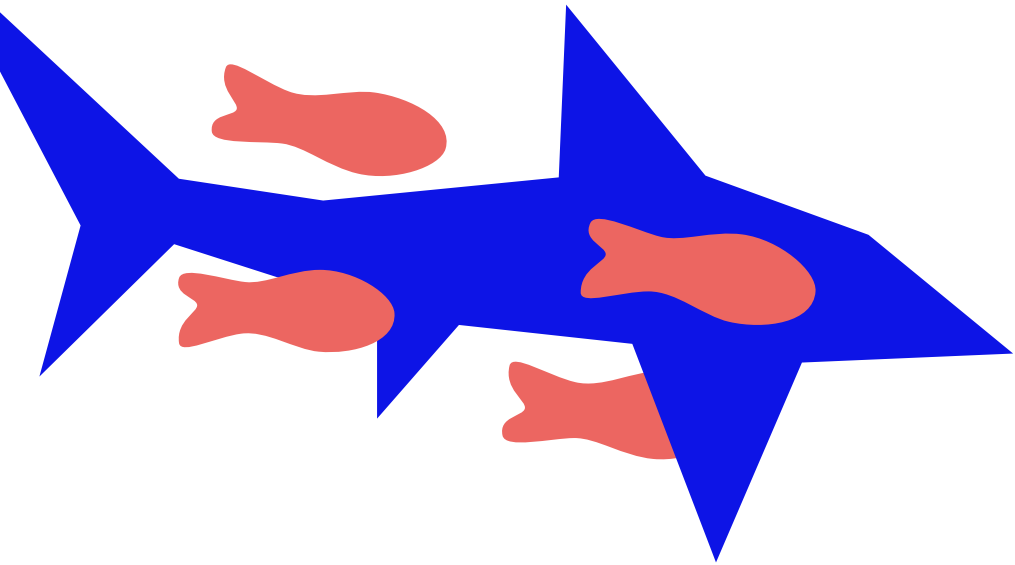
Approche humaine à l'intelligence artificielle et sa réglementation

Thomas Burri
Juliane Beck
François Fleuret
Serhiy Kandul
Markus Kneer
Vincent Micheli
Markus Christen



Professeur Dr iur. Thomas Burri, LLM., RA,
professeur de droit européen et international, Université de Saint-Gall

Juliane Beck,
doctorante en droit, Université de Saint-Gall

Dr François Fleuret,
professeur d'informatique (apprentissage automatique), Université de Genève

Dr Serhiy Kandul,
post-doctorant en économie comportementale, Université de Zurich

Professeur Dr Markus Kneer,
professeur d'éthique de l'intelligence artificielle, Université de Graz

Vincent Micheli,
doctorant en informatique (apprentissage automatique), Université de Genève

PD Dr Markus Christen,
directeur de la Digital Society Initiative, Université de Zurich

L'intelligence artificielle (IA) connaît actuellement un développement fulgurant. Les autorités réagissent en multipliant les projets de régulation, mais de nombreuses incertitudes subsistent et leur compliquent la tâche.

Les systèmes d'IA poussent comme des champignons et ne cessent de se diversifier. Des applications hautement spécialisées font leur entrée dans de multiples domaines spécifiques, déjà réglementés séparément. L'IA en elle-même est de surcroît très polyvalente. Pour nombre d'applications se pose la question du rôle de la personne en contact plus ou moins direct avec l'IA.

Ces dernières années, un projet soutenu par le Fonds national suisse s'est penché sur un aspect crucial de ce problème: quel contrôle les êtres humains peuvent-ils et doivent-ils exercer sur les applications d'IA. Les modes de réglementation actuels les plus pertinents pour l'IA ont été examinés dans un projet séparé.

La présente brochure résume les résultats scientifiques sous la forme de recommandations.

Celles-ci sont principalement adressées aux politiques, aux autorités et au grand public.

Surveillance et contrôle humains

L'idée d'encadrer les applications d'IA par une surveillance humaine avait déjà été avancée il y a une dizaine d'années, alors que les États tentaient de réguler les systèmes d'armes létales autonomes (également appelés «robots tueurs») dans le cadre de la Convention sur certaines armes classiques, ou CCAC, signée à Genève.

