Lecture Notes: Com B - Essor de l'I.A.

Taught by Prof. Floran Jaton, Rémi Lebret

Cours de SHS - Enjeux mondiaux de la communication en lien avec l'essor de l'intelligence.

Emmanuel Omont, Simon Lefort, Spring 2024

Outline

Chapter 1: La loi de Moore

1.1 Les promesses technoscientifiques

- du 16e siècle au 18e siècle ⟶ fonction de sensibilisation
- 19e siècle ⟶ fonction idéologique

1.1.1 Ces promesses sont toujours…

- non dystopiques
- imposent des solutions technologiques
- performatives[1], le fait de formuler la promesse contribue à la faire réaliser (orientent les moyens alloués à la recherche & innovation)

1.1.2 …et ont pour contraintes…

- la nécessité de nouveauté radicale (la promesse est la solution unique à un problème urgent)
- crédibilité (soutient des spécialistes, quitte à inventer ces soutiens)

1.1.3 Loi de Moore en microélectronique

est un modèle pour la fabrication de promesses technoscientifiques
↳ définit ce qui est pensable pour l'évolution des micro-processeurs

> 📖 **Contexte** : apparitions des premiers circuits intégrés
>
> - Fin 1950 ⟶ Fairchild Seminconductor fabrique des transistors (#FDS)
> - Pour inciter d'autres acteurs à faire le pari de l'ouverture de la société civile, Gordon Moore, directeur R&D chez Fairchild, publie un manifeste économique
> - promesse en faveur de l'intégration :
>
> The future of integrated electronics is the future of electronics itself…
>
> — Moore, 1965

Seconde formulation de la loi :

~~The density has increased at a rate of roughly a factor of two per year~~,
The density has increased at a rate of roughly a factor of two per two years

— Moore, 1975

Pour les CPU, en 2024, il est probable qu'on se détache de cette loi :
- l'énergie est chère
- coûts des lieux de fabrication
- la consommation change (+ d'économie d'énergie over + de performance)

[1] #olympedegouge

1.2 depuis 2010 : la course aux GPU

CPU	**GPU**
Quelques coeurs: entre 2 et 64	Plusieurs coeurs: entre 2'000 et 50'000
Faible latence	Haut débit de données
Bon pour le traitement en série	Bon pour le traitement en parallèle
Peut effectuer une poignée d'opérations à la fois	Peut effectuer des milliers d'opérations à la fois

Les G.P.U. (graphical processing units) sont très efficaces pour la multiplication de matrices (donc très utiles pour l'I.A.).

1.3 Loi de Huang

> Les performances des GPUs seront plus que doublées chaque 2 ans
>
> — Jensen Huang (CEO Nvidia, leader des cartes GPU)

1.4 Optimisation

Représentation des nombres : réduction de la précision pour accélerer les calculs. (2019, Google, grâce au brain floating point).

Nouvelles cartes :

- le T.P.U. (Tensor Processing Unit) par Google (2015), conçu pour les réseaux de neurones (fonctionne avec Tensorflow)
- le L.P.U. (Language Processing Unit) par Groq
- Apple Série M (système sur une unique puce SoC : CPU, GPU, mémoire unifié)

et beaucoup d'autres entreprises (AWS…)

1.5 Etat du marché

Nvidia domine, tensions géopolitiques liées aux semi-conducteurs.

Chapter 2: Les modèles de langue

LLM : Large Language Model

2.1 Recette d'un bon LLM

- bcp de paramètres (x10 chaque année, nouvelle loi de Moore[2] ?)
- de la puissance de calcul
- bcp (bcp) de données (seront épuisées en 2026 !)

[2] OMG la dinguerie !??!

Chapter 3: Des booms et des hivers de l'IA

Historiquement, les technologies IA ont traversé des phases de booms et de crises.

