Web & Mail, GT Communication, Brèves, Réseaux
	Gayo DIALLO	Collège Industriel
	Gaël DIAS	GT Enseignement, École saisonnière
	Bernard GEORGES	Collège Industriel
	Thomas GUYET	GT Plateforme, PFIA 2023
	Frédéric MARIS	GT Enseignement, Adhésions
	Davy MONTICOLO	GT Plateforme, PFIA 2023, PDIA 2023
	Gauthier PICARD	GT Enseignement, EFIA 2022
	Valérie REINER	Collège Industriel, École saisonnière, Représentation
	Céline ROUVEIROL	Collège Industriel, Cartographie
	Fatiha SAIS	GT Journées, PDIA 2022
	Ahmed SAMET	GT Plateforme, Tutoriels PFIA 2022

Conseil d'Administration 2024

Président	Benoit LE BLANC	GT Bureau, Prix Thèse, International, Représentation
V.-Président	Thomas GUYET	GT Bureau, GT Collèges, GT Plateforme, Représentation
Trésorière	Isabelle SESÉ	GT Bureau, Adhésions, Partenariats
Secrétaire	Grégory BONNET	GT Bureau, GT Rédaction, Adhésions, Prix Thèse
Porte-parole	Emmanuel ADAM	GT Communication, Brèves, Réseaux
Rédacteur	Dominique LONGIN	GT Rédaction, Collections HAL
Webmestre	Catherine ROUSSEY	GT Web & Mail, Réseaux
	Azzedine BENABBOU	Invité, GT Web & Mail
	Zied BOURAOUI	GT Collèges
	Gayo DIALLO	Collège Industriel
	Bernard GEORGES	Collège Industriel
	Domitile LOURDEAUX	GT Collèges
	Frédéric MARIS	GT Enseignement, ESIA
	Davy MONTICOLO	GT Plateforme, GT Journées
	Jose MORENO	Invité, GT Plateforme
	Gauthier PICARD	GT Enseignement
	Valérie REINER	Collège Industriel
	Céline ROUVEIROL	Collège Industriel
	Fatiha SAIS	GT Journées
	Ahmed SAMET	GT Plateforme, GT Journées

AfIA
Association française
pour l'Intelligence Artificielle

Collège Apprentissage Artificiel

Objectif du collège

L'objectif du Collège Apprentissage Artificiel (C2A) de l'AFIA est de contribuer à l'animation de la communauté de recherche française en apprentissage automatique (ou artificiel), et ce en synergie avec les structures d'animation déjà existantes. Les thématiques de recherche sont celles d'apprentissage artificiel, adossées aux principales conférences de la communauté que sont : Conférence d'Apprentissage Artificiel (CAp), Journées Francophones sur les Réseaux Bayésiens et les Modèles Graphiques Probabilistes (JFRB), Rencontres de la Société Francophone de Classification (SFC), Reconnaissance de Formes, Image, Apprentissage et Perception (RFIAP), etc. et concerne principalement l'ensemble des travaux autour de l'étude, la conception et l'évaluation d'algorithmes d'apprentissage artificiel, sous ses différentes formes : supervisé, non supervisé ou semi-supervisé ; statistique ou symbolique ; par renforcement ; par transfert.

Plus précisément, en lien avec le CA de l'AFIA, le collège a pour buts l'animation de la communauté autour de l'apprentissage artificiel (parrainage d'événements, organisation de journées bilatérales ou thématiques), et la représentation de la communauté en apprentissage artificiel au sein de l'AFIA (communications sur le thème de l'apprentissage artificiel, participation aux comités de programme).

Programme de travail

Le Collège Apprentissage Artificiel sera impliqué dans les activités suivantes :

- organisation de journées communes, par exemple :
  - Cla&IA : Classification et IA, en lien avec la SFC ;
  - Stats&IA : Statistiques et IA, en lien avec la SFdS ;
  - RF&IA : Reconnaissance de formes et IA (partie Apprentissage), en lien avec l'AFRIF ;
- organisation de journées thématiques, en particulier :
  - atelier sur la plateforme de l'AFIA ;
  - atelier en lien avec une conférence spécialisée (CAp, JFRB, SFC, RFIAP) ;
  - journée industrielle en lien avec le Collège Industriel de l'AFIA ;
- interaction avec des conférences (hors AFIA) pour lesquelles l'apprentissage est un thème central, parmi lesquels : CAp, JFRB, SFC, RFIAP (partie Apprentissage) ;
- interaction avec les autres associations existantes d'animation de la communauté de l'apprentissage artificiel :
  - Société Savante Francophone sur l'Apprentissage Machine (SSFAM) ;
  - Société Francophone de Classification (SFC) ;
  - Société Française de Statistique (SFdS).

Comité de pilotage

Le comité de pilotage est composé des personnes suivantes :

- Jérôme AZÉ, Université de Montpellier, LIRMM ;

— Isabelle BLOCH, Télécom Paris, LTCI ;
— Antoine CORNUÉJOLS, AgroParisTech, MMIP ;
— Elisa FROMONT, Université Rennes 1, IRISA ;
— Charlotte LACLAU, Université Jean Monnet à Saint-Etienne, LaHC ;
— Engelbert Mephu NGUIFO, Université Clermont Auvergne, LIMOS ;
— Amedeo NAPOLI, CNRS, LORIA ;
— Philippe PREUX, Université Lille 3, CRIStAL ;
— Céline ROUVEIROL, Université Paris 13, LIPN
— Christel VRAIN, Université d'Orléans, LIFO.

Le comité de pilotage peut être amené à inviter des membres de la communauté à participer aux discussions et réunions du collège.

Contacts

Coordinateur du collège : engelbert.mephu_nguifo@uca.fr.

Listes de diffusion : info-ic@inria.fr.

Deux membres du comité de pilotage sont membres du conseil d'administration de l'AFIA :

— Engelbert Mephu NGUIFO, engelbert.mephu_nguifo@uca.fr ;
— Céline ROUVEIROL, celine.rouveirol@lipn.univ-paris13.fr.

AfIA
Association française
pour l'Intelligence Artificielle

Collège Création d'Événements Collaboratifs, Inclusifs et Ludiques en IA

Objectif du collège

Le Collège Création d'Événements Collaboratifs, Inclusifs et Ludiques en IA (CECILIA) de l'AFIA défend l'apprentissage de l'IA grâce à la pratique coopérative et l'expérimentation. Il a deux objectifs :

- mettre à disposition des ressources pour l'Intelligence Artificielle par la pratique, en particulier pour les étudiant(e)s/lycéen(ne)s (par exemple lors de la Nuit de l'info) ;
- favoriser les rencontres dans la communauté IA au travers d'événements instructifs, ludiques et conviviaux (par exemple lors de PFIA).

Ce collège a repris les activités du Collège Compétition 2018-2020, pour les étendre à d'autres activités comme des ateliers inspirés des *game jams* pour être plus inclusives en termes de public et d'approches scientifiques. Ainsi le collège CECILIA à organisé un atelier « Jam Création de textes poétiques ou drôles » lors de PFIA'2022 et un atelier « Jam DriveToGæther » lors de PFIA'2023.

Pour mémoire, le Collège Compétition a organisé pour l'AFIA les compétitions et animations : « IA sur Robot » lors de RFIA'16 à Clermont-Ferrand, et « BotContest » lors de PFIA'17 à Caen, « DriveToGæther » lors de PFIA'19.

Programme de travail

À ce jour, les événements identifiées par le collège sont les suivants :

- **Nuit de l'Info.** Les participantes et les participants à la Nuit de l'Info répondent en une nuit sur un défi national de programmation, tout en relevant divers défis connexes. Leur programme et un document doivent être remis le lendemain matin avant 8h. Ces résultats sont évalués par un jury national et les jurys des défis connexes. Cette compétition a lieu tous les ans en décembre et l'AFIA y participe en y proposant un défi : intégrer de l'IA dans l'application développée. Le jury de ce défi se compose de membres du collège et d'autres personnes de l'AFIA.
- **IA sur Robots.** Le but de ce défi est de mettre en évidence l'IA au sein d'une ou plusieurs plateformes robotiques, dans un scénario figé avec des règles, et une mesure des performances. Ce défi est ouvert à toutes et tous : personnels travaillant dans la recherche, l'enseignement, l'ingénierie, étudiantes et étudiants ainsi que le grand public. Ce défi s'intitule « DriveToGæther » et dispose d'un scénario et d'un règlement. Les projets sélectionnés seront testés par le public (par exemple lors de PFIA). Un de nos souhaits est de formaliser le règlement et les supports de cet événement pour le rendre facilement organisable par des tierces organisations locales (lycées, clubs, etc.).
- **IA et Créativité.** Le but de cet événement est de faire expérimenter au public le potentiel créatif, l'originalité, la performance stratégique et/ou l'adaptabilité des méthodes d'IA. Il se déroule sur une ou plusieurs soirées (par exemple lors de PFIA), dans un cadre coopératif et festif.
- Un rassemblement annuel des membres du collège pour revenir sur les activités organisées (par exemple lors de PFIA).

Comité de pilotage

Le comité de pilotage se réunit au moins mensuellement et plus avant les événements.

— Anne-Gwenn BOSSER, École Nationale d'Ingénieurs de Brest, Lab-STICC ;
— Florence DUPIN DE SAINT-CYR, Université Toulouse 3 Paul Sabatier, IRIT ;
— Liana ERMAKOVA, Université de Bretagne Occidentale, Brest, HCTI ;
— Thomas GUYET, Inria, Lyon ;
— Philippe MORIGNOT, chercheur indépendant, Paris
— Nicolas PÉPIN-HERMANN, BA Healthcare, Rennes.

Un comité d'organisation est déterminé en fonction des événements, et intègre des membres du comité de pilotage et des membres de comité d'organisation de l'événement. Nous projetons de lancer un appel à participation pour constituer un comité consultatif afin de fédérer une communauté d'actrices et d'acteurs intéressés par l'IA pour tous (responsables d'enseignement ou de FabLab, etc.).

Contacts

Coordinateur du collège : Florence DUPIN DE SAINT-CYR

Adresse de contact : collegececilia-ca@googlegroups.com

Un membre du comité de pilotage est membre du conseil d'administration de l'AFIA :

— Thomas GUYET, thomas.guyet@afia.asso.fr.

Références

— Florence Dupin de Saint-Cyr, Nicolas Yannick Pepin, Nassim Mokhtari, Philippe Morignot, Julien Vianey, Anne-Gwenn Bosser, Liana Ermakova (2024) *DriveToGaether: a Turnkey Collaborative Robotic Event Platform*, 16th Int. Conf. on Agents and Artificial Intelligence (ICAART 2024), SciTePress, pages 404-411, doi : 10.5220/0012463800003636.

— Anne-Gwenn Bosser, Liana Ermakova, Florence Dupin de Saint-Cyr, Pierre De Loor, Victor Charpenay, Nicolas Yannick Pépin, Benoît Alcaraz, Jean-Victor Autran, Alexandre Devillers, Juliette Grosset, Aymeric Hénard, Florian Marchal (2022) *Poetic or Humorous Text Generation: Jam Event at PFIA2022* 13th Conf. and Labs of the Evaluation Forum (CLEF 2022), CEUR Workshop Proceedings, CEUR-WS.org, vol 3180, pages 1719-1726.

AfIA
Association française
pour l'Intelligence Artificielle

Collège Industriel

Objectif du collège

L'objet du Collège Industriel (CI) de l'AFIA est de favoriser les échanges en France dans le domaine de l'IA entre sa composante industrielle et sa composante académique ainsi que diverses actions de promotion de l'IA. Le rejoindre c'est, pour une société, en plus des bénéfices accordés à toutes les personnes morales de l'AFIA (pointeur vers site Web adhésions) :

- accroître la visibilité du CI de l'AFIA ;
- pouvoir faire état de ses relations académiques et leurs recherches de partenariat académique sur des problématiques ciblées ;
- proposer aux collèges thématiques des actions intéressant les membres du CI ;
- pouvoir discuter avec les autres sociétés et adhérents au CI de problématiques dans le domaine de l'IA, et partager des solutions en garantissant la confidentialité des échanges ;
- promouvoir l'IA auprès des décideurs et dirigeants industriels ;
- contribuer à équilibrer tous les domaines de l'IA et leurs hybridations ;
- témoigner auprès des collèges thématiques de cas d'usage qui intéressent le CI. ;
- témoigner de l'apport de l'IA dans l'industrie lors d'événements AFIA (FIIA, tutoriels). ;
- faire du lobbying au niveau français auprès des Ministères, des pôles de compétitivité, de l'ANR et tout autre organisme, également au niveau européen ;
- faire connaître aux académiques ses besoins en recrutement ;
- offrir des opportunités à la communauté académique de valoriser leurs formations.

