



Glider soaring via reinforcement learning in the field, Nature - 19/09/2018

Gautam Reddy, Jerome Wong-Ng, Antonio Celani, Terrence J. Sejnowski & Massimo Vergassola - University of California, San Diego, La Jolla, CA, USA.
Encadrant : Valentin Emiya

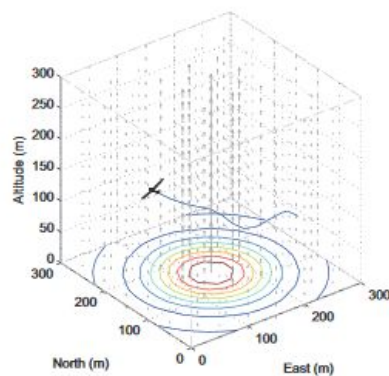
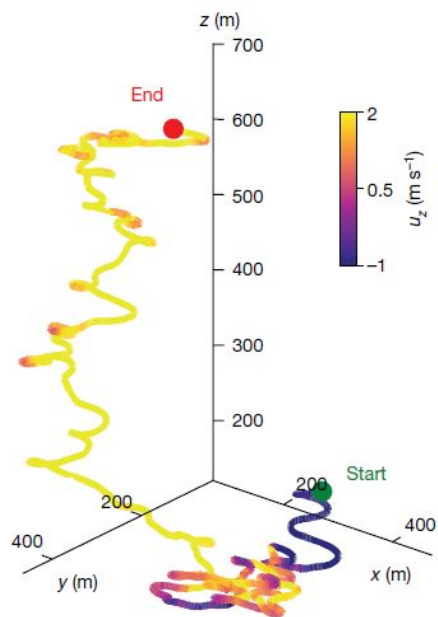
Introduction



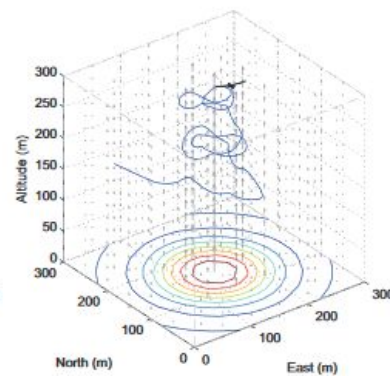
Contexte : Exploitation des ascendances atmosphériques naturelles par des aéronefs sans pilote (UAVs).

But affiché : Permettre à terme des économies d'énergie pour les aéronefs commerciaux en exploitant ces ascendances et mieux comprendre les migrations des oiseaux.