3.1 Genèse et 1er boom de l'IA (1940–1965)

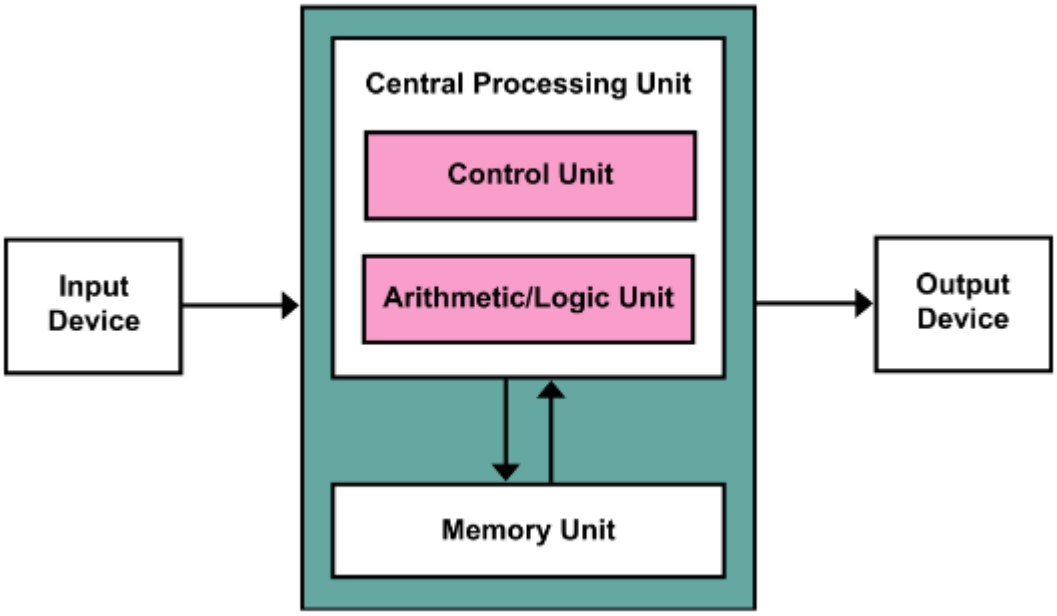
Seconde Guerre mondiale ⟶ augmentation de la demande en calcul
- Angleterre ⟶ décryptage ("Bomb", "Coloss")
- U.S.A. ⟶ calcul balistique (ENIAC)

1942 : Moore School of Electrical Engineering, pour accélérer la production de tables de tir.

1943 : $400k alloués à la constructeur de l'ENIAC (le dispositif prend le nom de "computer")
↳ très innovant mais problème d'architecture

1944 : nouveau projet dérivé de l'ENIAC, le **EDVAC** (notament grâce à John von Neumann).

Formalisation de l'**architecture Von-Neumann** (encore très utilisée aujourd'hui) :



Juin 1945 : Neumann publie un rapport sur EDVAC (utilise des analogies avec le cerveau pour la première fois)

1949 : 1er ordinateur BINAC (sert de référence à UNIVAC, 1951), les ordinateurs/calculateurs sortent progressivement de la recherche militaire ⟶ **industrie et administration**.

1956 : 1ère appartition du terme I.A. (John McCarthy)

1960 : premier boom de l'IA dite "symbolique"
↳ recherche logicielle visant à décrire les règles de pensée et les exprimer sous forme de code informatique (par ex. chatbot ELIZA).

Un groupe fermé de chercheurs s'arroge le monopole de la définition des enjeux de l'IA.
↳ capture l'essentiel des financements (75% de US Air Force)
↳ conserve l'accès aux grands systèmes informatiques

… cela conduit au premier hiver.

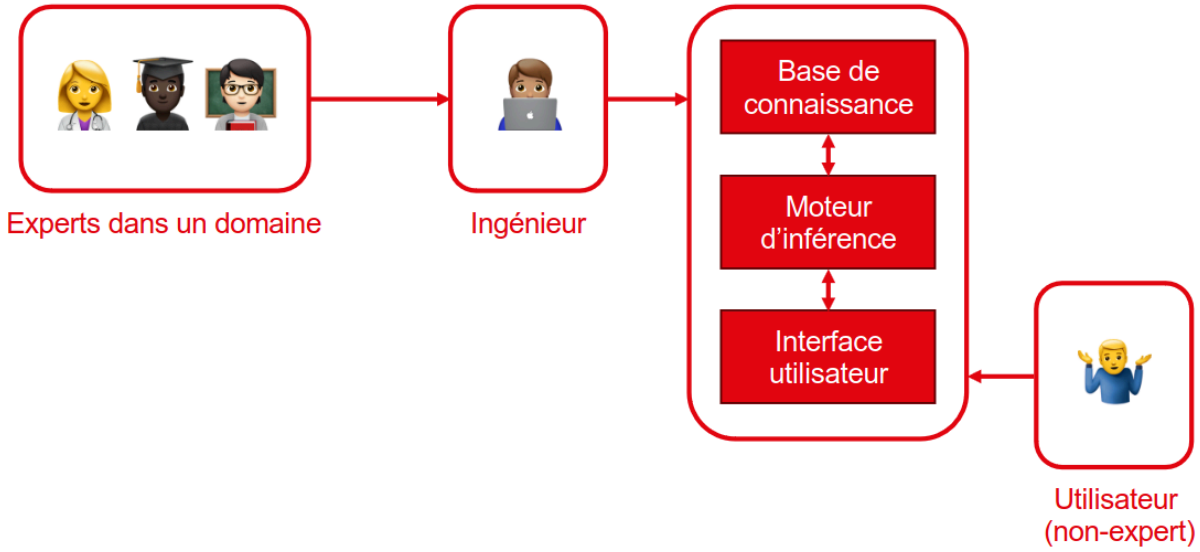
3.2 Premier hiver de l'IA (1965-1975)