En outre, les sociétés membres du CI à jour de leur cotisation mensuelle au printemps apparaissent comme partenaires de la Plateforme Intelligence Artificielle de la même année.

Programme de travail

En délégation du CA de l'AFIA, le programme de travail du CI consiste à :

- contribuer au pilotage d'événements annuels à forte visibilité, le Forum Industriel de l'IA (FIIA) et la Conférence sur les Applications Pratiques de l'IA (APIA) ;
- cartographier les relations académiques et industrielles (services du 1er Ministre, MA, MC, MEAE, MESRI, MI, MINEF, MJ, MS, MTES, ALLISTENE, CNRS, IMT, INRAE, INRIA, ONERA + IRT) ;
- solliciter les collèges thématiques de l'AFIA pour des contacts ou des interventions ;
- solliciter les collèges thématiques de l'AFIA pour des partenariats de projet ;
- organiser des réunions régulières au sein du collège ;
- organiser des réunions avec invités externes ;
- coprogrammer le prochain *AI Summit France* ;
- diffuser des bulletins ou des dossiers du collège en français avec résumés en anglais ;
- motiver les facilités d'accès à toutes les approches d'IA et leurs hybridations dans les formations, open-sourcer.

Ce programme est complété en début d'année civile par les membres du collège. Le CI se réunit mensuellement pour coordonner les avancées sur les actions engagées par le collège et en décider d'autres, échanger sur un sujet particulier et/ou sur l'actualité en IA sur le mois écoulé. Les réunions du CI font l'objet de comptes rendus qui distribués à ses membres et au CA de l'AFIA.

Comité de pilotage

Le CI est composé de l'ensemble des sociétés s'étant acquittées des droits d'adhésion pour l'année en cours et d'au moins deux académiques membres du CA de l'AFIA. Son comité de pilotage est constitué de dix personnes physiques, dont au moins : le coordinateur du CI, le responsable de la feuille de route et le responsable des séminaires. Ces responsables sont désignés pour une durée d'une année par les membres du CI lors de la première réunion annuelle. En 2022, le Collège Industriel est composé de :

- Bruno CARRON et Frédéric PERLANT, AIRBUS, Elancourt ;
- Alain BERGER, ARDANS, Montigny-le-Bretonneux ;
- Mustapha DERRAS, Youssef MILOUDI et Valérie REINER, BERGER-LEVRAULT, Boulogne Billancourt ;
- Stéphane DURAND et Bruno PATIN, DASSAULT Aviation, Saint-Cloud ;
- Pierre FEILLET et Christian DE SAINTE-MARIE, IBM FRANCE, Gentilly ;
- Ghislain ATEMEZING et Christophe PRIGENT, MONDECA, Paris ;
- Jean-Pierre DESMOULINS, Jean-Baptiste FANTUN et Véronique VENTOS, NUKKAI, Paris ;
- Julien BOHNE, Bernard GEORGES et Christelle LAUNOIS, SOCIETE GENERALE, Val de Fontenay ;
- Patricia BESSON, Juliette MATTIOLI et David SADEK, THALES, Palaiseau ;
- Yves DEMAZEAU et Céline ROUVEIROL, CA AFIA, Grenoble et Paris.

Contacts

Coordinateur et Responsable de la feuille de route : yves.demazeau@afia.asso.fr.

Responsable des séminaires : valerie.reiner@berger-levrault.com.

Quatre membres du comité de pilotage sont membres du conseil d'administration de l'AFIA :

- Yves DEMAZEAU, yves.demazeau@afia.asso.fr ;
- Bernard GEORGES, bernard.georges.777@gmail.com ;
- Valérie REINER, valerie.reiner@berger-levrault.com ;
- Céline ROUVEIROL, celine.rouverirol@afia.asso.fr.

AfIA
Association française
pour l'Intelligence Artificielle

Collège Interaction avec l'Humain

Objectif du collège

Le collège Interaction avec l'Humain (IH) de l'AFIA a pour mission de contribuer aux activités menées par l'AFIA, par des actions relatives au domaine de l'Intelligence Artificielle (IA), en lien avec les domaines de l'Interaction Homme-Machine (IHM), des Environnements Informatiques pour l'Apprentissage Humain (EIAH), de la Narration Interactive (NI) et des Environnements Virtuels Interactifs (EVI, incluant la réalité virtuelle, la réalité augmentée, ou encore la réalité mixte).

Les thématiques de recherche couvertes par ce collège sont celles relevant de :

- l'ingénierie et la modélisation des connaissances :
  - connaissances des utilisateurs d'environnement numérique : les apprenants dans les EIAH (tuteurs intelligents, jeux sérieux, MOOC, etc.), les utilisateurs d'une IHM ou d'un EVI, les lecteurs de récits élaborés en NI ;
  - connaissances nécessaires au bon fonctionnement de l'environnement numérique : les connaissances des domaines d'enseignements, les connaissances liées à la cognition humaine, les connaissances liées aux environnements virtuels ;
  - connaissances nécessaires à l'interaction entre les humains et leur environnement numérique ;
- les algorithmes d'apprentissage automatique et semi-supervisés utilisés notamment pour la fouille des données d'interaction : par exemple, l'*Educational Data Mining* et les *Learning Analytics* dans le cadre des EIAH ;
- les systèmes de diagnostic et de prise de décisions pour adapter l'environnement aux besoins, capacités et préférences de leurs utilisateurs ;
- les modèles de collaboration au sein des environnements complexes où l'humain a une place prépondérante ;
- l'Intelligence Artificielle pour soutenir l'interaction 3D informée en Environnements Immersifs ;
- les approches génératives pour la personnalisation de scénarios en environnements virtuels et narration interactive.

Ces thématiques seront coordonnées avec celles couvertes par les autres collèges, notamment pour les parties qui se trouvent aux intersections, comme par exemple la prise en compte de l'humain dans les systèmes à base de connaissances, également couverte par le collège Science de l'Ingénierie des Connaissances.

Programme de travail

Les missions du collège IH concernent le soutien à l'organisation de manifestations scientifiques (conférences, ateliers), l'animation de groupes de travail, l'édition de dossiers techniques ou de numéros spéciaux de journaux sur des thématiques d'intérêt pour la communauté, ainsi que la diffusion et la communication autour des recherches menées par les communautés françaises sur les disciplines ciblées.

Ce collège contribue aux actions initiées par l'AFIA sur la mise en place de journées bilatérales, notamment les journées communes Environnements Informatiques pour l'Apprentissage Humain et

Intelligence Artificielle (EIAH & IA), Interaction Homme-Machine et Intelligence Artificielle (IHM & IA), et Réalité Virtuelle et Intelligence Artificielle (RV & IA). Il proposera également un soutien similaire à d'autres événements, ponctuels ou récurrents, relevant de son périmètre scientifique.

Il contribue également au Bulletin de l'AFIA en proposant notamment des dossiers sur l'état des lieux des recherches combinant IHM et IA, comme cela vient d'être fait pour la thématique EIAH & IA.

Les actualités du collège IH, et plus largement des communautés françaises de recherche associées aux journées, sont publiées régulièrement sur un site web dédié au collège.

Comité de pilotage

Le comité de pilotage du collège IH est constitué de chercheurs spécialisés dans les domaines des IHM, des EIAH et de la RV avec une approche IA :

- Armelle BRUN, Université de Lorraine, LORIA, Nancy ;
- Pierre CHEVAILLIER, ENI Brest, CERV, Brest ;
- Nadine COUTURE, ESTIA, LABRI, Bordeaux ;
- Catherine FARON, Université Côte d'Azur, I3S, Sophia Antipolis ;
- Benoit LE BLANC, ENSC Bordeaux INP, IMS, Bordeaux ;
- Marie LEFEVRE, Université Lyon 1, LIRIS, Lyon ;
- Domitile LOURDEAUX, Université de Technologie de Compiègne, HEUDIASYC, Compiègne ;
- Vanda LUENGO, Sorbonne Université, LIP6, Paris ;
- Nicolas SABOURET, Université Paris-Saclay, LIMSI, Saclay.

Contacts

Coordinateur du collège : benoit.leblanc@ensc.fr.

Trois membres du comité de pilotage du collège sont membres du Conseil d'Administration de l'AFIA :

- Catherine FARON, faron@i3s.unice.fr ;
- Marie LEFEVRE, marie.lefevre@liris.cnrs.fr ;
- Domitile LOURDEAUX, domitile.lourdeaux@hds.utc.fr.

AfIA
Association française
pour l'Intelligence Artificielle

Collège Représentation et Raisonnement

Objectif du collège

L'objectif du Collège Représentation et Raisonnement (R&R) de l'AFIA est d'animer les communautés de recherche françaises dans ce domaine. Les thématiques de recherche sont relatives aux méthodes et outils fondamentaux de l'Intelligence Artificielle. Elles portent sur :

- la définition de modèles de représentation des informations (croyances, connaissances, préférences, obligations et permissions, actions, incertitude, confiance, réputation) comme les langages des logiques classiques ou non classiques, les modèles possibilistes, les ontologies, les langages à base de contraintes, les représentations graphiques, etc. ;
- la définition et l'automatisation de raisonnements sur ces informations : raisonnement spatio-temporel, dynamique des informations, révision de croyances, fusion d'informations symboliques, raisonnement par argumentation, raisonnement causal, raisonnement abductif, raisonnement à partir de cas, etc. ;
- la perspective algorithmique et de représentation pour des concepts utilisés dans des théories connexes comme la théorie des jeux ou la théorie du choix social (équilibre, stratégie gagnante, manipulation, etc.) : théorie des jeux algorithmique et choix social computationnel ;
- la mise au point de méthodes de codage des informations et d'algorithmes de traitement efficaces : compilation de connaissances, SAT, ASP, etc. ;
- la modélisation formelle de l'interaction : entre utilisateurs et systèmes informatiques, entre entités informatiques autonomes (agents) ;
- et généralement le lien avec différentes techniques liées à la décision, la planification, l'ordonnancement, le diagnostic, l'apprentissage, les sciences des données, etc.

Ces thématiques couvrent de très nombreux contextes d'application, comme par exemple le Web sémantique, le Web des données, les systèmes de recommandation ou d'aide à la décision, les agents conversationnels et assistants personnels, la programmation des jeux, la robotique, etc.

Programme de travail

Le collège R&R est impliqué dans les activités suivantes :

- IAF : Journées d'Intelligence Artificielle Fondamentales. Ces journées ont lieu tous les ans, à l'initiative du comité IAF. Elles sont articulées autour de 3 ou 4 exposés de synthèse invités, ainsi que d'un programme constitué après appel à communication ;
- JFPDA : Journées Francophones sur la Planification, la Décision et l'Apprentissage pour la conduite de systèmes. Ces journées ont lieu tous les ans, à l'initiative des anciens présidents de comités de programme JFPDA et des participants à la liste de diffusion jfpda@loria.fr. Elles sont articulées autour de 2 ou 3 exposés de synthèse invités, ainsi que d'un programme constitué après appel à communication ;
- JFPC : Journées Francophones de Programmation par Contraintes. Ces journées ont lieu tous les ans à l'initiative de l'AFPC (Association Française pour la Programmation par Contraintes). Elles sont articulées autour de 2 ou 3 exposés de synthèse invités, ainsi que d'un programme constitué après appel à communication ;

— des ateliers thématiques, qui ont lieu lors d'autres évènements et la contribution à l'organisation de journées communes.

Le collège consacre une partie de son budget de fonctionnement à l'attribution de bourses permettant à des étudiants d'assister à l'un des évènements organisés par le collège, ou d'autres manifestations pertinentes au regard des thématiques scientifiques.