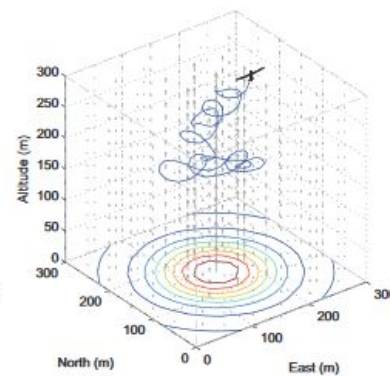
Objectif



(a) Episode 1



(b) Episode 20



(c) Episode 50

Vocabulaire

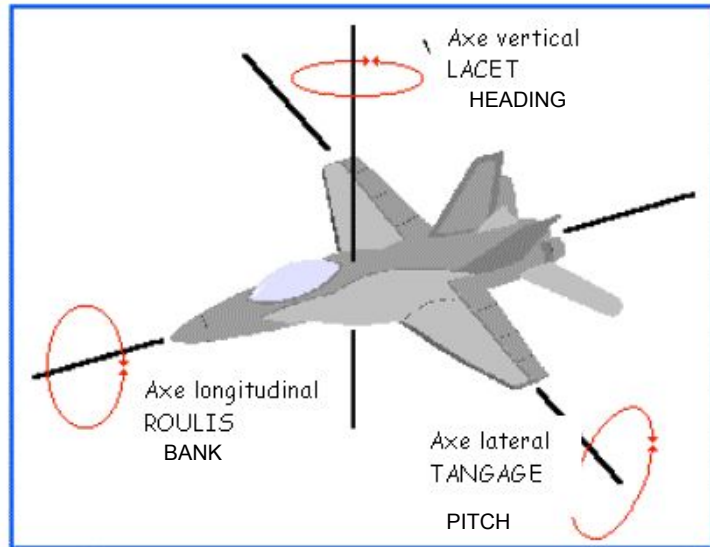
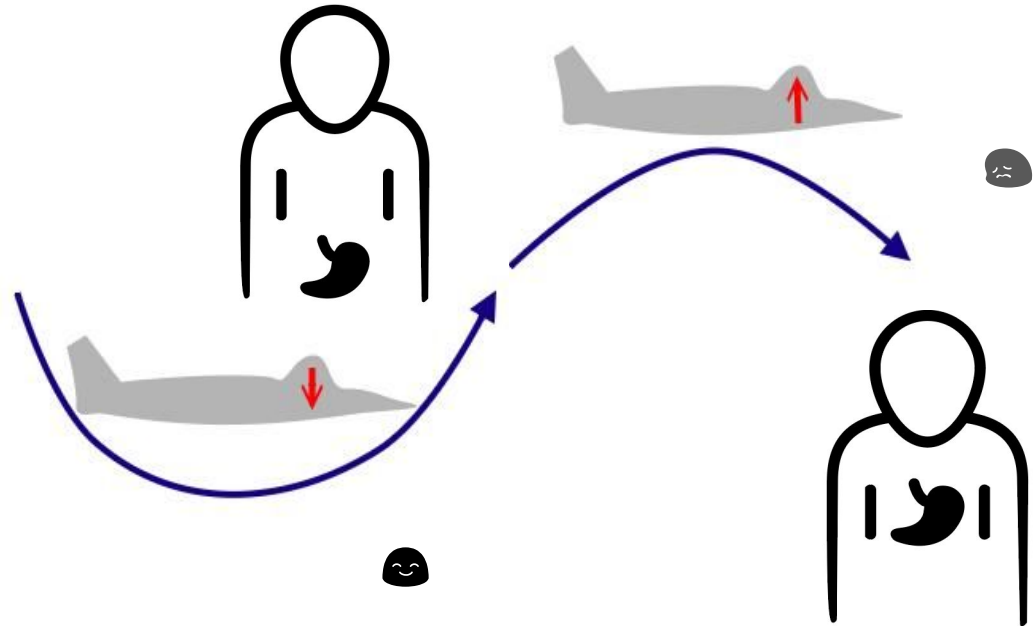
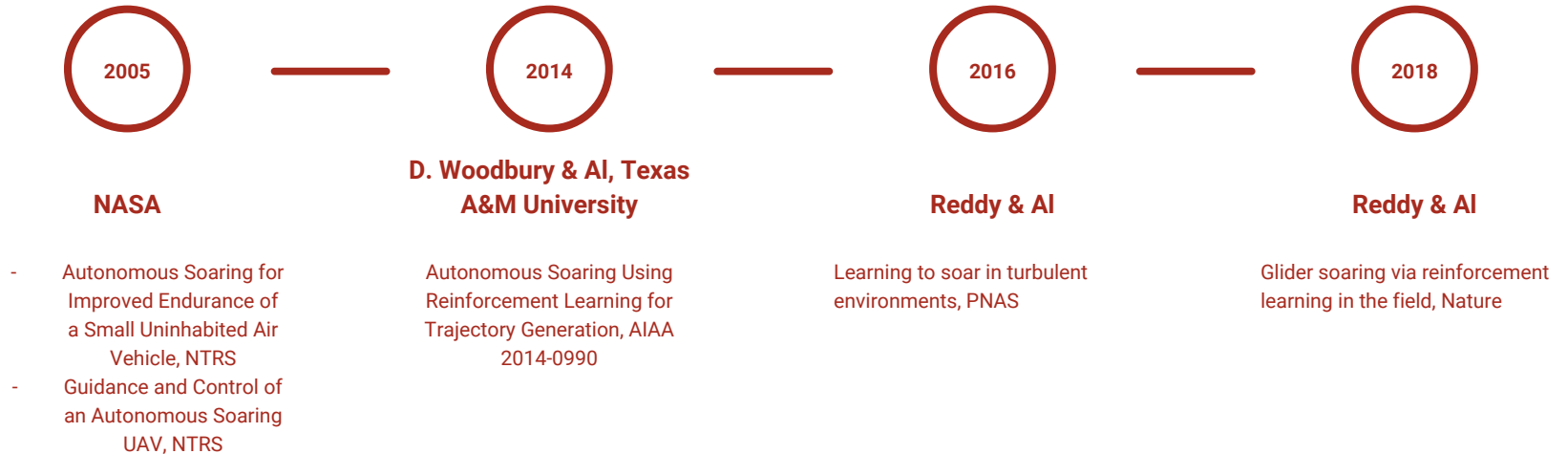