⟶ promesses des promoteurs de l'IA symbolique n'ont pas été tenues
⟶ à partir de 1970, baisse des financements (notamment militaires)
⟶ accusés de se concentrer sur des "mondes jouets"

3.3 Systèmes experts experts et 2ème boom (1975-1985)

Renouveau de l'IA symbolique (ordi + puissant et décomposition des processus de raisonnement en briques élémentaires)

Système expert :



1970 : système MYCIN
⟶ série de questions au médecin
⟶ env. 600 règles
⟶ produit une liste de bactéries candidates

1980 : XCON pour aider à configurer les ordinateurs

1984 : DELTA pour identifier les pannes sur les locomotives

Limites des systèmes expert :

> Dans les décennies à venir, nous devrons disposer de moyens plus automatiques pour remplacer ce qui est actuellement une procédure très fastidieuse, longue et coûteuse.

— Edward Feigenbaum, 1983

3.4 Deuxième hiver et travail de l'ombre

Des promesses non tenues (encore…)
⟶ problème de hardware
⟶ problème de maintenance des logiciels
⟶ la plupart des startups ont fait faillite

1990 : IA symbolique est si affaiblie que le terme disparaît quasiment du vocabulaire de recherche.

3.4.1 Parallel Distributed Processing

1986 : à l'écart de l'IA symbolique, un groupe de chercheurs travaille sur les **réseaux de neurones** (notamment reconnaissance des codes postaux)

notion de rétropropagation du gradient (ajuster les paramètres du modèle en fonction des erreurs qu'il commet, en fait le gradient c'est la dérivée, on l'utilise pour savoir dans quelle direction aller pour minimiser l'erreur).

LeNet-1, reconnaissance de chiffres - 1989
LeNet-5, reconnaissance de caractères (reconnaissances de ZIP codes par ex.) - 1998

3.5 Réseau de neurones profonds et 3ème boom (2005-2024)

3.5.1 Avènement du Deep Learning :

- puissance de calcul augmente (performances des cpu ↗ + GPU)
- réseaux de neurones + profonds

3.5.2 Numérisation et essort d'Internet

- quantité de données ↗
- mise en place de plateformes de crowdsourcing (pour labelliser des données)

Exemple : ImageNet (1k catégories d'objets, 1.2M d'images)

2012 : AlexNet, reconnaissance d'objets, meilleurs performances sur ImageNet (25% ⟶ 16%) grâce aux **filtres de convolution**.

2015 : ResNet

⟶ application d'une même transformation linéaire sur différentes zones de l'image
⟶ on part de petites matrices (3x3) en demandant au modèle de générer de grandes matrices (5x5) **, chaque groupe est une couche de convolution** (pour éviter de se concentrer sur les détails).

Il y a plusieurs **cartes d'activation** (filtre = feature, appris à l'entraînement) à la sortie de chaque couche.

Réduction des cartes d'activation par **opération de pooling**.

Chaque convolution est suivie d'une **fonction d'activation** non linéaire.

La dernière couche est la couche **de classification**, qui détermine la sortie avec poids (appris pendant l'entraînement).

Chapter 4: Des booms et des hivers de l'IA (II)

4.1 Genèse: memorandum de Weaver et démos. publiques (1950-1965)

1949 : Weaver suggère une meilleure approche (statistique et probabiliste) que celle de la traduction linéaire

1954 : première démonstration publique à New York, traduction de russe à anglais en public.