Comité de pilotage

Le comité de pilotage du collège est constitué des personnes suivantes :

— Elise BONZON, Université Paris-Descartes, LIPADE, Paris ;
— Zied BOURAOUI, co-président du comité de programme IAF, Université d'Artois, CRIL, Lens ;
— Sylvie DOUTRE, co-présidente du comité de programme IAF, Université Toulouse 1 Capitole, IRIT, Toulouse ;
— Sébastien KONIECZNY, ancien directeur du GDR IA, CNRS, CRIL, Lens ;
— Fréderic MARIS, membre du CA et représentant pour l'AFIA, Université Toulouse 3 Paul Sabatier, IRIT, Toulouse ;
— Nicolas MAUDET, Université Pierre et Marie Curie, LIP6, Paris ;
— François SCHWARZENTRUBER, ancien président du comité de programme JFPDA, ENS Rennes, IRISA, Rennes ;
— Laurent SIMON, membre du CA de l'AFIA, président de l'AFPC, Université de Bordeaux, LaBRI, Bordeaux ;
— Elise VAREILLES, membre du CA de l'AFPC ISAE Supaero, Toulouse ;
— Bruno ZANUTTINI, ancien président du comité de programme JFPDA, Université Caen Normandie, GREYC, Caen.

La composition du comité est amenée à être modifiée bi-annuellement.

Contacts

Coordinateur du collège : frederic.maris@irit.fr.

Listes de diffusion : bull-i3@irit.fr, bull-ia@gdria.fr, jfpda@loria.fr.

Deux membres du comité de pilotage du collège sont membres du Conseil d'Administration de l'AFIA :

— Fréderic MARIS, frederic.maris@irit.fr ;
— Laurent SIMON, lsimon@labri.fr.

AfIA
Association française
pour l'Intelligence Artificielle

Collège Science de l'Ingénierie des Connaissances

Objectif du collège

L'Ingénierie des Connaissances (IC) est une thématique de l'Intelligence Artificielle (IA). Sa finalité est la production de systèmes « intelligents et explicables », capables d'aider l'humain dans ses activités et ses prises de décisions.

La construction des modèles de connaissances (principalement à base d'ontologies) et leur évaluation reposent sur une prise en compte des contextes applicatifs des différents domaines d'étude et des cas d'usage qui leur correspondent. La représentation formelle permet des raisonnements automatiques sur ces connaissances et sur les données qui leur sont associées, pouvant être complexes, hétérogènes, évolutives et distribuées. Elle permet aussi des tests de validité et de cohérence des modèles développés. L'opérationnalisation des modèles pose des questions de passage à l'échelle et d'explicabilité des résultats destinés aux utilisateurs des systèmes construits.

Les thématiques de recherche de l'IC sont relatives aux méthodes et outils de l'IA. Elles traitent un ensemble de thèmes portant, entre autres, sur les graphes et les modèles de connaissances dont nous listons les principaux :

- construction, réutilisation et mise à jour de ressources sémantiques, liage de données, alignement complexe, gestion des données et connaissances évolutives ;
- découverte de connaissances : fouille de ressources structurées et non structurées, découverte de classes et de propriétés, découverte de règles ;
- validation et évaluation : métriques d'évaluation, explicabilité des résultats, détection d'erreurs, interaction avec les utilisateurs ;
- gestion de données complexes, spatiales et temporelles,
- gestion des flux de données hétérogènes : imprécision, incertitude, interopérabilité sémantique ; mise en œuvre dans le contexte de l'internet des objets (IoT) ;
- éthique et Ingénierie des Connaissances.

Programme de travail

Le collège Science de l'Ingénierie des Connaissances (SIC) de l'AFIA a premièrement un rôle de diffusion de l'information. Il gère une liste de diffusion (comptant plus de 800 inscrits) sur l'ingénierie des connaissances et maintient une page web sur le site Web de l'AFIA. Il contribue également à l'animation et au dynamisme de la communauté de recherche en ingénierie des connaissances. Il est moteur dans l'organisation de différentes manifestations, comme :

- la conférence nationale Ingénierie des Connaissances (IC) qui a lieu chaque année lors de PFIA et les ateliers qui lui sont associés ;
- les sessions spéciales Interaction Management in Digitized Ecosystems organisées dans le cadre du chapitre français IEEE Systems Man and Cybernetics (SMC) ;
- la conférence internationale « *Terminology & Ontology : Theories and applications* » (TOTh) ;
- les journées dédiées au web sémantique dans le monde professionnel (SemWebPro) ;

— l'édition de numéros spéciaux en lien avec les thématiques du collège.

Les membres du collège contribuent à la mise en place de plusieurs « Journées Communes » de l'AFIA avec d'autres sociétés savantes, d'ateliers sur la Plate-Forme Intelligence Artificielle (PFIA) et de journées thématiques internes au collège. Ils participent activement à des événements internationaux tels que EKAW, K-CAP, ESWC, ECAI, TheWebConf ou encore ISWC.

Comité de pilotage

Le comité de pilotage du collège est constitué de chercheurs ou enseignants chercheurs spécialisés dans le domaine de l'ingénierie des connaissances, tous membres des comités de programme de la conférence nationale « journées francophones d'Ingénierie des Connaissances » (IC). Ils font également le lien entre des sociétés savantes (ARIA (Association Francophone de Recherche d'Information (RI) et Applications), ATALA (Association pour le Traitement Automatique des Langues), AIM (Association d'Informatique Médicale)) et avec les GDRs RADIA et MADICS, humanité numérique :

— Marie-Hélène ABEL, Université de Technologie de Compiègne, HEUDIASYC ;
— Jean CHARLET, Assistance Publique-Hôpitaux de Paris & INSERM, LIMICS ;
— Sylvie DESPRÉS, Université Sorbonne Paris Nord,, LIMICS ;
— Catherine FARON, Université Nice Sophia Antipolis, I3S et Inria ;
— Nathalie HERNANDEZ, Université Toulouse 2 Jean Jaurès , IRIT ;
— Nathalie PERNELLE, Université Sorbonne Paris Nord, LIPN ;
— Maxime LEFRANÇOIS, Mines Saint Etienne, IMT, LIMOS ;
— Nathalie PERNELLE, EURECOM, Sophia Antipolis ;
— Catherine ROUSSEY, INRAE Centre Occitanie Montpellier, MISTEA ;
— Fatiha SAÏS, Université Paris Saclay, LISN ;
— Cassia TROJAHN, Université Jean Jaurès Toulouse 2, IRIT ;
— Haifa ZARGAYOUNA, Université Sorbonne Paris Nord, LIPN.

Contacts

Coordinateur du collège : sylvie.despres@univ-paris13.fr.
Listes de diffusion : info-ic@inria.fr.
Mail pour contacter le comité de pilotage du collège : collegeSIC@afia.asso.fr.
Deux membres du comité de pilotage sont membres du conseil d'administration de l'AFIA :

— Fatiha SAÏS, fatiha.sais@lisn.fr ;
— Catherine ROUSSEY, catherine.roussey@inrae.fr.

AfIA
Association française
pour l'Intelligence Artificielle

Collège Systèmes Multi-Agents et Agents Autonomes

Objectif du collège

Le collège Systèmes Multi-Agents et Agents Autonomes (SMAA) de l'AFIA a pour mission d'organiser les activités du collège au sein de l'AFIA et d'assurer l'interaction entre l'AFIA et la communauté francophone SMA et Agents Autonomes, concernant leur volet IA. Il participe à l'animation scientifique au sein de l'AFIA, en coordination avec les communautés concernées, pour susciter l'implication des membres du collège dans les événements organisés par l'AFIA (p. ex. PFIA, CNIA) tant en participant aux comités scientifiques, qu'en organisant des manifestations d'intérêt pour la communauté IA en France.

Le collège SMAA évolue dans deux directions :

— Accentuation des interactions avec les communautés robotique, automatique d'une part et simulation, systèmes complexes d'autre part. Concrètement cela se traduira par l'organisation d'événements joints sur des travaux associant SMA et/ou agents conversationnels animés à ces disciplines. Lorsque d'autres champs d'IA seront ciblés, ces événements seront préparés en association avec les autres collèges de l'AFIA concernés.

— Mise en place de webinaires réguliers, issues des équipes impliquées dans le collège SMAA. Les séminaires d'équipes, soutenances de thèses, les soutenances de HDR pourront, sur base du volontariat être diffusées à l'ensemble du collège SMAA.

Programme de travail

Les missions du collège SMAA concernent l'organisation de manifestations scientifiques (conférences, journées thématiques), l'édition de dossiers techniques ou de numéros spéciaux de journaux sur des thématiques d'intérêt pour la communauté.

Le collège SMAA accompagne notamment l'organisation régulière des JFSMA et de WACAI :

— JFSMA 2021 à PFIA Bordeaux ;

— JFSMA 2022 à Saint-Etienne ;

— WACAI 2022 (lieu à définir).

Le collège SMAA accompagne également l'organisation des journées communes suivantes :

— Journée commune MACS & IA ;

— Journée commune Robotique & IA ;

— Journée commune Simulation & IA.

Il propose également un soutien similaire à d'autres événements, ponctuels ou récurrents, relevant de son périmètre scientifique.

Plus largement le collège envisage des actions d'animation à destination des doctorants, et des actions de médiation scientifique et de communication à destination du public.

Comité de pilotage

Le comité de pilotage du collège SMAA est constitué de chercheurs spécialisés dans le domaine des systèmes multi-agents et des agents autonomes, tous membres des comités de programme ou du comité consultatif des Journées Francophones en Systèmes Multi-Agents (JFSMA) et/ou du Workshop Affect, Compagnon Artificiel, Interaction (WACAI).

— Emmanuel ADAM, Université Polytechnique Hauts-de-France, LAMIH, Valenciennes ;
— Fabien MICHEL, Université de Montpellier, LIRMM, Montpellier ;
— Frédéric MIGEON, Université Toulouse 3 Paul Sabatier, IRIT, Toulouse ;
— Maxime MORGE, Université de Lille 1, CRIStAL, Lille ;
— Magalie OCHS, Université Aix-Marseille, LSIS, Marseille ;
— Gauthier PICARD, ONERA, Toulouse ;
— Nicolas SABOURET, Université Paris-Sud, LIMSI, Saclay ;
— Olivier SIMONIN, INSA Lyon, CITI, Lyon ;
— Mahdi ZARGAYOUNA, IFSTTAR, Paris.

Contacts

Coordinateur du collège : emmanuel.adam@uphf.fr.

Listes de diffusion : sma@loria.fr, acai@poleia.lip6.fr.

Deux membres du comité de pilotage sont membres du conseil d'administration de l'AFIA :

— Emmanuel ADAM, emmanuel.adam@uphf.fr ;
— Gauthier PICARD, gauthier.picard@onera.fr.

AfIA
Association française
pour l'Intelligence Artificielle

Collège Technologies du Langage Humain

Objectif du collège

Les Technologies du Langage Humain (TLH) proposent des méthodes permettant une communication homme-machine naturelle, pouvant s'étendre à une interaction homme-homme médiée. Ainsi, les TLH permettent d'analyser, d'interpréter et de produire des actes du langage écrit, parlé ou signé, mais aussi d'interagir avec des données langagières. Ainsi, les TLH englobent traditionnellement le Traitement Automatique des Langues (TAL), la Communication Parlée (CP) et leurs applications les plus emblématiques comme la Recherche d'Information (RI) et la Traduction Automatique.

