

## Journée d'étude

Université Bourgogne Europe, CNRS, LEAD UMR5022

Patrick Bard, Michel Paindavoine & Fan Yang-Song



# DE L'IA BIO-INSPIRÉE À L'IA FRUGALE



### Marina REYBOZ

CEA-LIST à Grenoble

Les sciences cognitives pour une IA soutenable

### Martial MERMILLOD

Laboratoire de Psychologie et NeuroCognition, Université Grenoble Alpes

L'intelligence artificielle pour les sciences cognitives

### Maxime AMBARD

Université Bourgogne Europe, CNRS, LEAD UMR5022, Dijon

L'intelligence artificielle pour la remédiation du handicap

### Constance DOUWES

LIS, CNRS UMR7020 / AMU / UTLN, Aix Marseille Université

Energy and environmental impacts of deep learning



### Nasser DANDANA

LISPEN Chalon-sur-Saône & Université Bourgogne Europe, CNRS, LEAD UMR5022, Dijon & CHU Dijon Bourgogne

Modèles de machine learning pour la transformation, segmentation et visualisation 3D d'images D-FF-OCT de biopsies rénales

### Timothée MASQUELIER

CerCo - CNRS UMR5549, Toulouse

Learning delays with backpropagation in spiking neural networks

### Benoit MIRAMOND

LEAT, CNRS UMR7248, Université Côte d'Azur

All in one timestep: Enhancing sparsity and energy efficiency in multi-level spiking neural networks

### Florian FIZAINÉ

Archives départementales de la Côte d'Or & Université Bourgogne Europe, CNRS, LEAD UMR5022, Dijon

Une nouvelle approche de la transcription de textes historiques par réseaux de neurones impulsionnels



# INTRODUCTION

Depuis quelques années, l'intelligence artificielle (IA) a fait des progrès majeurs au point d'être maintenant un fait de société aux implications considérables. Si un grand nombre d'entre elles apportent un réel gain au service de l'Humain, l'intelligence artificielle, de par son empreinte écologique et plus particulièrement sa consommation énergétique, interroge au regard de la transition énergétique nécessaire depuis le début du XXIème Siècle.

Diverses approches émergent pour proposer une IA frugale. Une solution pourrait bien être celle de s'inspirer du vivant. En effet, avec une consommation d'à peine 30 W, soit plus de 10 fois moins qu'une carte graphique couramment utilisée dans les ordinateurs dédiés au calcul, le cerveau humain est capable d'une performance globale encore inégalée par les meilleurs modèles d'intelligence artificielle.

Cette journée d'étude se propose de s'intéresser de façon plus précise au coût énergétique de l'IA, d'étudier plus en profondeur les liens entre IA bio-inspirée et IA frugale, de présenter un paradigme incontournable en termes d'IA bio-inspirée, à savoir les réseaux de neurones impulsionnels, mais aussi de présenter quelques applications de l'IA au bénéfice de la société.





## DE L'IA BIO-INSPIRÉE À L'IA FRUGALE

8h00 : Inscriptions, accueil café

8h30-8h45 :

Présentation de la journée par Michel Paindavoine, Patrick Bard & Fan Yang-Song

08h45-9h30 : **Marina REYBOZ**

CEA-LIST à Grenoble

Les sciences cognitives pour une IA soutenable

9h30-10h15 : **Martial MERMILLOD**

Laboratoire de Psychologie et NeuroCognition, Université Grenoble Alpes

L'intelligence artificielle pour les sciences cognitives

10h15-10h30 : Pause

10h30-11h00 : **Maxime AMBARD**

Université Bourgogne Europe, CNRS, LEAD UMR5022, Dijon

L'intelligence artificielle pour la remédiation du handicap

11h00-11h45 : **Constance DOUWES**

LIS, CNRS UMR7020 / AMU / UTLN, Aix Marseille Université

Energy and environmental impacts of deep learning

11h45-12h15 : **Nasser DANDANA**

LISPEN Chalon-sur-Saône & Université Bourgogne Europe, CNRS, LEAD UMR5022, Dijon & CHU Dijon Bourgogne

Modèles de machine learning pour la transformation, segmentation et visualisation

3D d'images D-FF-OCT de biopsies rénales

12h15-13h15 : Pause déjeuner

13h15-14h00 : **Timothée MASQUELIER**

CerCo - CNRS UMR5549, Toulouse

Learning delays with backpropagation in spiking neural networks

14h00-14h45 : **Benoit MIRAMOND**

LEAT, CNRS UMR7248, Université Côte d'Azur

All in one timestep: Enhancing sparsity and energy efficiency in multi-level spiking neural networks