Illustration pour l'accélération verticale

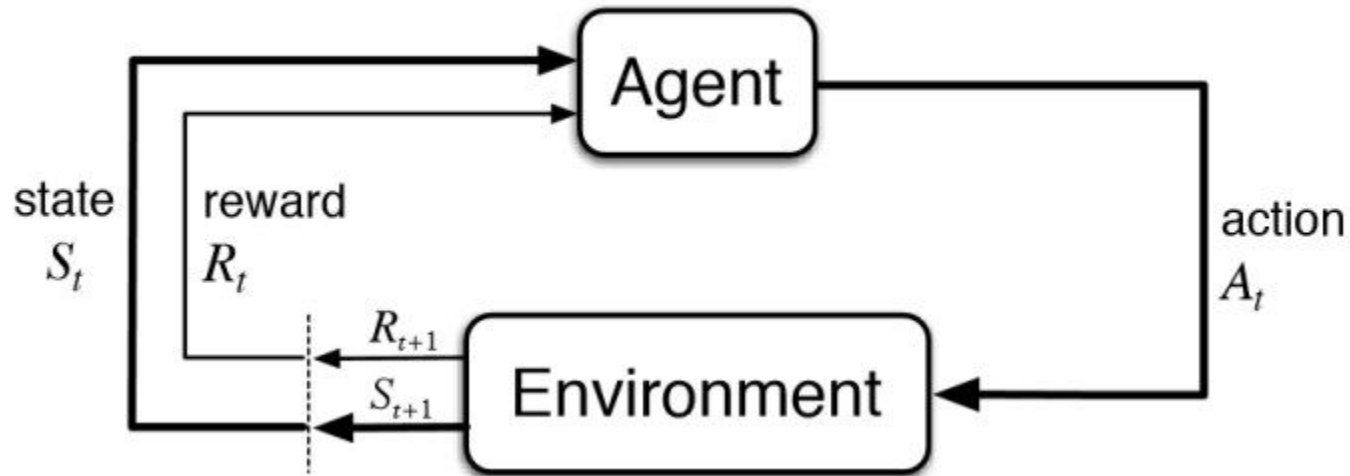




Timeline des travaux sur le sujet



Markov Decision Process (MDP)



$$P(R_{t+1} = r, S_{t+1} = s' | S_0, A_0, R_1, \dots, S_{t-1}, A_{t-1}, R_t, S_t, A_t) = p(s', r | s, a)$$

Premiere publication

Learning to soar in turbulent environments

Gautam Reddy^a, Antonio Celani^b, Terrence J. Sejnowski^{c,d,1}, and Massimo Vergassola^a

^aDepartment of Physics, University of California, San Diego, La Jolla, CA 92093; ^bThe Abdus Salam International Center for Theoretical Physics, I-34014 Trieste, Italy; ^cHoward Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037; and ^dDivision of Biological Sciences, University of California, San Diego, La Jolla, CA 92093

Contributed by Terrence J. Sejnowski, April 28, 2016 (sent for review March 21, 2016; reviewed by Andrew G. Barto and Gregory Falkovich)

Birds and gliders exploit warm, rising atmospheric currents (thermals) to reach heights comparable to low-lying clouds with a reduced expenditure of energy. This strategy of flight (thermal soaring) is frequently used by migratory birds. Soaring provides a remarkable instance of complex decision making in biology and requires a long-term strategy to effectively use the ascending thermals. Furthermore, the problem is technologically relevant to extend the flying range of autonomous gliders. Thermal soaring is commonly observed in the atmospheric convective boundary layer on warm, sunny days. The formation of thermals unavoidably generates strong turbulent fluctuations, which constitute an essential element of soaring. Here, we approach soaring flight as a problem of learning to navigate complex, highly fluctuating turbulent environments. We simulate the atmospheric boundary layer by numerical models of turbulent convective flow and combine them with model-free, experience-based, reinforcement learning algorithms to train the gliders. For the learned policies in the regimes of moderate and

the velocity amplitudes of a glider or bird are of the same order of magnitude as the fluctuating flow they are immersed in.

It has been frequently observed and attested by glider pilots that birds are able to identify and navigate thermals more accurately than human pilots endowed with modern instrumentation (11). It is an open problem, however, what sensorimotor cues are available to birds and how they are exploited, which constitutes a major motivation for the present study.