L'étude du langage humain est une entreprise multidisciplinaire qui nécessite une expertise dans les domaines de la Linguistique, de la Psychologie, des Sciences Cognitives, des Sciences du Numérique, et notamment de l'Intelligence Artificielle (IA). Les TLH occupent une part entière en IA sous le prisme du Test de Turing. Ainsi, elles regroupent tous les axes de recherche de l'IA dans le cadre des données langagières comme la représentation (*e. g.* plongements lexicaux, analyse syntaxique), le raisonnement (*e. g.* systèmes de question-réponse), la planification (*e. g.* argumentation), l'apprentissage (*e. g.* analyse de sentiments), ou même l'intelligence collective (*e. g.* détection de communautés). Créer des modèles pouvant interagir de manière élégante et naturelle en utilisant la langue nécessite une compréhension profonde de l'acoustique, de la phonétique et de la prosodie pour le langage oral d'une part (dans le cadre de la reconnaissance et de la synthèse de la parole), et d'autre part, de la morphologie, de la syntaxe, de la sémantique et de la pragmatique pour le langage écrit ou signé (dans le cadre de l'analyse et de la génération de textes). Seulement à cette condition les applications des TLH peuvent offrir la promesse d'un accès « universel » à l'information, notamment à travers les moteurs de recherche, la traduction automatique, le résumé de textes, la veille automatique ou les systèmes de question-réponse. La compréhension globale du langage permet également de caractériser les textes suivant leurs objectifs communicationnels. Ainsi, l'analyse des sentiments et des émotions, l'identification de discours haineux, la détection de plagiats, l'identification et la vérification du locuteur proposent autant de cadres applicatifs importants pour les sociétés numériques.

Puisqu'à l'ère du numérique les données langagières prolifèrent dans des quantités toujours plus importantes et accessibles (*big data*), les champs d'applications des TLH sont nombreux. Ainsi, les archives numériques, les réseaux sociaux, les plateformes collaboratives, les entretiens clients, les textos, les courriels, les commentaires sur des plateformes de vente en ligne sont autant de matières premières pour le développement d'applications des TLH. En particulier, quelques domaines privilégiés sont la santé, l'éducation, le droit, le journalisme et le handicap, mais d'autres émergent comme la finance, l'agriculture, la sécurité, le marketing et les humanités numériques.

Parallèlement au développement de modèles des TLH, de nombreux défis connexes doivent être pris en compte dans le cadre d'une démarche globale, comme la confiance (*e. g.* reproductibilité, explicabilité, confidentialité), l'éthique (*e. g.* biais d'apprentissage, représentativité) et l'évaluation (*e. g.* métriques dédiées) des systèmes proposés.

Le collège Technologies du Langage Humain (TLH) de l'AFIA a donc pour mission de promouvoir l'animation et l'interaction scientifiques entre les communautés TAL, CP et RI, et l'ensemble des communautés en IA ayant des intérêts communs dans le but de consolider les collaborations transversales.

Programme de travail

Afin de mener à bien sa mission, le collège TLH s'engage à soutenir l'organisation de manifestations scientifiques (conférences, ateliers), animer des groupes de travail, éditer des dossiers techniques, organiser des journées thématiques et diffuser et communiquer autour des recherches des communautés françaises du TAL, de la CP, de la RI et de l'IA.

Le collège TLH s'engage à servir de canal de communication entre l'AFIA et ses collèges ainsi qu'avec l'ATALA (Association pour le Traitement Automatique des Langues), l'ARIA (Association Francophone de Recherche d'Information et Applications), et l'AFCP (Association Francophone de la Communication Parlée).

Comité de pilotage

Le comité de pilotage du collège est constitué de 10 chercheuses et chercheurs spécialisés dans le domaine du TAL, de la CP et de la RI.

- Florian BOUDIN, Université de Nantes, LS2N, Nantes ;
- Davide BUSCALDI, Université Paris 13, LIPN, Villetaneuse ;
- Gaël DIAS, Université de Caen Normandie, GREYC, Caen ;
- Emmanuelle ESPERANÇA-RODIER, Université Grenoble Alpes, LIG, Grenoble ;
- Corinne FREDOUILLE, Université d'Avignon, LIA, Avignon ;
- José MORENO, Université Toulouse 3 Paul Sabatier, IRIT, Toulouse ;
- Aurélie NÉVÉOL, CNRS, LISN, Saclay ;
- Yannick PARMENTIER, Université de Lorraine, LORIA, Nancy ;
- Mathieu ROCHE, CIRAD, TETIS, Montpellier ;
- Marie TAHON, Le Mans Université, LIUM, Le Mans.

Contacts

Coordinateurs du collège : corinne.fredouille@univ-avignon.fr et jose.moreno@irit.fr.

Listes de diffusion :

- ln@cines.fr (TAL) ;
- parole@listes.afcp-parole.org (CP) ;
- info-aria@lsis.org (RI).

Un membre du comité de pilotage est membres du conseil d'administration de l'AFIA :

- Gaël DIAS, gael.dias@afia.asso.fr.

AfIA
Association française
pour l'Intelligence Artificielle

Prix de Thèse IA 2023

Le jury

— Présidente : Hélène FARGIER ;
Isabelle BLOCH, Antoine CORNUEJOLS, Sébastien DESTERCKE, Jean-Gabriel GANASCIA, Jérôme LANG, Pierre MARQUIS, Edouard PAUWELS, Marie-Christine ROUSSET, Thomas SCHIEX, Mathieu SERRURIER, Bruno ZANUTTINI.

Les lauréats

— 1er prix (ex-æquo) : Vincent GRARI, « *Adversarial mitigation to reduce unwanted biases in machine learning* », dir. : Marcin DETYNIECKI et Sylvain LAMPRIER, 22 Mai 2022, Sorbonne Université ;
— 1er prix (ex-æquo) : Munyque MITTELMANN, « Logiques pour la représentation et la conception d'enchères », dir. : Laurent PERRUSSEL, 01/09/2022, Université Toulouse Capitole, IRIT.

Prix de Thèse IA 2024

Le jury

— Président : Andreas HERZIG ;
— Membres : Thierry ARTIERES, Isabelle BLOCH, Tristan CAZENAVE, Élisa FROMONT, Thomas GUYET, Pierre MARQUIS, Nicolas MAUDET, Fatiha SAIS, Thomas SCHIEX, Nicolas THOME.

Les lauréats

— 1er prix : Virginie DO, « *Fairness in recommender systems : insights from social choice* », dir. : Nicolas USUNIER, Jérôme LANG et Jamal ATIF, 11/07/2023, Meta AI / Université Paris Dauphine-PSL ;
— accessit (ex-æquo) : Pierre MARION, « *Mathematics of deep learning : generalization, optimization, continuous-time models* », dir. : Gérard BIAU et Jean-Philippe VERT, 20/11/2023, LPSM / Sorbonne Université ;
— accessit (ex-æquo) : Yuan YIN, « *Physics-aware deep learning and dynamical systems : hybrid modeling and generalization* », dir. : Patrick GALLINARI et Nicolas BASKIOTIS, 28/06/2023, ISIR / Sorbonne Université.

AfIA
Association française
pour l'Intelligence Artificielle

Bulletins

Le Bulletin de l'AFIA fournit un cadre de discussions et d'échanges au sein de la communauté. Toutes les contributions, pour peu qu'elles aient un intérêt général, sont les bienvenues. Le bulletin contient des rubriques régulières de comptes rendus des conférences, journées et autres événements que l'AFIA organise ou parraine, les résumés d'HDR et de Thèses de Doctorat, et un Dossier qui dresse un état de l'art sur un domaine particulier de l'IA, présente des équipes françaises de recherche en IA (académiques ou industrielles), ou PFIA. Les bulletins de l'AFIA sont accessibles librement depuis le site de l'AFIA.

Le comité de rédaction

Le comité de rédaction 2023 avait Dominique LONGIN comme rédacteur en chef, aidé de Emmanuel ADAM, Grégory BONNET (rédacteur adjoint) et Gaël LEJEUNE. En 2024, c'est la même équipe qui a souhaité continuer à s'investir dans le comité de rédaction.

Bulletin 118 — Dossier « PFIA 2022 » — octobre 2022

- Le dossier de ce Bulletin est dirigé par Engelbert Mephu NGUIFO. Il est consacré à PFIA 2022 (Saint-Étienne, du 27 juin au 1er juillet 2022) qui a hébergé 8 conférences (APIA, CNIA, IC, JFPC, JFPDA, JFSMA, JIAF et RJCIA), 4 journées (EIAH & IA, IoT & IA, Résilience & IA, Santé & IA), des tutoriels, des conférences invitées et la remise des 2 prix de thèse AFIA, en plus d'un événement consacré à l'année PROLOG.
- Un compte rendu des journées PFIA autour de l'IoT, de la santé et de la résilience est également disponible, ainsi qu'un résumé des exposés invités. Ce Bulletin recence également les thèses et HDR du 3e trimestre de 2022 dont nous avons eu connaissance.

Bulletin 119 — Dossier « Équipes académiques en IA » — janvier 2023

- Ce Bulletin a été dirigé par Dominique LONGIN qui également dirigé son dossier. Ce dernier dévoile les travaux de 11 équipes de recherche en IA réparties dans toute la France.
- Ce bulletin présente également la liste des thèses et HDR soutenues durant le 4e trimestre 2022 et dont nous avons eu connaissance.

Bulletin 120 — Dossier « IA & Normes » — avril 2023

- Le dossier très riche de ce Bulletin, dirigé par Nathalie NEVEJANS, était autour de la thématique « IA et normes » déclinée au travers de 16 contributions.
- Ce Bulletin contient également un compte rendu des Nuits de l'Info 2022, ainsi que la liste des thèses et HDR soutenues pendant le 1er trimestre 2023 et dont nous avons eu connaissance.

Bulletin 121 — Dossier « IA & Énergies » — juillet 2023

- Ce second dossier thématique de 2023 a été dirigé par Nouredine HADHJ-SAID et rassemble près d'une dizaine de contribution différentes.

— Ce Bulletin contient également les thèses et HDR dont nous avons eu connaissance et soutenue durant la période du 2e trimestre 2023.

Bulletin 122 **Dossier « PFIA 2023 »** **octobre 2023**

— Ce bulletin portait sur la Plate-Forme Intelligence Artificielle (PFIA 2023) qui s'est déroulée à Strasbourg du 3 au 7 juillet 2023. Il regroupe les comptes rendus : des 9 conférences hébergées (APIA, CAp, CNIA, IC, JFPC, JFSMA, JIAF, RJCIA et SFC.) de cette belle édition de PFIA ; de ses 4 journées communes (« Jeux et IA », « Santé et IA », « Affects, Compagnons Artificiels et Interactions » et « Résilience et IA ») ; d'un résumé des travaux menés par les jeunes docteurs à qui a été décerné un prix pour leur thèse.

— Ce bulletin contient également les comptes rendus de la journée AFIA/GRD RADIA sur le thème « Modèles hybrides & IA » et de la journée « Jam DriveToGaether ».

Bulletin 123 **Dossier « Équipes industrielles en IA »** **janvier 2024**

— Ce bulletin, comme tous les deux ans, contient un dossier sur les équipes de recherche industrielles en IA. Ce dossier, monté par Christophe BORTOLASO (Berger-Levrault) du Collège Industriel de l'AFIA, a l'originalité de débuter par un texte collectif du collège sur les LLM et l'IA générative. Il est suivi par les contributions de 4 entreprises différentes (Jeolis Solutions, Thalès, Cosling et son Choco Solver et enfin CS Group).

— Comme de tradition, les thèses en IA soutenues durant le trimestre écoulées sont présentées en fin de bulletin.

Bulletin 124 **Dossier « Regards croisés : AFIA & GDR RADIA »** **avril 2024**

— Ce bulletin est un peu particulier par rapport aux précédents car son dossier, que nous devons à Meghyn BIENVENUE et Nicolas MAUDET, présente et met en mirroir les 10 GT du GDR RADIA avec les 8 collèges de l'AFIA. Un travail très intéressant qui met en lumière toute la complémentarité de ces deux structures.

— Ce bulletin présente également les comptes rendus de FIIA 2023 et PDIA 2024, ainsi que les thèses et HDR soutenue sur le trimestre écoulé.

Bulletin 125 **Dossier « IA & Création artistique »** **juillet 2024**

— Ce bulletin est aussi un peu particulier puisque son dossier, sous la responsabilité de Samuel BIANCHINI, est résolument interdisciplinaire. Il regroupe 7 contributions, pour plonger dans l'univers passionnant de la création artistique sous toutes ses formes.

— Ce bulletin présente aussi une liste des (très nombreuses) thèses soutenues en fin d'année, et en relation avec l'IA.