14h45-15h15 : **Florian FIZAINE**

Archives départementales de la Côte d'Or & Université Bourgogne Europe, CNRS, LEAD UMR5022, Dijon

Une nouvelle approche de la transcription de textes historiques par réseaux de neurones impulsifs

15h15-15h30 : Conclusion





# Marina REYBOZ

CEA-LIST à Grenoble

## Les sciences cognitives pour une IA soutenable

*Marina Reyboz est directrice de recherche en intelligence artificielle au CEA-LIST à Grenoble. Ses travaux portent sur l'apprentissage profond embarqué, avec le développement et la co-optimisation d'algorithmes innovants inspirés du fonctionnement du cerveau. La question centrale de ses recherches est la suivante : « Que peuvent nous apporter les connaissances du cerveau humain pour concevoir des circuits intelligents à très faible consommation énergétique? ». Ses activités s'inscrivent également dans une démarche d'éco-innovation et de sobriété numérique, notamment à travers le développement de réflexion autour de l'IA pour la soutenabilité.*

L'essor rapide de l'intelligence artificielle pose des défis éthiques et environnementaux majeurs. Pour y répondre, le développement de modèles inspirés du fonctionnement du cerveau apparaît comme une voie prometteuse. Dans les travaux que nous évoquerons, nous nous inspirons du fonctionnement de la mémoire humaine pour privilégier l'adaptabilité des modèles qui permet leur réutilisation ainsi que l'apprentissage local qui permet la protection des données et une meilleure efficacité énergétique. Le développement de ces méthodes est spécifié pour des applicatifs dans le domaine de la santé et aidant à la transition écologique.



## Martial MERMILLOD

Laboratoire de Psychologie et NeuroCognition, Université Grenoble Alpes  
**L'intelligence artificielle pour les sciences cognitives**

*Martial Mermillod est Professeur des Universités à l'UGA en Sciences Cognitives, Intelligence Artificielle et Neurosciences Cognitives. Ses recherches visent à utiliser notre compréhension du cerveau et de la cognition humaine pour améliorer l'IA fondée des réseaux neuronaux artificiels (Deep Learning). Ses travaux se concentrent actuellement sur l'utilisation de l'intelligence humaine et artificielle pour accélérer la transition écologique et permettre de répondre aux limites planétaires. Plus d'information ici: <https://lpnc.univ-grenoble-alpes.fr/fr/martial-mermillod>.*

Les réseaux de neurones convolutifs à l'origine de la nouvelle révolution de l'IA sont inspirés de l'organisation du cerveau humain. Nous verrons l'origine de cette révolution, les succès de l'IA mais aussi les dangers et limites de l'IA que l'on peut dépasser par la bio-inspiration. Réciproquement, ces outils d'IA offrent aujourd'hui un cadre puissant pour mieux comprendre la cognition humaine. En psychologie cognitive, le deep learning permet de générer des cartes de saillance qui indiquent quelles régions d'un stimulus visuel attirent l'attention ou influencent la prise de décision. Appliqués aux neurosciences, ces modèles permettent le brain decoding qui vise à déterminer la fonction cognitive réalisée à partir d'enregistrements cérébraux (IRMf, EEG, sEEG, MEG). Ces méthodes permettent de largement dépasser les techniques d'analyses de données actuellement utilisées en psychologie et neurosciences cognitives (modèle linéaire général, machine learning). Cette double interaction — l'IA inspirée du cerveau et l'IA utilisée pour analyser le cerveau — constitue aujourd'hui des méthodes innovantes pour comprendre les bases computationnelles de la cognition humaine.



# Maxime AMBARD

Université Bourgogne Europe, CNRS, LEAD UMR5022, Dijon

## L'intelligence artificielle pour la remédiation du handicap

*Maxime Ambard est Maître de Conférences des Universités en informatique au LEAD CNRS UMR5022 et à l'IUT de Dijon. Ses thèmes de recherche concernent la conception et la réalisation de dispositifs d'aide aux personnes porteuses de handicap, principalement les personnes mal ou non-voyantes.*

Les algorithmes d'intelligence artificielle offrent de vastes possibilités pour aider les personnes ayant des handicaps. Ces algorithmes peuvent être tunés, combinés et embarqués afin d'offrir des systèmes facilitant par exemple les apprentissages moteurs ou la perception de l'environnement extérieur.

Nous verrons leurs applications au travers de 2 projets menés au laboratoire LEAD CNRS UMR5022 en lien avec le handicap.