An active agent navigating a turbulent environment has to gather information about the fluctuating flow while simultaneously using the flow to ascend. Thus, the problem faced by the agent bears similarities to the general problem of balancing exploration and exploitation in uncertain environments, which has been well studied in the reinforcement learning framework (12). The general idea of reinforcement learning is to selectively reinforce actions that are highly rewarding and thereby have the reinforced



Etat

- Accelération verticale du vent
- Torque
- Angle d'incidence (tangage)
- Temperature
- Angle d'attaque



Actions

- Augmenter l'angle d'incidence (5 / 2.5 deg)
- Baisser l'angle d'incidence (-5 / -2.5 deg)
- Augmenter l'angle d'attaque (5 / 2.5 deg)
- Baisser l'angle d'attaque (-5 / -2.5 deg)
- Ne rien changer (0 deg)



Recompense

$$R = u_z + Ca_z$$

- Vitesse verticale du vent
- Constante en parametre
- Accelération verticale du vent

Touche la surface -> grosse penalite

Recompense globale -> probleme de backup pour les longs trajets



L'algorithme

SARSA

TD algorithm

On-policy, Online

```
Initialize  $Q(s, a), \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$ , arbitrarily, and  $Q(\text{terminal-state}, \cdot) = 0$ 
Repeat (for each episode):
  Initialize  $S$ 
  Choose  $A$  from  $S$  using policy derived from  $Q$  (e.g.,  $\epsilon$ -greedy)
  Repeat (for each step of episode):
    Take action  $A$ , observe  $R, S'$ 
    Choose  $A'$  from  $S'$  using policy derived from  $Q$  (e.g.,  $\epsilon$ -greedy)
     $Q(S, A) \leftarrow Q(S, A) + \alpha [R + \gamma Q(S', A') - Q(S, A)]$ 
     $S \leftarrow S'; A \leftarrow A'$ 
  until  $S$  is terminal
```