AfIA
Association française
pour l'Intelligence Artificielle

FIIA 2022

L'Association Française pour l'Intelligence Artificielle (AFIA) organise sa son septième « Forum Industriel de l'IA » sur le thème « Hybridation & IA ». Il prolonge le quatrième forum tenu en 2019 dont le thème était les « Systèmes mixtes ».

La journée est constituée de présentations pour permettre des échanges accrus entre académiques et industriels. Une dernière session est réservée à des présentations rapides par des industriels invités, dont candidats à rejoindre le Collège Industriel de l'AFIA.

Date et lieu

— Date : 6 Octobre 2022
— Lieu : TOTEM / Institut des Systèmes Complexes Paris IdF, 11 Place Nationale, 75013 Paris

Programme

— 09h00 Ouverture de David CHAVALARIAS (Directeur de l'ISC IdF) et Benoit LE BLANC (Président de l'AFIA). Introduction de la journée.
— 09h15 « *PINNs and other approaches to accelerate simulations* » par SINCLAIR Laboratory (Total Energy, EDF, Thales).
— 10h15 Pause.
— 10h45 « *Quantum machine learning at Pasqal* » par PASQAL.
— 11h25 « *Hybrid AI in the ANITI projects and for natural language processing* » par Nicholas ASHER (CNRS, IRIT).
— 12h30 Buffet.
— 13h30 « *Symbolic and data-driven AI for image interpretation* » par Céline HUDELOT (Centrale Supélec).
— 14h10 « *Designing rule-based decision policy with reinforcement learning* » par Thales Research & Technology.
— 14h50 « *Rule-based-model discovery for business automation* » par IBM France.
— 15h30 Pause.
— 16h00 Présentations de sociétés concernées par l'IA en « Trois planches ».
— 16h05 « Une approche hybride pour accompagner la mémoire du REx dans l'élaboration d'avis technique » par Jean-Pierre COTTON et Alain BERGER (Ardans).
— 16h25 « De l'analyse d'images jusqu'au traitement des données numériques et textuelles ; au cœur des travaux de recherche appliquée » par Valérie REINER (Berger-Levrault).
— 17h00 Clôture.

Organisation

Cet événement est organisé par Patricia BESSON (Thales Research & Technology), Remy KUSTERS (IBM France) et Patrick FABIANI (Dassault Aviation) pour le Collège Industriel de l'AFIA.

Inscriptions

Les inscriptions à la journée (gratuites pour les membres AFIA, 30 € sinon) sont obligatoires et à effectuer sur le site : https://afia.asso.fr/inscription-fiia/. Un lien de connexion à une salle visio-conférence sera envoyé aux inscrits.

Institut des Systèmes Complexes

cnrs

AfIA
Association française
pour l'Intelligence Artificielle

EFIA 2023

L'Association Française pour l'Intelligence Artificielle (AFIA) organise sa quatrième journée « Enseignement et Formation en IA » sur le thème « Compétences et métiers d'avenir en IA ».

La demi-journée (en distanciel) a pour but d'échanger autour des nouvelles compétences et métiers d'avenir en IA, avec les projets lauréats de l'Appel à Manifestations d'Intérêt « Compétences et Métiers d'Avenir » (AMI CMA 2022). Ce sera l'occasion d'identifier les innovations pédagogiques, les évolutions de contenus ainsi que la projection sur de nouveaux métiers d'avenir, dans le cadre de ces projets de grande envergure.

Date et lieu

— Date : 05 janvier 2023
— Lieu : Journée organisée en distanciel

Programme

— 13h30 Ouverture de Benoit LE BLANC (Président de l'AFIA). Introduction par Gauthier PICARD (membres du CA de l'AFIA et responsable du GT Enseignement).
— 13h45 « Stratégie d'accélération française sur la formation en IA » par Anne BOYER (MESR Paris).
— 14h00 « *Artificial Intelligence for All* Sorbonne Université » par Xavier FRESQUET (Sorbonne Université Paris) et Agnès DUDYCH (Sorbonne Université Paris).
— 14h30 « *SaclAI-School : Paris-Saclay AI Education Program* » par Frédéric PASCAL (Université Paris-Saclay) et Sarah COHEN-BOULAKIA (Université Paris-Saclay).
— 15h00 « Massification et accélération des Formations en IA par IP Paris et HEC Paris » par Dominique ROSSIN (Ecole Polytechnique Paris).
— 15h30 Pause.
— 15h45 « École Française de l'Intelligence Artificielle » par Chantal SOULÉ-DUPUY et Julie AL-ATRACH (Université Toulouse Capitole) et Jean-Louis ROCH (Université Grenoble Alpes).
— 16h45 Table ronde.
— 17h30 Clôture.

Organisation

Cet événement est organisé par Laurent VERCOUTER pour l'AFIA et Gauthier PICARD pour le Conseil d'Administration de l'AFIA.

Inscriptions

Les inscriptions à la journée (https://afia.asso.fr/inscription-efia/) sont gratuites mais obligatoires.

AfIA
Association française
pour l'Intelligence Artificielle

PDIA 2023

L'Association Française pour l'Intelligence Artificielle (AFIA) organise sa neuvième journée « Perspectives et Défis en Intelligence Artificielle » sur le thème « IA et écologie ».

La journée est composée d'interventions regroupées selon deux sessions : Impact de l'IA sur l'écologie en matinée ; l'IA au service de l'écologie l'après-midi.

Date et lieu

— Date : 04 Avril 2023
— Lieu : CNAM, Amphi Friedmann, Accès 33, 2 rue Conté, 75003 Paris

Programme

— 09h00 Accueil.
— 09h15 Présentation de l'AFIA . Introduction.
— 09h30 « Bénéfices/risques de l'IA pour l'environnement : que faire en contexte d'incertitude ? » par Julien LEFEVRE (Aix-Marseille Université, INT).
— 10h30 « Une dystopie de l'apprentissage sobre » par Denis TRYSTAM (Université Grenoble Alpes, Institut Universitaire de France).
— 11h30 Pause.
— 11h45 Courtes présentations de doctorants.
— 12h45 Buffet.
— 14h00 « L'IA pour l'agro-écologie » par Christine DILLMANN (Université Paris Saclay, INRAE).
— 15h00 « *Machine Learning for Climate Change and Environmental Sustainability* » par Claire MONTELEONI (University of Colorado Boulder, INRIA Paris).
— 16h00 Pause.
— 16h15 Courtes présentations de doctorants.
— 17h15 Discussion.
— 17h30 Clôture.

Organisation

Cet événement est organisé par Anne-Laure LIGOZAT (LISN, Université Paris Saclay), Fatiha SAIS (LISN, Université Paris Saclay) pour le Conseil d'Administration de l'AFIA, et Fayçal HAMDI (CEDRIC, CNAM Paris).

Inscriptions

Les inscriptions à la journée (gratuites pour les membres AFIA, 30 € sinon) sont obligatoires et à effectuer sur le site : https://www.linscription.com/pro/activite.php?P1=128448. Le déjeuner est garanti à toutes les personnes qui se seront inscrites avant le 31/03/2023, et seulement à celles-ci.

AFIA
Association française
pour l'Intelligence Artificielle

Résilience & IA 2023

L'Association Française pour l'Intelligence Artificielle (AFIA) organise au travers de son collège industriel sa seconde journée « Résilience et IA » au sein de la Plate-Forme Intelligence Artificielle 2023 (PFIA 2023).

Cette journée a pour objectif de faire un point sur les travaux menés sur la gestion de crise et la sortie de crise, communément appelé résilience, avec l'appui du traitement des données, de l'information et du Big Data. Outre le constat souvent appuyé sur des chiffres d'une crise ou d'un dysfonctionnement, les sorties de crise et de situations difficiles imposent la collecte et un traitement efficace des données pour une meilleure prise de décision traduite par des politiques de développement, de mise en place des outils, des réponses précises. Des techniques d'Intelligence Artificielle peuvent ou pourraient aider à la résilience.

Il sera aussi intéressant de voir pour des situations précises quelles sont les données massives ou non qui sont à disposition (accès libre par exemple) et des desiderata en terme des connaissances à extraire.

Date et lieu

— Date : 5 juillet 2023
— Lieu : PFIA 23, Strasbourg

Programme

— 10h30 « *AI Research for Climate Change and Environmental Sustainability* » par Claire MONTELEONI (INRIA-Paris).
— 11h30 « Les défis du glissement de contexte géographique » par Theophile BAYET, Christophe DENIS, Alassane BAH, Jean-Daniel ZUCKER.
— 12h30 Repas.
— 14h40 « Fonder le concept de résilience sur la théorie de la viabilité dans le cas de dynamiques incertaines » par Jean-Denis MATHIAS (INRAE Clermont ARA).
— 15h40 Panel autour de la résilience et l'apport de l'IA.
— 16h30 « Escape-SG : Un jeu sérieux pour mieux préparer les évacuations de masse. Simulation à base d'agents » par Arnaud SAVAL, Mathieu BOURGAIS, Éric DAUDÉ, Pierrick TRANOUEZ, Oliviet GILLET.
— 17h00 « Gestion des connaissances partagées par des agents à ressources et connectivité limitées : étude, analyse et expérimentation » par Mohamed LIMAME, Julien HENRIET, Christophe LANG, Nicolas MARILLEAU.
— 17h30 Clôture.

Organisation

Cet événement est organisé par Ghislain A. ATEMEZING (Mondeca, Paris) et Mihaela JUGANARU-MATHIEU (Mines de Saint Etienne, IMT).

AfIA
Association française
pour l'Intelligence Artificielle

FIIA 2023

L'Association Française pour l'Intelligence Artificielle (AFIA) organise son huitième « Forum Industriel de l'IA » sur le thème « Large Language Models (LLM) & IA ». La date a été en effet différée en 2024 pour des raisons d'organisation.

La journée est constituée de présentations pour permettre des échanges accrus entre académiques et industriels. Une dernière session permettra des échanges entre les industriels et universitaires invités, sur les verrous technologiques identifiés, et les perspectives de résolutions.

Date et lieu

— Date : 9 février 2024
— Lieu : TOTEM / Institut des Systèmes Complexes Paris IdF, 11 Place Nationale, 75013 Paris

Programme

— 09h00 « Mot d'accueil de l'AFIA » par Valérie REINER (Coordinatrice du collège industriel de l'AFIA).
— 09h10 « Introduction de la journée » par Davy MONTICOLO (Représentant du collège industriel de l'AFIA) et Bruno CARRON (Airbus Defence and Space).
— 09h25 « *Connaissances et IA* » par Guilherme ALVES (INRIA).
— 10h10 « *Approches IA hybrides et applications au domaine de la défense* » par Claude FENDZI et Geraud FAYE(Airbus Defence and Space).
— 10h55 Pause.
— 11h10 « *ChatDOC : une IA Générative interne pour interroger des documents clients* » par Aymen SHABOU et Mohamed DHOUIB (DataLab Groupe Crédit Agricole).
— 11h55 « *Mathématiques et IA* » par Marianne CLAUSEL (Institut Elie Cartan Nancy).
— 12h30 Buffet.
— 13h55 « *Ingénierie des connaissances, ontologie* » par Cécilia ZANNI (INSA Rouen).
— 14h40 « *Applications de l'IA* » par Christophe BORTOLASO (Berger Levrault).
— 15h25 Table ronde : « *Apports de la journée, Perspectives (verrous et ouvertures)* » par l'ensemble des orateurs et des oratrices.
— 16h10 Conclusion et prochains travaux (talks, bulletin, PFIA, APIA) par Davy MONTICOLO (Représentant du collège industriel de l'AFIA) et Bruno CARRON (Airbus Defence and Space).
— 16h30 Clôture.

Organisation

Cet événement est organisé par Davy MONTICOLO (Représentant du collège industriel de l'AFIA) et Bruno CARRON (Airbus Defence and Space).

Inscriptions

Les inscriptions à la journée (gratuites pour les membres AFIA, 30 € sinon) sont obligatoires et à effectuer sur le site : https://afia.asso.fr/inscription-fiia/.

Institut des Systèmes Complexes

cnrs

PDIA 2024

L'Association Française pour l'Intelligence Artificielle (AFIA) organise sa neuvième journée « Perspectives et Défis en Intelligence Artificielle » sur le thème « NEUROSCIENCE ET IA ».

La journée est hybride, pouvant être donc suivie à distance.