4.2 Crise: le rapport ALPAC et ses conséquences (1965-1990)

1966 : le rapport ALPAC (Automatic Language Processing Advisory Committee)

> Il n'y a aucune urgence dans le domaine de la traduction automatique. Le problème n'est pas de satisfaire un besoin inexistant à travers des systèmes de traduction automatiques inexistants
>
> — National Research Council, 1966

peu de bénéfices à court-terme → **chute drastique des financements.**

4.3 Renouveau et tradition statistique (1990-2015)

1990 : apparition de corpus parallèles[3] (utiles pour la traduction)

1992 : rapport JTEC (Japan Technology Evaluation Center) et incitations politiques (convaincre les gouvernements d'utiliser les nouvelles technologies)

mais aussi puissance de calcul et stockage ↗
et nouvelle culture statistique, probabiliste

4.3.1 Focus sur modèles de traduction basés sur les groupes de mots

1993 : IBM introduit plusieurs modèles statistiques pour la traduction (corpus parallèles issus du parlement canadien)

2006 : Google Translate, basée sur cette méthode

4.3.1.1 Comment ça marche ?

- segmentation des phrases en groupe de mots (tokens)
- recherche de correspondances les + probables
- assemblage des correspondances

4.3.1.2 Limitations de l'approche statistique

- traduction fausse si syntaxe non courante
- utilisation de l'anglais comme "langue pivot" (FR → EN → IT)[4]

[3] Merci au parlement canadien d'avoir traduit gratuitement des textes anglais-français. Sinon, on utilise aussi la Bible, vu que c'est un texte traduit dans quasiment 100% des langues

4.4 Traduction automatique par réseau de neurones (2014-2024)

4.4.1 Boom des ConvNets

Entre 2012 et 2015 : boom des ConvNets pour traitement des images

4.4.1.1 Comment ça marche ?

Comment apprendre le langage naturel avec des réseaux de neurones?

- les machines comprennent le langage binaire
- les réseaux de neurones doivent recevoir en entrée des données continues
- le texte est représenté par des symboles discrets (lettres, chiffres, caractères spéciaux, etc.)

Problème : avec ASCII, l'encodage binaire, un mot n'est pas défini par ses lettres, impossibilité d'apprendre le sens d'un mot avec binaire. (ex chouette ≠ brouette).

Solution : encodage one-hot : grâce au word embeddings un graph basé sur le sens tous les mots va se former.

Pour cela : on prend la probabilité de coocurrence P(c/w), puis on réduit les dimensions trouvées pour être + ou - précis dans la compréhension (grâce à la SVD).

⟶ une méthode efficace mais peu efficiente (ça a pris 4 mois pour s'entraîner sur Wikipedia).

4.4.1.2 word2vec (2013)

- modèle linéaire
- tricks pour améliorer l'apprentissage des mots rares
- code open source en C
- apprentissages de word embeddings en quelques minutes

Exemple : déterminer si un avis est positif ou négatif.

4.4.2 Réseaux récurrents

- adapté au langage, qui est séquentiel

2014 : premiers réseaux de neurones récurrents pour la traduction automatique

4.4.3 Long-Short Term Memory (LSTM) Network (traduction)

Problème de l'époque : apprentissage des réseaux récurrents difficile pour les longues séquences.

Une idée des années 90 refait surface: LSTM networks.

2016: Google Translate opte pour un modèle de traduction neuronal basé sur les LSTM

4.4.3.1 Limites des LSTM-RNN

- problème d'optimisation
- modèles séquentiels difficilement parallélisables
- architecture peu efficace sur GPU

[4] le truc drôle en plus, c'est si vous le mettez en PLS il sortira une phrase de la Bible

- temps d'apprentissage + long

4.4.4 Transformer: Attention Is All You Need

- 2017: nouvelle architecture basée uniquement sur le mécanisme d'attention
- efficace sur GPU (entraînement + rapide)
- les RRN n'ont pas dit leur dernier mot (transformer + rapide en entraînement mais gourmands en mémoire)
- Mars 2024: Google DeepMind présente deux nouveaux modèles de langue basés sur des RNNs

Mécanisme d'attention : fait attention à n'importance des mots en fonction du contexte des mots en l'entourant.