L'entraînement

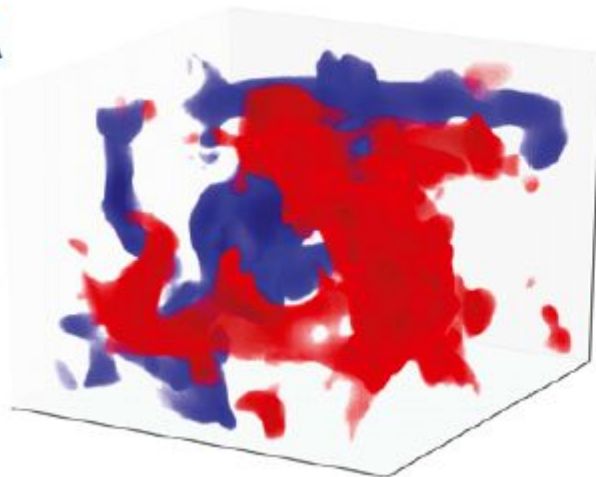
Epsilon greedy

Plusieurs valeurs de C

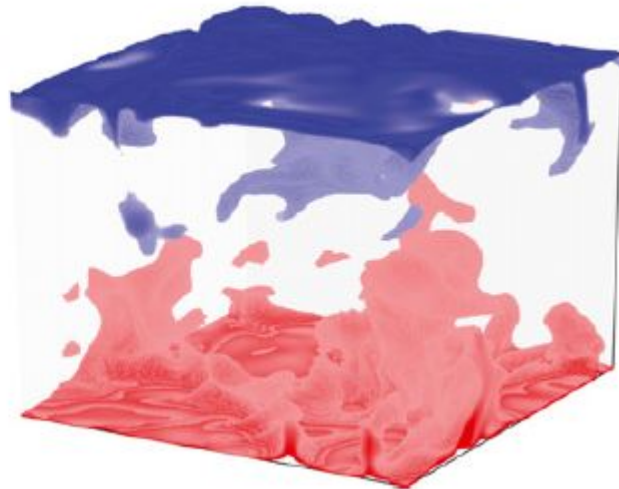
L'environnement

Simulation

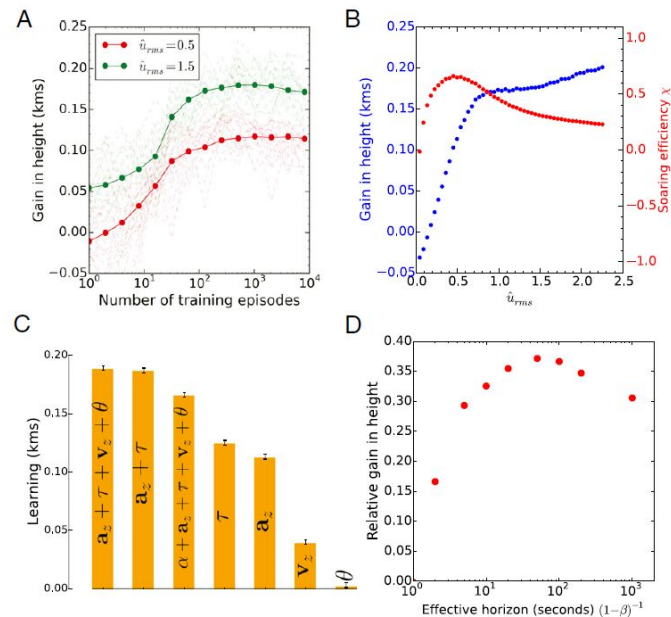
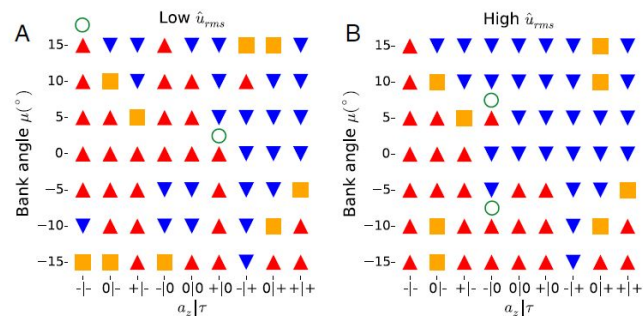
A



B



Les resultats





Publication de 2018

LETTER

<https://doi.org/10.1038/s41586-018-0533-0>

Glider soaring via reinforcement learning in the field

Gautam Reddy^{1,5}, Jerome Wong-Ng^{1,5}, Antonio Celani², Terrence J. Sejnowski^{3,4} & Massimo Vergassola^{1*}

Soaring birds often rely on ascending thermal plumes (thermals) in the atmosphere as they search for prey or migrate across large distances^{1–4}. The landscape of convective currents is rugged and shifts on timescales of a few minutes as thermals constantly form, disintegrate or are transported away by the wind^{5,6}. How soaring birds find and navigate thermals within this complex landscape is unknown. Reinforcement learning⁷ provides an appropriate framework in which to identify an effective navigational strategy as a sequence of decisions made in response to environmental cues. Here we use reinforcement learning to train a glider in the field to navigate atmospheric thermals autonomously. We equipped a glider of two-metre wingspan with a flight controller that precisely controlled the bank angle and pitch, modulating these at intervals with the aim of gaining as much lift as possible. A navigational strategy was determined solely from the glider's pooled experiences, collected

is unrealistic, or have applied learning methods in highly simplified simulated settings^{17–19}.

Using the reinforcement learning framework⁷, we may describe the behaviour of the glider as an agent traversing different states (s) by taking actions (a) while receiving a local reward (r). The goal is to find a behavioural policy that maximizes the 'value': that is, the mean sum of future rewards up to a specified horizon. We seek a model-free approach, which estimates the value of different actions at a particular state (called the Q function) solely through the agent's experiences during repeated instances of the task, thereby bypassing the modelling of complex atmospheric physics and aerodynamics (see Methods). The optimal policy is subsequently derived by taking actions with the highest Q value at each state, where the state includes sensorimotor cues and the glider's aerodynamic state.

To identify mechanosensory cues that could guide soaring, we



Etat

(a_z, ω, μ)

- Accelération verticale du vent (signe ou 0)
- Torque (signe ou 0)
- Angle d'incidence (tangage) (de -30deg à 30deg avec un pas de 15deg)

Seulement 45 états possibles !



Actions

- Augmenter l'angle d'incidence (15deg)
- Baisser l'angle d'incidence (-15deg)
- Ne rien changer (0deg)



Recompense

- Accelération verticale du vent



L'algorithme

Value iteration

Off-policy, Offline

Initialize array V arbitrarily (e.g., $V(s) = 0$ for all $s \in \mathcal{S}^+$)

Repeat

$\Delta \leftarrow 0$

For each $s \in \mathcal{S}$:

$v \leftarrow V(s)$

$V(s) \leftarrow \max_a \sum_{s',r} p(s',r|s,a) [r + \gamma V(s')]$

$\Delta \leftarrow \max(\Delta, |v - V(s)|)$

until $\Delta < \theta$ (a small positive number)