# Constance DOUWES

LIS, CNRS UMR7020 / AMU / UTLN, Aix Marseille Université

## Energy and environmental impacts of deep learning

*Constance Douwes a soutenu sa thèse en 2023 à l'Ircam, dans le laboratoire STMS, où elle a étudié la consommation énergétique des modèles d'IA pour la génération audio. Elle a ensuite fait un postdoctorat à l'Inria Nancy, où elle a poursuivi ses recherches sur ces thématiques dans le cadre du challenge DCASE de détection et classification d'événements sonores. Elle est maintenant MCF en informatique à Centrale Méditerranée depuis la rentrée, et membre de l'équipe Qarma du laboratoire LIS à Marseille, où elle se spécialise dans les impacts environnementaux de l'intelligence artificielle moderne.*

In recent years, artificial intelligence based on deep learning models has become increasingly widespread across a wide range of applications, achieving ever-greater performance with ever-increasing computational cost. This growth drives higher energy consumption and carbon emissions, running counter the emission reduction pathways recommended by the IPCC. In this presentation, I will present and explore the environmental impacts associated with developing, training, and deploying deep learning models using the Life-Cycle Assessment methodology. I aim to provide useful methods for estimating the impacts and to raise awareness of the environmental challenges posed by the rapid growth of AI, thereby encouraging more efficient and responsible practices.



## Nasser DANDANA

LISPEN Chalon-sur-Saône & Université Bourgogne Europe, CNRS, LEAD  
UMR5022, Dijon & CHU Dijon-Bourgogne

### Modèles de machine learning pour la transformation, segmentation et visualisation 3D d'images D-FF-OCT de biopsies rénales nels

*Nasser Dandana est doctorant au Laboratoire d'Ingénierie des Systèmes Physiques Et Numériques (LISPEN) de Chalon-sur-Saône en collaboration avec le LEAD et le CHU Dijon - Bourgogne. Ses directeurs de thèses sont Frédéric Mérienne du LISPEN, Manon Ansart du LEAD et Mathieu Legendre, médecin au CHU Dijon-Bourgogne et chercheur associé au LEAD. Sa thèse porte sur de l'IA générative pour la médecine.*

Les maladies rénales chroniques sont habituellement considérées comme l'une des principales causes de mortalité à travers le monde. Leur diagnostic repose sur l'analyse microscopique des anomalies de la structure du rein après biopsie rénale, appelée analyse histopathologique, permettant alors la sélection du traitement le plus adapté à la pathologie identifiée. Cependant, dans sa version conventionnelle, cette analyse microscopique du tissu rénal nécessite plusieurs étapes de colorations et traitements susceptibles de retarder (jusqu'à quelques jours) le rendu diagnostique, et l'initiation du bon traitement.

Ainsi, l'ARMSL (Association pour la Recherche Médicale en Saône-et-Loire) a récemment participé au lancement d'un projet de recherche clinique visant à l'utilisation d'une nouvelle technique d'imagerie appelée D-FF-OCT (conjointement déployée au Centre Hospitalier Universitaire (CHU) Dijon Bourgogne, au centre Hospitalier William Morey (Chalon-sur-Saône) et au Centre Hospitalier de Mâcon) pour l'analyse ultra-rapide des biopsies rénales, permettant ainsi l'enregistrement des premières images de biopsies en moins d'une heure, cela pour les patients transplantés et non transplantés du rein.

Ce projet vise à aider les professionnels de santé dans l'interprétation des images D-FF-OCT grâce à des algorithmes de machine learning capables de les transformer et de les segmenter. Un outil de visualisation 3D des résultats viendra compléter cette approche afin de renforcer l'aide au diagnostic.



# Timothée MASQUELIER

CerCo - CNRS UMR5549, Toulouse

## Learning delays with backpropagation in spiking neural networks

*Tim Masquelier is a CNRS Research Director in AI & Computational Neuroscience. He works at the Centre de Recherche Cerveau et Cognition (CERCO) in Toulouse, where he leads the NeuroAI team, which embodies the cross-pollination between neuroscience and AI. He has been doing research on spiking neural networks for more than 20 years and published 50+ peer-reviewed paper on this topic.*

*Spiking neural networks (SNNs) are appealing for AI because they are particularly energy-efficient when implemented on event-driven neuromorphic chips. Interest in them has surged recently due to a major breakthrough, known as surrogate gradient learning (SGL), which enables training them using backpropagation and thus solving real-world problems with SNNs. In my group, we have demonstrated that SGL enables the learning of not only connection weights but also connection delays. These delays represent the time needed for one spike to travel from the emitting to the receiving neurons. Delays matter because they shift the spike arrival times, which should be synchronous to trigger an output spike. Thus, plastic delays significantly enhance the expressivity of SNNs. If this fact is well established theoretically, efficient algorithms to learn these delays have been lacking. We proposed a family of such algorithms, based on differentiable interpolation techniques, which outperform previous proposals. Our results show that learning delays, in addition to the weights, in fully connected, convolutional, and recurrent spiking layers strongly increases the accuracy on several temporal vision and audition tasks, leading to new state-of-the-arts.*