Chapter 5: Supervision et apprentissage automatique

5.1 Apprentissage supervisé

Une fois qu'un jeu de données est disponible, les chercheurs la divise en 2 sous-ensembles :
- jeu d'entraînement
- jeu d'évaluation

5.1.1 Des avancées et des limites

- biais de sélection des données
- biais des annotations des données
- coût de l'annotation des données (contraîntes de temps : ex. annotation de contrats juridiques 1/4 de temps juste pour les annoter).

5.2 Apprentissage auto-supervisé

Yann Le Cun "Cake Analogy" : la grosse partie du gâteau est faite avec un apprentissage non supervisé, et le glaçage, la cerise, est faite avec un peu d'apprentissage supervisé.

5.2.1 Apprentissage des données d'entrée

- apprendre à prédire le mot d'après, ou les mots cachés
- GPT (Generative Pre-Trained Transformer) ⟶ prédire le mot suivant
- apprendre des images avec des tâches "prétextes" prédéfinies (rotation, mettre en couleur)

5.2.2 Apprentissage discriminatif

utilise un signal discriminant entre les images, permet de classifier les images

5.2.3 Apprentissage contrastif

Apprendre à déterminer si une paire d'image est positive ou pas (donc si les deux images sont de la même catégorie)
On donne des paires d'images positives, c'est-à-dire d'un même set, et des paires d'images négatives. Ensuite pour continuer à l'entraîner on lui donne des paires avec une image A et cette image A retournée, et on vérifie si le résultat est bien "positif" (c'est les tâches prétextes).

Exemple : **modèle CLIP**, auto-supervision avec 400M de paires (image, description) collectées sur le web. Performances équivalentes à ImageNet. Meilleures généralisation des données inconnues.

5.3 Apprentissage par instruction (objectif: alignement)

Les modèles basés sur l'apprentissage auto sont limités, ils ne savent que prédire le mot suivant.

Solution : s'aligner sur les attentes des utilisateurs :

Instruction: Traduit la phrase suivante en Français.
Observation: The cat sat on the mat.
Label: [...prédit par le modèle...]

Cependant, les données annotées restent coûteuses et limitées. Ainsi, on peut utiliser d'autres modèles pré-entraînés pour générer des instructions.

Exemple : **Alpaca dataset,** 52 000 instructions générées avec GPT-3.5 à partir de 175 instructions.

Solution améliorée : s'aligner sur les domaines spécialisés

Exemple : Google Med-PaLM (modèle sous license propriétaire) adapté depuis PaLM.

A intégrer à la solution : s'aligner sur les valeurs humaines

- le web contient des données toxiques ou mensongères
- human in the loop : humains qui vérifient les données

5.3.1 Les limites de l'alignement

- Gemini image était woke
- il est actuellement dans les faits impossible de sortir de la supervision
- les modèles d'IA qu'on utilise ont été supervisés avec des biais arbitraires
- la seule piste qu'on a est de rendre explicite les valeurs qui sous-tendent le process d'annotation et d'alignement[5]

[5] basiquement on va juste dire "oui j'assume que notre modèle est biaisé et pense que la terre est plate, désolé on y peut rien nos données venaient des platistes"

Chapter 6: IA Generative & sphère informelle

6.1 Quelques exemples

- Pape François en doudoune, généré par un employé du BTB à Chicago, relayée sur twitter
- Donald Trump arrêté par la police de NY, dans un contexte tendu
- Emmanuel Macron au milieu de la réforme des retraites

→ À Davos, tous les regards sont sur les deepfakes.

6.2 Generateur d'images par IA

6.2.1 IA generative VS IA discriminative

- Discriminatif : focus sur les caractéristiques qui distinguent les différentes catégories (objectif : reconnaître une image)
- Génératif : modélisation de la distribution des données

6.2.2 Type de modèles pour la distribution d'images.

Auto-Encodeur :
Apprendre une distribution approximée d'un ensemble de points ayant une distribution inconnue.

Cool : compresser des images dans un espace réduit, apprentissage auto-supervisé.
Pas cool : Qualité mauvaise, manque de contrôle sur la génération (on ne peut pas dire où sont les yeux d'une image par ex).