Output a deterministic policy, π , such that

$\pi(s) = \arg \max_a \sum_{s',r} p(s',r|s,a) [r + \gamma V(s')]$



L'entraînement

15 jours d'entraînement

12 jours en choisissant les actions aléatoirement

3 jours en utilisant softmax

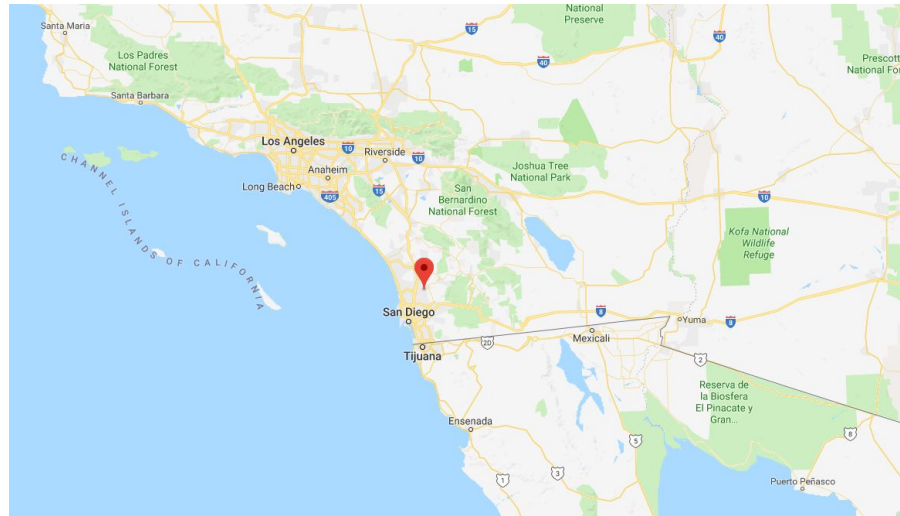
$$P_t(a) = \frac{\exp(q_t(a)/\tau)}{\sum_{i=1}^n \exp(q_t(i)/\tau)}$$

Estimation de $p(s', r \mid s, a)$ puis de V

Symétrie en faisant la moyenne de (a_z, ω, μ) et $(a_z, -\omega, -\mu)$

L'environnement

Environnement reel ($p(s', r \mid s, a)$ défini naturellement).



Objet d'étude

- Planeur Radian Pro PNP (envergure 2m) (grand-public)
- Firmware et controleurs grand-public

Acquisition de données à une fréquence de 10Hz passées dans un filtre EKF pour débruiter :

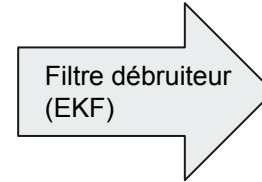
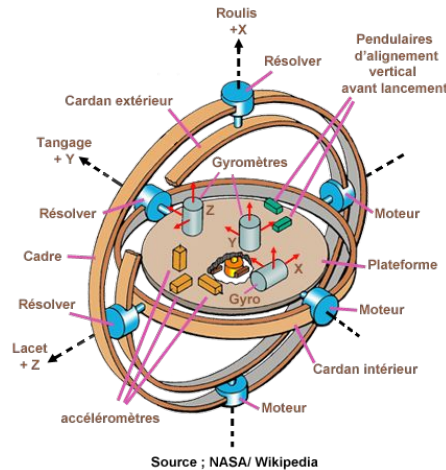
- pression statique et dynamiques => altitude air, vitesse air
- GPS (vitesse sol, altitude GPS) et d'inertie (pour mesurer rouli et tangage)