*Our approach is already highly relevant to the neuromorphic engineering community, as most existing digital neuromorphic chips (e.g., Intel Loihi, IBM True North, Spinnaker, SENECA) feature programmable delays, and many analog chips can also implement delays using circuit physics (e.g., BrainScaleS-2, DYNAP-SE2). In the long term, our research should also shed light on the role of myelin plasticity in the brain, which tunes the conduction velocities and thus the delays.*



# Benoit MIRAMOND

LEAT, CNRS UMR7248, Université Côte d'Azur

## All in one timestep: Enhancing sparsity and energy efficiency in multi-level spiking neural networks

*Benoit Miramond is Full Professor at LEAT lab from Université Côte d'Azur. Pr. Miramond is head of the eBRAIN research group working on embedded Bio-inspiRed Artificial Intelligence and Neuromorphic architectures. He holds the Chair of Bio-Inspired AI at the 3IA Côte d'Azur Institute. Pr. Miramond is head of research and innovation of Polytech Nice Sophia and research director at AICO Technology.*

*Spiking Neural Networks (SNNs) are one of the most promising bio-inspired neural networks models and have drawn increasing attention in recent years. The event-driven communication mechanism of SNNs allows for sparse and theoretically low-power operations on dedicated neuromorphic hardware. However, the binary nature of instantaneous spikes also leads to considerable information loss in SNNs, resulting in accuracy degradation. To address this issue, we propose a multi-level spiking neuron model able to provide both low-quantization error and minimal inference latency while approaching the performance of full precision Artificial Neural Networks (ANNs). Experimental results with popular network architectures and datasets, show that multi-level spiking neurons provide better information compression, allowing therefore a reduction in latency without performance loss. When compared to binary SNNs on image classification scenarios, multi-level SNNs indeed allow reducing the energy consumption depending on the number of quantization intervals. On neuromorphic data, our approach allows us to drastically reduce the inference latency to 1 timestep, which corresponds to a compression factor of 10 compared to previously published results. At the architectural level, we propose a new residual architecture that we call Sparse-ResNet. Through a careful analysis of the spikes propagation in residual connections we highlight a spike avalanche effect, that affects most spiking residual architectures. Using our Sparse-ResNet architecture, we can provide state-of-the-art accuracy results in image classification while reducing by more than 20% the network activity compared to the previous spiking ResNets. We conclude on the future integration of this execution model into the SPLEAT neuromorphic architecture developed in the eBRAIN research group from LEAT Lab.*



## Florian FIZAINE

Archives départementales de la Côte d'Or & Université Bourgogne Europe,  
CNRS, LEAD UMR5022, Dijon

### Une nouvelle approche de la transcription de textes historiques par réseaux de neurones impulsionnels

*Florian Fizaine a réalisé sa thèse au LEAD en collaboration avec les Archives départementales de la Côte d'Or sous la direction de Michel Paindavoine, Annie Vinter et moi-même. Elle portait sur la transcription de textes historique par Intelligence Artificielle. Florian et maintenant post-doctorant au LEAD en collaboration avec les Archives départementales sur le même sujet.*

L'arrivée des Transformers, et plus récemment des LLM multimodaux, a permis des avancées majeures dans la reconnaissance de caractères manuscrits grâce à la robustesse de leurs modèles de langue et à leur capacité à traduire des caractéristiques visuelles en représentations sémantiques. Cependant, ces progrès s'accompagnent d'un coût énergétique et calculatoire élevé.

Une approche alternative aux réseaux de neurones artificiels classiques (ANN), qui opèrent en calculs décimaux, réside dans les Spiking Neural Networks (SNN) ou réseaux de neurones à impulsions. Fonctionnant avec des calculs binaires et temporels, les SNN permettent des traitements plus simples tout en offrant une plausibilité biologique accrue. Intégrés aux architectures de type Transformer, ils permettent de simplifier le calcul d'attention. De plus, lorsqu'ils sont déployés sur du matériel neuromorphique, ces Transformers basés sur des SNN peuvent exploiter la sparsité de l'attention — c'est-à-dire la focalisation sur l'information réellement pertinente — pour réduire encore davantage le coût computationnel.

Appliquée à la transcription de documents, cette approche ouvre la voie à la production massive de transcriptions directement à la sortie du scanner, avec un coût énergétique nettement réduit.