Date et lieu

— Date : 09 Avril 2024
— Lieu : ENSC, 109 avenue Roul, 33400 Talence

Programme

— 09h30 Accueil.
— 10h00 Présentation de l'AFIA par Benoit LEBLANC (Président de l'AFIA) et introduction de la journée PDIA – Perspectives et Défis en IA – Neuroscience et IA par Thomas BORAUD (Campus Neuroscience).
— 10h15 « *IVirtual Brain Twins at the interface of AI and the brain* » par Viktor JIRSA (Institut de Neurosciences des Systèmes, Marseille).
— 11h15 « *IDeep Learning in Medical Imaging : What's Needed for Training Data ?* » par Francesca GALASSI (Empenn lab, Inria Rennes).
— 12h15 Pause.
— 13h45 « *IHippocampal cells inside a random but embodied computational model* » par Naomi CHAI-EICHEL (Institut des Maladies Neurodégénératives, CNRS, Inria, U. Bordeaux).
— 14h45 « *IA neuro-inspirée pour le codage prédictif et neurosymbolique dans l'acquisition du langage et la planification* » par Alex PITTI (ETIS Lab, CY Alliance, Cergy Paris).
— 15h45 Pause.
— 16h15 « Algorithmes d'apprentissage et de prise de décision à la croisée des chemins entre l'Intelligence Artificielle et les Neurosciences » par Mehdi KHAMASSI (ISIR, Sorbonne Université).
— 17h15 « Comment exploiter et faire évoluer les relations IA-Neurosciences ?» » Table ronde.
— 18h00 Clôture.

Organisation

Cet événement est organisé par Frédéric ALEXANDRE (LABRI, INRIA), Arthur LEBLOIS (Université de Bordeaux, Neurocampus).

Inscriptions

Les inscriptions à la journée sont obligatoires uniquement pour les personnes souhaitant être présentes dans l'amphithéâtre. Pour bénéficier des pauses et du déjeuner, il est nécessaire de s'inscrire avant le 5 avril 2024.

— L'inscription est gratuite pour les personnes adhérentes à l'AFIA.
— Elle est de 30€ pour les personnes non adhérentes.

EGC & IA 2023

L'Association Française pour l'Intelligence Artificielle (AFIA) et l'association internationale francophone d'Extraction et de Gestion des Connaissances (Association EGC) organisent, avec l'aide locale du groupe de travail Gestion et l'Analyse de données Spatiales et Temporelles (GAST), une journée « Gestion et Analyse des données Maritimes » (GAM' 23) sur le thème de la représentation, de la gestion, de l'analyse et du stockage des données maritimes. Cette journée réunira les acteurs de la recherche académique ou industrielle autour de cette thématique.

Avec la multiplication des capteurs, des satellites et des systèmes d'émission, une grande variété de données liées au monde de la mer est désormais disponible et nécessite d'être analysée. Cette analyse est cruciale pour répondre à différentes problématiques, par exemple celles liées à l'activité des navires en mer (sécurité maritime, routage, détection d'activités illégales) ou aux enjeux environnementaux (réchauffement climatique, préservation de la biodiversité, pollution en mer). Le but de cette journée thématique est donc de réunir les gens intéressés par le traitement de données maritimes, celles-ci peuvent être par exemple des positions de navires, des données météo, des données d'images ou satellitaires, des données de qualité de l'eau ou de pollution qu'il s'agit d'assimiler et de traiter afin d'extraire de l'information.

Date et lieu

— Date : 11 mai 2023
— Lieu : EPITA, 14-16 rue Voltaire, 94270 Le Kremlin-Bicêtre

Programme

— 08h30 Accueil.
— 09h30 Mot d'accueil par Aurélie LEBORGNE (membre du comité de programme GAM) et par Thierry GERAUD (directeur du LRE-EPITA).
— 09h45 « Traitement et analyse de données de positions de navires pour le suivi du trafic maritime à différentes échelles » par Loïc SALMON (ISEA—Université de la Nouvelle-Calédonie).
— 10h30 « De l'AIS à la prédiction de temps d'arrivée : détection de routes maritimes » par Jacques EVERWYN (SINAY – Caen).
— 11h15 « Exploitation et croisement de données complexes d'aquaculture » par Nazha SELMAOUI (ISEA—Université de la Nouvelle-Calédonie).
— 12h00 Buffet.
— 13h30 « Les navires à l'escale : une approche des dynamiques maritimes caribéennes au regard des fréquentations portuaires » par Clément IPHAR (LETG – Université de Bretagne Occidentale).
— 14h00 « Maritime 4.0 : Challenges et opportunités dans la marine marchande » par Pedro MERINO LASO (ENSM — Nantes).
— 14h45 « Segmentation d'objets mobiles pour les images sous-marines en utilisant des réseaux de neurones de graphes » par Wieke PRUMMEL (MIA Lab — Université de La Rochelle).
— 15h30 « Surveillance en mer » par Olivier RICOU (LRE — Le Kremlin-Bicêtre).
— 16h15 Mot de la fin par Aurélie LEBORGNE, Nida MEDDOURI et Loïc ALMON (membre du comité de programme GAM).

Organisation

Cet événement est organisé par Nida MEDDOURI (LRE/EPITA Kremlin-Bicêtre), Aurélie LEBORGNE (ICube/UNISTRA Strasbourg) et Loïc ALMON (ISEA/UNC (Nouméa – Nouvelle-Calédonie).

IHM & IA 2023

L'Association Française pour l'Intelligence Artificielle (AFIA) et l'Association Francophone d'Interaction Homme-Machine (AFIHM) organise une sixième journée commune « IHM et IA » sur les thème « Concevoir des interactions avec et pour l'IA » (en matinée) et « Construire, évaluer, et déployer des IA de confiance » (l'après-midi).

Cette journée s'articulera autour de présentations, débats, posters et démos mettant en évidence les liens entre les deux disciplines.

Date et lieu

— Date : 7 juin 2023
— Lieu : Sorbonne Center for Artificial Intelligence (SCAI), 4 Pl. Jussieu, 75005 Paris

Programme

— 09h30 Café d'accueil.
— 10h00 Présentation de l'AFIA et de l'AFIHM.
— 10h15 « Composer les interactions avec les systèmes d'apprentissage automatique » par Jules FRANÇOISE (CNRS, Université Paris-Saclay).
— 10h45 « Évaluation et utilisations expertes des systèmes d'apprentissage automatique interactifs » par Nadia BOUKHELIFA (INRAE, Université Paris-Saclay).
— 11h15 « Concevoir par des voies détournées : comment les designers parviennent à façonner les systèmes algorithmiques » par Nolwenn MAUDET (ACCRA, Université de Strasbourg).
— 11h45 Débat« Quelle est la place du design et des designers dans la fabrique des systèmes intelligents ? » : échange entre Jules FRANÇOISE, Nadia BOUKHELIFA et Nolwenn MAUDET, modéré par Téo SANCHEZ.
— 12h30 Buffet.
— 14h00 « Les traces d'interaction comme support d'explications » par Béatrice FUCHS (LIRIS, Université Jean Moulin Lyon III).
— 14h30 « Confiance humain-IA pour la prise de décision : définitions, facteurs et évaluation au travers de prisme académique et industriel » par Oleksandra VERESCHAK (Sorbonne Université).
— 15h00 « Analyse et coordination de l'automatisation de décisions pour les professionnels » par Thomas BAUDEL (IBM, Université Paris-Saclay).
— 15h30 Débat « L'explicabilité : condition nécessaire ou suffisante pour mériter la confiance des utilisateurs ? » : échange entre Béatrice FUCHS, Oleksandra VERESCHAK et Thomas BAUDEL, modéré par Benoît LE BLANC.
— 16h00 Posters et démos.
— 17h30 Clôture.

Organisation

Cet événement est organisé par Benoît LE BLANC et Eya BEN CHAABEN pour l'AFIA, et Téo SANCHEZ pour l'AFIHM.

Inscriptions

Les inscriptions à la journée (gratuites pour les membres AFIA ou AFIHM, 30 € sinon) sont obligatoires et à effectuer sur le site : https://www.linscription.com/pro/activite.php?P1=135557. Le déjeuner est garanti à toutes les personnes qui se seront inscrites avant le 2 juin 2023, et seulement à celles-ci.

Jeux & IA 2023

L'Association Française pour l'Intelligence Artificielle (AFIA) et le groupe de travail Jeux et Planification Multi-Agents, Flexible, Temporelle, Épistémique et Contingente (MAFTEC) du groupement de recherche Raisonnement, Apprentissage, et Décision en Intelligence Artificielle (GDR RADIA) organisent conjointement une quatrième journée commune « Jeux et IA » au sein de la Plate-Forme Intelligence Artificielle 2023 (PFIA 2023).

Date et lieu

— Date : 3 juillet 2023
— Lieu : PFIA 2023, Strasbourg

Programme

— 10h30 Présentation des Groupes de Travail.
— 10h45 « Vers plus de raisonnement dans EL-O : l'exemple de Hanabi » par Elise PERROTIN (CRIL).
— 11h10 « Évaluation de méthodes d'XAI diverses sur une tâche de pronostic d'e-sport » par Corentin BOIDOT (ENIB, Lab-STICC).
— 11h35 « Prolog et ontologies, une autre approche pour les comportements des PNJ » par Sylvain LAPEYRADE (Université Clermont-Auvergne, LIMOS).
— 12h00 Repas.
— 14h40 « Perspectives sur l'automatisation de l'évaluation de l'expérience de jeu » par Thomas CONSTANT (CNAM, CEDRIC).
— 15h05 « Modélisation récursive d'opposants dans les jeux à information incomplète » par Junkang LI (NukkAI et GREYC, Université de Caen Normandie).
— 15h30 Discussion.
— 16h10 Clôture.

Organisation

Cet événement est organisé par Anne-Gwenn BOSSER (Lab-STICC, Ecole Nationale d'Ingénieur de Brest) et Tristan CAZENAVE (LAMSADE, Université Paris-Dauphine) pour le Conseil d'Administration de l'AFIA, et Tiago DE LIMA (CRIL, Université d'Artois) et Bruno ZANUTTINI (GREYC, Université Caen Normandie) pour le GT MAFTEC.

Santé & IA 2023

L'Association Française pour l'Intelligence Artificielle (AFIA) et l'Association française d'Informatique Médicale (AIM) organisent conjointement une sixième journée commune « Santé et IA » au sein de PFIA 2023 afin de faire un point sur les travaux actuels en ingénierie des connaissances pour la santé.

Date et lieu

— Date : 6 Juillet 2023
— Lieu : PFIA 23, Strasbourg

Programme

— 10h40 « Conception et visualisation d'un graphe de connaissances à partir de données en hématologie : Prise en charge de l'anémie chez l'adulte » par Soulaymane HODROJ, Sylvie DESPRES, Jérôme NOBECOURT, Florence CYMBALISTA.
— 11h10 « Graphe de connaissance et ontologie pour la représentation des données de la LLC » par Chiraz PIRIOU, Sylvie DESPRES, Jérôme NOBECOURT, Claudine IRLES, Christine LE ROY.
— 11h25 « *Enhancing a Biomedical Ontology with Knowledge from Discharge Summaries* » par Sylvie DESPRES, Catherine DUCLOS, Chan LE DUC, Pascal VAILLANT.
— 11h40 « *Weak Controllability of multi-agent plans with uncertainty : towards temporal flexibility negotiation* » par Ajdin SUMIC, Thierry VIDAL, Mohamed HEDI KARRAY.
— 11h55 « *Towards Trustworthy-AI-by-Design Methodology for Intelligent Radiology Systems* » par Clotilde BRAYE, Jérémy CLECH, Arnaud GOTLIEB, Nadjib LAZAAR, Patrick MALLEA.
— 12h10 Repas.
— 14h40 « Détection automatique des pics d'un signal de pression intracrânienne : comparaison d'algorithmes combinant apprentissage profond et fonction de courbure » par Donatien LEGÉ, Marion PRUD'HOMME, Julien HENRIET.
— 15h10 « Analyse automatique de négations pour la radiologie et autres textes cliniques en français par modèles de langage » par Salim SADOUNE, Antoine FRABOULET, Antoine RICHARD, François TALBOT, Loïc BOUSSEL, Hugues BERRY.
— 15h40 « Anonymisation de documents médicaux en textes libres et en français via réseaux de neurones » par Antoine RICHARD, François TALBOT, David GIMBERT.
— 16h10 Pause.
— 16h30 « Un système d'aide au dialogue en santé intime des femmes » par Xingyu LIU, François PORTET, Didier SCHWAB.
— 16h45 « *Automatic detection of schwa in French hypersomniac patients using Automatic Speech Recognition* » par Colleen BEAUMARD, Vincent MARTIN, Yaru WU, Jean-Luc ROUAS, Pierre PHILIP.
— 17h15 « Explorer des mentions d'interventions non médicamenteuses dans des données issues des médias sociaux » par Alexis DELAFORGE, Jérôme AZÉ, Sandra BRINGAY, Caroline MOLLEVI, Arnaud SALLABERRY, Maximilien SERVAJEAN.
— 17h45 « *Heterogeneous incomplete multi-view data for Neurotoxicity biomarkers Identification* » par Quentin RUIN, David BALAYSSAC, Issam FALIH, Engelbert MEPHU NGUIFO.
— 18h00 Clôture.