Schéma de processing des données

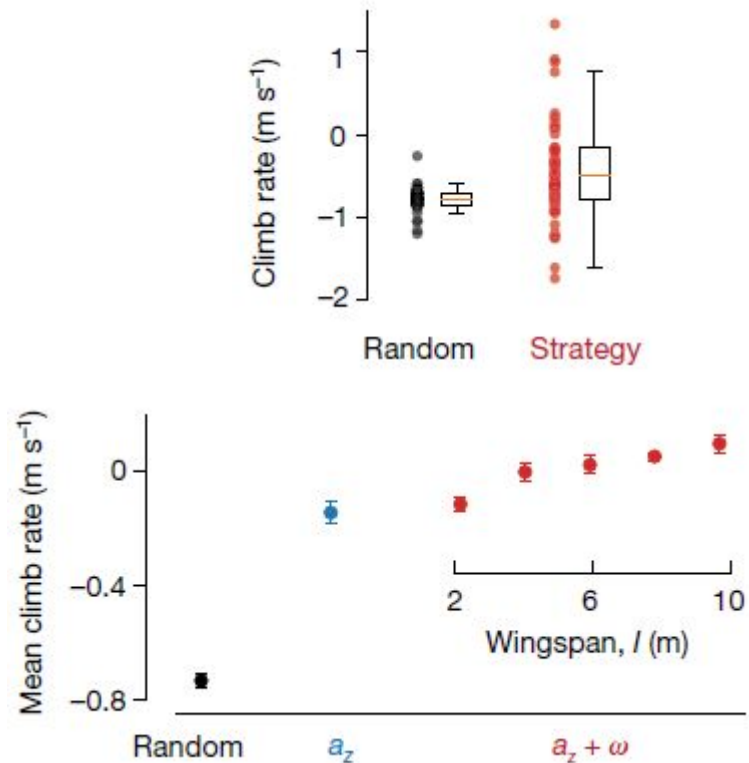
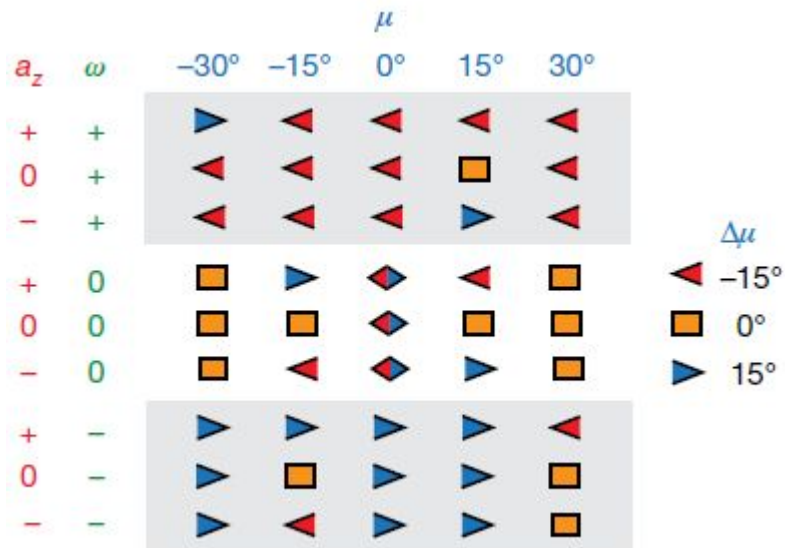


GPS: Shows pilots their position
but not normally used by ATC



$$(a_z, \omega, \mu)$$

Les resultats



— —

Ils en ont parlé



ScienceDaily®



Discover
SCIENCE FOR THE CURIOUS



Il en ont parlé

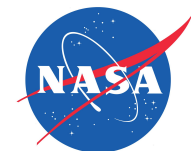




Ils l'ont cité

- Deep Reinforcement Learning, Yuxi Li
- Artificial Lumbered Flight for Autonomous Soaring, TC Powers
- Making BREAD: Biomimetic strategies for Artificial Intelligence Now and in the Future, Jeffrey L. Krichmar, William Severa, Salar M. Khan, James L. Olds

Autres recherches sur le sujet



Autres recherches sur le sujet



Autres recherches sur le sujet





Autres recherches sur le sujet

En simulation :

- GP-SARSA(λ)
- Q-Learning avec reseaux de neurones
- Heuristiques

Autre approche pour l'économie de carburant

Optimisation des trajectoires de montée et descente :

- Murrieta Mendoza, Alejandro and Botez, Ruxandra, 2015, Aircraft Vertical Route Optimization Deterministic Algorithm for a Flight Management System
- Alessandro Gardi, Roberto Sabatini, Subramanian Ramasamy, 2016, Multi-objective optimisation of aircraft flight trajectories in the ATM and avionics context

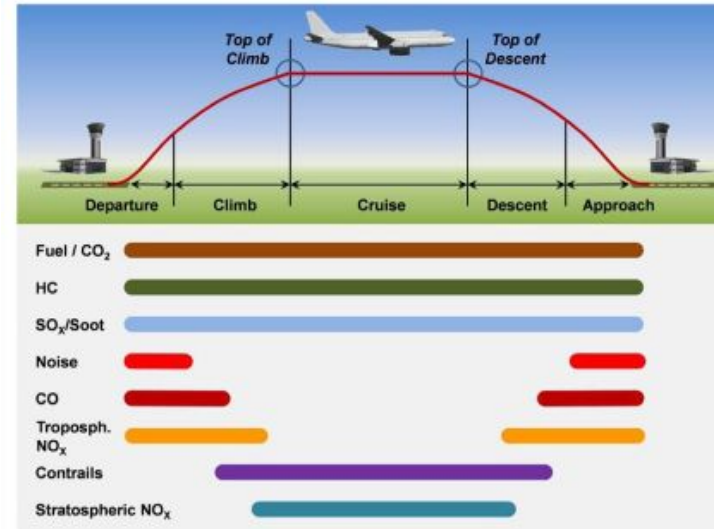


Fig. 4. Repartition by flight phases of the environmental impacts associated with the aircraft operation.