Auto-Encodeur variationnel

Apprentissage d'une représentation probabiliste et continue de l'entrée
Le décodeur génère une image à partir d'une variable latente

Cool : Représentation plus structurée, images + diversifiées, bcp plus de contrôle
Pas cool : On génère l'image en une étape (VAE)

Modèle de diffusion

On génère l'image en plusieurs étapes.
↳ On apprend à prédire le bruit d'une image avec un encodeur, qui "bruite" l'image au fur et à mesure (à la fin on a un espace réduit ducoup), après, on apprend à reconstruire l'image avec un décodeur.

forward diffusion : obtenir le bruit à partir de l'image
reverse diffusion : l'inverse.

On a une chaîne de markov, avec des étapes petit à petit.

6.2.3 Entrainement pour modèle de diffusion

1. Prend une image
2. Générer du niveau de bruit
3. On choisit un niveau de bruit

4. Ajouter le bruit à l'image
5. On lui demande de prédire le bruit qui a été ajouté à l'image.

↳ Une fois entraîné, on peut prédire l'image suivante en retirant le bruit que le modèle pense qui a été ajouté.

6.2.4 Conditionner la génération d'image

On va ajouter du texte à notre jeu d'entrainement.
↳ On utilise des modèles qui ont déjà labelisé les images (modèle CLIP) ⟶ Chaque mot est représenté par du text embedding.
↳ Le prédicteur de bruit avec mécanisme d'intention va générer une image aléatoire conditionnée avec le texte entré
↳ On peut aussi conditionner en disant "eh à cet endroit tu me met un train"[6]

6.2.5 Principaux modèles de diffusion text-to-image

1. Midjourney (v1 ⟶ v6)
2. Dalle-E (2 ⟶ 3)
3. Imagen (⟶ 2)
4. Stable diffusion (v1 ⟶ v3)

6.2.6 Controverse de génération d'image

Des scientifiques ont publié un papier avec des images générées par Midjourney, ils se sont fait taper dessus

6.3 Fausses images et sphère informationnelle

Jusqu'à présent, il n'y a pas eu de répercussions géopolitiques majeures[7], mais ces images génères des inquiétudes.
↳ Elles mettent à mal la recherche d'image inversée et l'analyse des retouches (les 2 check des sécurités)

Ces fausses images continuent de mettre de l'huile sur le feu dans un monde avec déjà beaucoup de guerres.

Point de départ de la guerre informationnelle : **Guerre du Golfe**, quand le gouvernement du Koweit engage de nombreuses ressources pour mobiliser l'opinion publique contre Saddam Hussein (ex. témoignage vidéo de Nahira, vu par 60,000,000 d'américains **qui était un faux !**).
↳ aide bcp le Koweit

Tous les pays commencent à prendre au sérieux cette domination informationnelle.

Début années 1990 : Iran, mise en place de l'IRIB. un peu comme CNN, 1Md de budget, 15k employés, mais l'information est orientée pour aller dans le sens de l'islam

1996 : fondation d'Al Jazeera par le Qatar, sur le modèle aussi de CNN, constitue un canal pour les déclarations du Hamas, du Hezbollah (sud Liban), et d'Al-Qaïda.

[6] Si vous voulez voir à quoi ça ressemble concrètement, vous pouvez aller voir Nvidia Canvas
[7] où on ne le sait pas encore MDR

1998 : opération bouclier doré (contrôle de l'accès des internautes chinois aux contenus étrangers). En 2022 : 73% du trafic internet chinois provient de l'intérieur de ses frontières !

1994 : système SORM (Système pour Activité d'Enquête Opératoire) permet d'intercepter l'ensemble des communications sur l'ensemble des territoires russophones.

2005 : lancement du média Russia Today (propagande)

2016 : grande campagne de déstabilisation informationnelle, lors de la campagne électoral entre Trump et Hillary Clinton (piratage des boîtes mails d'Hillary Clinton)

Situation préoccupante de chaos informationnelle.

2024 : Annonce faite par le Hamas (bombardement de l'hopital Al-Ahli) ⟶ relayée par la presse internationale… alors que c'était faux. Cela a conduit à une très forte augmentation des manifestations propalestine.

6.4 Mais il reste de l'espoir !

Le propriétaire du journal The Guardian est dirigé par une fondation, à but non lucratif. Il assure une indépendance d'écriture, car les journalistes n'ont pas besoin de "rechercher le scoop".

De même Le Temps est sorti d'une entreprise pour se faire racheter par une fondation, Aventinus, créée par des personnes riches pour soutenir la presse romande. **le comité de rédaction est séparé du conseil de fondation !** pas de conflit d'intérêt.