Organisation

Fleur MOUGIN (U. de Bordeaux) et Lina SOUALMIA (Normandie Universités).

AfIA
Association française
pour l'Intelligence Artificielle

cnrs GDR Groupement de recherche
RADIA Raisonnement, Apprentissage, et Décision en Intelligence Artificielle

Modèles hybrides & IA 2023

L'Association Française pour l'Intelligence Artificielle (AFIA) et le groupe de travail Modèles Hybrides d'IA (MHyIA) du groupement de recherche Raisonnement, Apprentissage, et Décision en Intelligence Artificielle (GDR RADIA) organisent conjointement une 1re journée commune « Modèles hybrides & IA » sur le thème « IA neuro-symbolique » au sein de la conférence « *European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty* » (ECSQARU 2023).

Date et lieu

— Date : 19 septembre 2023
— Lieu : la salle des colloques, Bâtiment "La Maison de la Recherche", Université d'Artois, Rue Maurice Schuman, 62000 Arras

Programme

— 09h00 Ouverture de Fatiha SAÏS (représentant l'AFIA) et Zied BOURAOUI (représentant le GDR RADIA).
— 09h15 « *Aligning embeddings with symbolic knowledge : towards a tight integration of learning and reasoning* » par Steven SCHOCKAERT (Cardiff University).
— 10h15 Pause.
— 10h30 « *Interpretable Neural-Symbolic Concept Reasoning* » par Giuseppe MARRA (KU Leuven).
— 11h30 « *Integrating Combinatorial Solvers and Neural Models* » par Pasquale MINERVINI (University of Edinburgh).
— 12h30 Clôture.

Organisation

Cet événement est organisé par Zied BOURAOUI (CRIL, Université d'Artois), Pierre MONNIN (I3S, Université Côte d'Azur) et Fatiha SAÏS (LISN, Université Paris Saclay).

Inscriptions

La participation est gratuite mais l'inscription est obligatoire, à effectuer à partir de ce lien, pour recevoir le lien Zoom de la demi-journée.

EGC & IA 2024

L'Association Française pour l'Intelligence Artificielle (AFIA) et l'association internationale francophone d'Extraction et de Gestion des Connaissances (Association EGC) organisent, avec l'aide du groupe de travail « Gestion et Analyse des données Spatiales et Temporelles » (GAST), une journée commune « Gestion et Analyse des données Aériennes et Satellitaires » (G2AS' 24) sur le thème de la représentation, de la Gestion, de l'Analyse et du stockage des données Aériennes et Satellitaires. Cette journée réunira les acteurs de la recherche académique ou industrielle autour de cette thématique.

Date et lieu

— Date : 17 avril 2024
— Lieu : EPITA, 14-16 rue Voltaire, 94270 Le Kremlin-Bicêtre

Programme

— 09h00 Mot d'accueil par Aurélie LEBORGNE (GT GAST), par Thierry GERAUD (LRE EPITA) et par Thomas GUYET (AFIA).
— 09h30 « Présentation du projet SESAME Surveillance de l'État de SAnté des Mangroves » par Franck NIVOLE (IRD Nouvelle-Calédonie, UMR Espace-DEV).
— 10h00 « Machine and deep learning for earth observation : advanced approaches and practical use cases » par Roberto INTERDONATO (CIRAD, UMR TETIS, INRIA, Montpellier).
— 10h30 Pause.
— 10h45 « Données satellitaires : de l'application locale à des modèles de fondation » par Sylvain LOBRY (LIPADE, Université Paris Cité).
— 11h15 « Interplay between data assimilation and Machine Learning for time series of satellite images » par Lucas DRUMET (Lab-STICC, IMT Atlantique).
— 11h45 Posters.
— 12h00 Pause.
— 13h30 Discussion Posters.
— 14h00 « Analyse d'images aériennes historiques : application à une étude épidémiologique » par Laure TOUGNE (LIRIS, Université Lumière Lyon 2).
— 14h30 « Utilisation de l'imagerie satellitaire – L'Espace au service de la Terre » par Henri GIRAUD (SERTIT, Université de Strasbourg).
— 15h00 « La Terre vue par la géodésie spatiale : des mesures massives pour une meilleure compréhension des processus géophysiques sous-jacents » par Samuel NAHMANI (UMR IPGP / Université Paris Cité, CNRS, IGN ; Université Gustave Eiffel, ENSG, IGN).
— 15h30 Pause.
— 15h45 « Apprentissage profond pour le traitement de séries temporelles d'images satellites » par Loïc LANDRIEU (IMAGINE – LIGM / École Nationale des Ponts et Chaussées).
— 16h15 « Traitement d'images de télédétection SAR et optiques pour la cartographie de l'occupation des sols » par Flora WEISSGERBER (ONERA/DTIS/SAPIA).
— 16h45 Clôture.

Organisation

Cet événement est organisé par Clément IPHAR (LETG, UBO, Brest), Guillaume TOCHON (LRE, EPITA, Kremlin-Bicêtre), Aurélie LEBORGNE (ICube, UNISTRA, Strasbourg), Loïc SALMON (ISEA, UNC, Nouméa, Nouvelle-Calédonie), et Nida MEDDOURI (LRE, EPITA, Kremlin-Bicêtre).

Humanités Numériques & IA 2024

L'Association Française pour l'Intelligence Artificielle (AFIA) et les groupes de travail Masses de Données, Informations et Connaissances en Sciences (MADICS) et Méthodes et Applications pour la Géomatique et l'Information Spatiale (MAGIS) organisent conjointement une première journée commune sur le thème « Humanité Numérique et IA ».

Le vocable humanités numériques s'est imposé pour désigner les travaux de recherche relevant tantôt : de la création, la gestion et la formalisation de processus sociaux à l'aide d'outils mathématiques et informatiques ; ou encore de la formaliser de processus humains. L'analyse automatique de documents anciens, de traitement automatique du langage naturel, de recherche d'informations ou encore de simulation, posent de véritables défis scientifiques aux approches développées dans le domaine de l'intelligence artificielle.

Date et lieu

— Date : 3 mai 2024
— Lieu : Datalab, BnF, Quai François Mauriac, 75706 Paris (aussi en distanciel)

Programme

— 09h15 Ouverture par Fatiha SAÏS (AFIA), Nathalie HERNANDEZ(GdR MADICS), Nathalie ABADIE (GdR MAGIS), Tiphaine VACQUE (BnF) et Marie CARLIN (Datalab, BnF).
— 09h55 « La politique dans la machine : identifier, mesurer et limiter l'information politique apprise par les algorithmes » par Tim FAVERJON(médialab Sciences Po) et Pedro RAMACIOTTI(médialab Sciences Po, ISC-PIF, INSHS-CNRS).
— 10h15 «ISIDORE 2030 : refactorisation d'un moteur de recherche à l'ère des IA de traitement et des IA génératives» par Stéphane POUYLLAU (Huma-Num)..
— 12h30 Clôture.

Organisation

Cet événement est organisé par Zied BOURAOUI (CRIL, Université d'Artois), Pierre MONNIN (I3S, Université Côte d'Azur) et Fatiha SAÏS (LISN, Université Paris Saclay).

Inscriptions

La participation est gratuite mais l'inscription est obligatoire, à effectuer à partir de ce lien, pour recevoir le lien Zoom de la demi-journée.

Santé & IA 2024

L'Association française d'Informatique Médicale (AIM) et le collège Science de l'Ingénierie des Connaissances de l'(AFIA) organisent une septième journée commune Santé et IA, inscrite au sein de la Plate-Forme Intelligence Artificielle 2024 (PFIA 2024).

Cette journée a pour objectif de faire un point sur les travaux menés actuellement en ingénierie des connaissances dans le domaine de la santé.

En effet, l'ingénierie des connaissances peut permettre de répondre aux enjeux majeurs tels que la progression du savoir médical, l'aide à la décision (qu'elle soit diagnostique, thérapeutique ou pronostique), et plus largement d'apporter des solutions permettant de favoriser l'accès aux informations et connaissances médicales. Ces méthodes peuvent être appliquées à de nombreux cas d'usage au service des patients (que ce soit à l'échelle individuelle ou d'une population), pour les professionnels de santé, étudiants en santé, chercheurs, décideurs et le grand public. Les données de santé ont de multiples caractéristiques qui soulèvent des problématiques liées à l'extraction d'information, à la sécurité des données à caractère personnel, à l'intégration de données réparties dans des systèmes hétérogènes, à la recherche d'information, au traitement de données massives et à la compréhension des données.

Articles et présentations sont disponibles sur cette page : afia.asso.fr/les-journees-communes/sante-et-ia-2024

Date et lieu

— Date : 1er Juillet 2024
— Lieu : PFIA 24, La Rochelle

Programme (session matinale)

— 9h30 « Chirurgie métabolique de précision » par François PATTOU *(conférence invitée)*(Université de Lille, Inserm, CHU Lille, Institut Pasteur de Lille, U1190-EGID).
— 10h30 Pause.
— 10h50 « Deep Reinforcement Learning for Controlled Piecewise Deterministic Markov Process in Cancer Treatment Follow-up » par Alice CLEYNEN, Benoîte DE SAPORTA, Orlane ROSSINI, Régis SABBADIN, Meritxell VINYALS.
— 11h15 « Contributions à l'Ordonnancement des Interventions en Chirurgie Ambulatoire : Q-learning et Flow-Shop Hybride » par Lydia BOUCHLAGHEM, Adnène GUESSOUM, Fatima GHEDJATI.
— 11h30 « Amélioration de la sécurité chirurgicale avec un jumeau numérique prédictif : le rôle des systèmes multi-agents et de l'apprentissage par renforcement » par Bruno PEREZ.
— 11h45 « Interpretable AI for Dermoscopy Images of Pigmented Skin Lesions » par Marianne DEFRESNE, Élise COUTIER, Paul FRICKER, Folkert BLOK, Hang NGUYEN.
— 12h10 « INM-Explain – Expliquer les controverses médicales : Application au cas des interventions non médicamenteuses » par Houria SAYAH, Alya ZOUZOU, Jonathan DUCKES, Audric GIRONDIN, Maéva MAÏO, Maximilien SERVAJEAN, Sandra BRINGAY.
— 12h30 Pause.

Programme (session post-méridienne)

— 14h50 « Une Ontologie du Parcours de Soins » par François-Élie CALVIER, Thomas GUYET, Nolwenn LEMEUR-ROUILLARD.

— 15h15 « Entrepôts de Données de Santé et Protection de la Vie Privée : Synthèse de discussions Inter-CHU » par Antoine RICHARD, Manal AHIKKI, Marc BERARD, Camille BOIN, Antoine BOUTET, Stéphane BREANT, Alice CALLIGER, Ariel COHEN, Jean-François COUCHOT, Denis DELAMARRE, Caroline DUNOYER, Thibaut FABACHER, Lucas GAUTHIER, David GIMBERT, Camille GIRARD-CHANUDET, Faustine GLAIS, Romain GRIFFIER, Martin HILKA, Yannick JACOB, Vianney JOUHET, David LAIY-MANI, Leonardo MOROS, Joris MULLER, David PELLECUER, Thomas PETIT-JEAN, Maxime SALAUN, François TALBOT, Perceval WAJSBURT, Kevin YAUY.