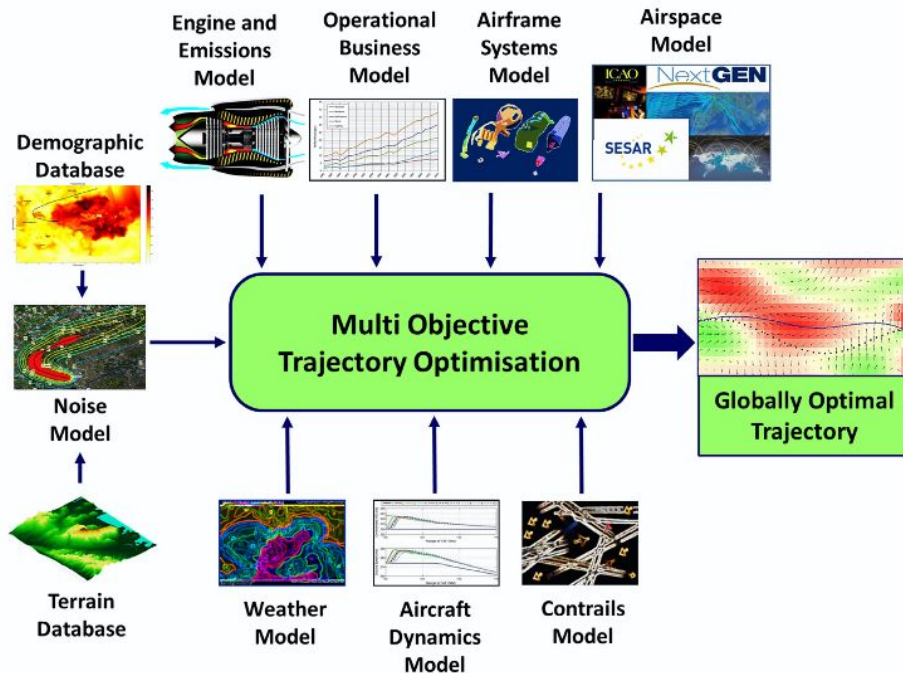


Fig. 7. Multi-objective trajectory optimisation concept [19].

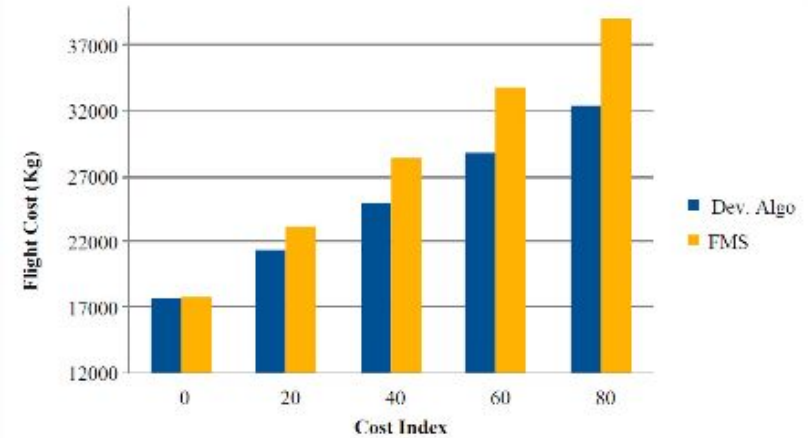


Figure 10. Flight Cost Comparison with FMS

Competitions





Le probleme de la comparaison

- Besoin du même environnement réel
- Pas les mêmes mesures et récompenses
- Pas de compétition exclusivement sur l'endurance

Pourquoi ça fait tant de bruit ?



- Apprentissage par renforcement au top de sa “hype”
- Sujet évocateur pour l’imaginaire collectif
- Solution diablement simple donc élégante



Progrès apportés relatifs aux buts affichés

Oiseaux :

- Difficile à évaluer

Économies pour l'aviation commerciale :

- L'exploitation thermique est encore illusoire pour l'aviation commerciale
- D'autres types de recherche (y compris en IA) sur l'optimisation des trajectoires sont plus abouties et déjà expérimentés