— 15h30 « Chain Classifier pour le transcodage LOINC » par Théodore MICHEL-PICQUE, Sandra BRINGAY, Pascal PONCELET, Namrata PATEL, Guilhem MAYORAL.

— 15h45 « Is DTW resilient to noise and effective for EEG functional connectivity assessment ? » par Maxime BEDOIN, Nesma HOUMANI, Rita YABOURI, Jérôme BOUDY, Kiyoka KINUGAWA.

— 16h00 Pause.

— 16h20 « Récentes avancées de l'inférence en langue naturelle pour les essais cliniques » par Mathilde AGUIAR, Pierre ZWEIGENBAUM, Nona NADERI.

— 16h45 « Des pipelines faciles à réutiliser pour comparer les performances d'outils de reconnaissance d'entités nommées sur les textes cliniques en français » par Thibault HUBERT, Ghislain VAILLANT, Olivier BIROT, Camila ARIAS, Antoine NEURAZ, Bastien RANCE, Adrien COULET.

— 17h10 « Équilibrer qualité et quantité : comparaison de stratégies d'annotation pour la reconnaissance d'entités nommées en cardiologie » par Virgile BARTHET, Laura MONCEAUX-CACHARD, Christine JACQUIN, Cyril GROUIN, Joconde WELLER, Pascal DEGroote, Emmanuel MORIN, Pierre ZWEIGENBAUM.

Organisation

Cet événement est organisé par Adrien COULET (HeKA, Inria Paris & Inserm, Université Paris Cité), Fleur MOUGIN (AHeaD, Université de Bordeaux & Inserm) et Lina SOUALMIA (LITIS & LIMICS, Normandie Universités & Inserm).

Société & IA 2024

L'AFIA, le groupe de travail Aspects Computationnels de l'Éthique (ACE) du GDR RADIA et Inria Bordeaux ont organisé les journées « Société et IA » les 1er et 2 juillet 2024 dans le cadre de la Plate-Forme Intelligence Artificielle de l'AFIA (PFIA 2024).

Depuis plusieurs années, des comités réunis à l'initiative d'université, d'États, de puissances supra-étatiques comme la Commission Européenne, de sociétés savantes ou d'organisation non gouvernementales réfléchissent aux questions d'éthique de l'Intelligence Artificielle et à sa régulation. Ces réflexions ont abouti entre autres sur la notion de systèmes informatiques dignes de confiance qui sont à mettre en perspective avec les problématiques en éthique artificielle.

Ces journées avaient pour objectif de réunir les communautés travaillant sur l'Intelligence Artificielle de confiance, l'éthique artificielle et plus généralement sur tout ce qui est en lien avec l'impact social de l'Intelligence Artificielle. Dans une volonté d'ouverture tant aux communautés de recherche travaillant déjà sur ces problématiques, qu'aux non-spécialistes intéressés, nous avons encouragé toutes les contributions relatives à ces sujets, qu'elles portent sur les aspects techniques, juridiques, philosophiques ou sociologiques de l'Intelligence Artificielle ou sur les impacts industriels de son déploiement.

Date et lieu

— Date : 1-2 Juillet 2024
— Lieu : PFIA 24, La Rochelle

Programme (Lundi 1er Juillet 2024)

— 14h50 « Ouverture des journées » par Frédéric ALEXANDRE, Grégory BONNET, Ikram CHRAIBIKaadoud, Jean-Gabriel GANASCIA.
— 15h00 « Un besoin de Confiance Artificielle pour l'Intelligence Artificielle » *(conférence invitée)* par Laurent SIMON.
— 16h00 Pause.
— 16h20 « Détection de biais et intégration de connaissances expertes pour l'explicabilité en IA » par Matthieu DELAHAYE, Lina FAHED, Florent CASTAGNINO et Philippe LENCA.
— 16h50 « Modéliser la confiance d'un agent décisionnel » par Baptiste PESQUET et Frédéric ALEXANDRE.
— 17h20 « L'explicabilité appliquée aux modèles de diffusion » par Raphael TEITGEN, Jeanine HARB et Jeanne LE PEILLET.

Programme (Mardi 2 Juillet 2024)

- 09h00 « La réglementation de l'intelligence artificielle dans l'Union européenne » *(conférence invitée)* par Nathalie NEVEJANS.
- 10h00 Pause.
- 10h20 « La normalisation de l'IA : un déluge de réinterprétations de l'AI Act » par Hélène HERMAN et Mélanie GORNET.
- 10h50 « IA générative et désinformation : quel impact sur les rapports de force en géopolitique ? » par Alice MARANNE, Clara FONTAINE-SAY et Ikram CHRAIBIKaadoud.
- 11h20 « Quel sens donner à l'IA de confiance ? » *(conférence invitée)* par Cédric BRUN.
- 12h30 Pause.
- 14h50 « L'intelligence artificielle à la lumière de la mythologie grecque : rendre compréhensible les impacts de l'IA pour le grand public » par Fabrice MUHLENBACH.
- 15h20 « Cadre conceptuel pour agents autonomes éthiques : application aux agents conversationnels » par Robert VOYER, Thierno TOUNKARA..
- 16h00 Pause.
- 16h20 « Définition de la compatibilité pour des préférences morales : une condition basée sur la cohérence de Suzumura » par Guillaume GERVOIS, Gauvain BOURGNE et Marie-Jeanne LESOT.
- 16h50 « Modèle d'éthique pour les MDP multi-agents » par Mihail STOJANOVSKI, Nadjet BOURDACHE, Grégory BONNET et Abdel-Illah MOUADDIB.
- 17h20 « Équité subjective par les explications » par Sarra TAJOURI et Alexis TSOUKIÀS..

Organisation

Cet événement est organisé par Frédéric ALEXANDRE (Inria Bordeaux), Grégory BONNET (GREYC, Université de Caen), Ikram CHRAIBI KAADOUD (Inria Bordeaux), Jean-Gabriel GANASCIA (LIP6, Sorbonne Université)

AfIA
Association française
pour l'Intelligence Artificielle

Agent & IA 2024

L'AFIA, le groupe de travail Affects, Compagnons Artificiels et Interactions (ACAI) et la communauté des Journées Francophones des Systèmes Multi-Agents (JFSMA) ont organisé la journée « Agent et IA » le 4 juillet 2024 dans le cadre de la Plate-Forme Intelligence Artificielle de l'AFIA (PFIA 2024).

Cette journée a porté sur les agents et avait pour thème l'humain dans la boucle. Elle était composée de deux parties :

- la matinée a été dédiée à un tutoriel animé par le groupe de travail ACAI : Affects, Compagnons Artificiels et Interactions, avec un focus sur l'interaction humain-agent ;
- l'après-midi a été dédiée à la présentation de systèmes multi-agents dans le domaine des *Smart-Cities*, avec un focus sur l'aspect mobilité.

Cette journée s'est voulue accessible et a visé également un objectif pédagogique, vous donnant quelques pistes pour développer vous-mêmes vos agents logiciels.

Date et lieu

- Date : 4 Juillet 2024
- Lieu : PFIA 24, La Rochelle

Programme (session ACAI)

- 10h20 « Interaction entre une personne et un ou plusieurs agent(s) conversationnel(s) animé(s) » par Brian RAVENET et Nicolas SABOURET.
- 12h30 Pause.

Programme (session JFSMA)

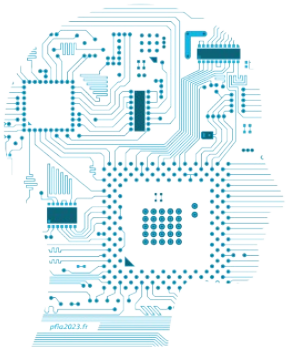
- 14h50 « Travaux et plateformes de la communauté Mobilité Intelligente de l'IMT » par Arnaud DONIEC et Flavien BALBO.
- 15h20 « Architecture et comportements pour la simulation multiagent de véhicules et de l'infrastructure connectée en condition atmosphérique difficiles » par Stéphane GALLAND.
- 15h40 « Vers des véhicules autonomes socialement désirables » par Joris DINNEWETH.
- 16h00 Pause.
- 16h20 « Résolution de conflits entre trajectoires de vol planifiées (contrats 4D) de drones dans le trafic urbain » par Gauthier PICARD.
- 16h40 « Projet VILAGIL et action VILAGIL-MaaS » par Valérie CAMPS et Elsy KADDOUM.
- 17h20 « Projet autOCampus » par Marie-Pierre GLEIZES.
- 17h40 Table ronde.

Organisation

Cet événement est organisé par Valérie CAMPS (Universite Paul Sabatier), Elsy KADDOUM (Université Toulouse 2 Jean-Jaurès), Brian RAVENET (IUT Orsay) et Nicolas SABOURET (Université Paris-Saclay).

PFIA 2023

Plate-forme Intelligence Artificielle



Strasbourg

03-07 juillet 2023

La rencontre annuelle de la communauté Intelligence Artificielle autour de conférences et ateliers thématiques

pfia2023.pfia.fr

PROGRAMME

APIA : Applications Pratiques de l'Intelligence Artificielle
CAp : Conférence sur l'Apprentissage automatique
CNIA : Conférence Nationale en Intelligence Artificielle
IC : Journées francophones d'Ingénierie des Connaissances
JFPC : Journées Francophone de Programmation par Contraintes
JFSMA : Journées Francophones sur les Systèmes Multi-Agents
JIAF : Journées d'Intelligence Artificielle Fondamentale
RJCIA : Rencontres des Jeunes Chercheurs en Intelligence Artificielle
SFC : Rencontres de la Société Francophone de Classification

Tutoriels

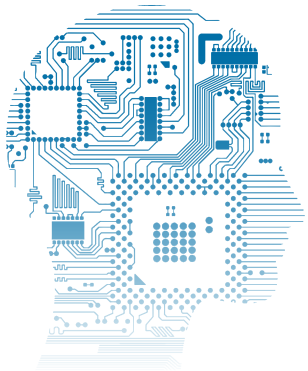
Journée thématique **ACAI** : Affects, Compagnons Artificiels et Interactions
Journée thématique **Jeux & IA**
Journée thématique **Santé & IA**
Journée thématique **Résilience & IA**

Soirée IA en Alsace

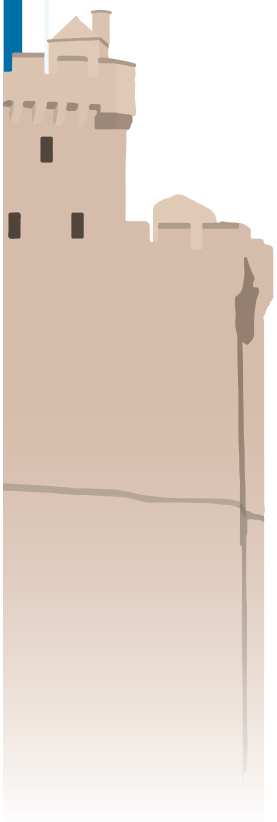
PFIA 2024

Plate-forme Intelligence Artificielle

La rencontre annuelle de la communauté **Intelligence Artificielle** autour de conférences et ateliers thématiques



LA ROCHELLE
01-05 JUILLET 2024



pfia2024.univ-lr.fr

PROGRAMME

CONFÉRENCES

APIA - Applications Pratiques de l'Intelligence Artificielle

CNIA - Conférence Nationale en Intelligence Artificielle

IC - Journées francophones d'Ingénierie des Connaissances

JIAF - Journées d'Intelligence Artificielle Fondamentale

RJCIA - Rencontres des Jeunes Chercheurs en Intelligence Artificielle

JOURNÉES

Agents & IA

Santé & IA

Société & IA

ATELIERS

Défense et IA

Jeux et IA

MAFTEC

SOSEM

CÉCILIA

IA en Nouvelle-Aquitaine

GdR RADIA

TUTORIELS

8 tutoriels répartis sur 5 jours

En partenariat avec

L'AVENIR EST AUX VALEURS SÛRES