Aucun accord n'avait cependant été trouvé sur l'amplitude du contrôle humain légalement requis pour les systèmes d'armes en question, de plus en plus basés sur l'IA. Néanmoins, la nécessité d'un tel contrôle est prise en compte dans les projets de réglementation actuels. Le principe d'une maîtrise humaine de l'IA, qui ne s'applique toutefois pas aux systèmes d'armes, est au cœur même du règlement sur l'intelligence artificielle (également appelé Acte sur l'IA ou loi sur l'IA, EU AI Act en anglais) qui est désormais entré en vigueur au sein de l'Union européenne.

Ce règlement contient notamment une disposition sur le contrôle humain des applications d'IA «à haut risque» (cf. art. 14 de la loi sur l'IA).

Quoiqu'il en soit, que l'on parle du contrôle de l'IA au sens large ou de la surveillance de l'interaction concrète entre une personne et la machine plus spécifiquement, le débat porte toujours sur les rapports entre intelligences humaine et artificielle. Le contexte de cette approche réglementaire est le suivant: les êtres humains dirigent ou dominent toujours l'IA d'une façon ou d'une autre.

Il leur incombe par conséquent la responsabilité (juridique) des résultats (outputs) obtenus. Les dispositifs théoriquement à même de concrétiser le contrôle ou la surveillance de l'IA comprennent typiquement un mécanisme d'arrêt (stop button) ou une observation par échantillonnage du comportement de l'IA. Les formes de contrôle ou de surveillance humains peuvent néanmoins varier. Citons, par exemple, la transparence, la gestion des données, ou encore la gestion du risque, qui doivent permettre de prévenir les menaces pesant sur les applications d'IA. La loi sur l'IA de l'UE prévoit déjà de tels dispositifs pour certains systèmes d'IA particulièrement dangereux (c.-à-d. «à haut risque»). La présente brochure et les recherches qui la sous-tendent sont largement consacrées aux possibilités d'intervention concrètes dans les processus en cours basés sur l'IA.

De la température de l'intelligence artificielle

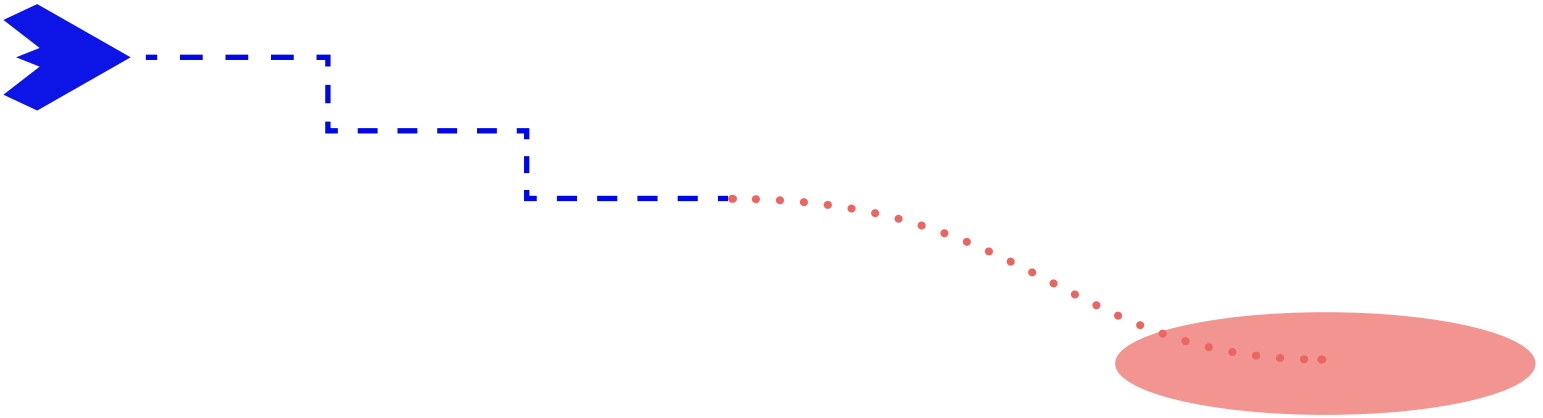
Pour bien comprendre le sens du contrôle de l'IA, il est utile de distinguer les IA «chaudes» des IA «froides». Une IA est dite «chaude» lorsqu'elle implique un processus d'interaction continu, gourmand en ressources et nécessitant des réactions rapides d'une personne, qui n'a parfois que quelques secondes pour décider d'une éventuelle intervention. Pensons à des situations telles que le pilotage d'un véhicule ou la défense contre une attaque de missiles. Les IA «froides», en revanche, laissent beaucoup plus de temps et demandent bien moins de ressources pour vérifier leur résultat.

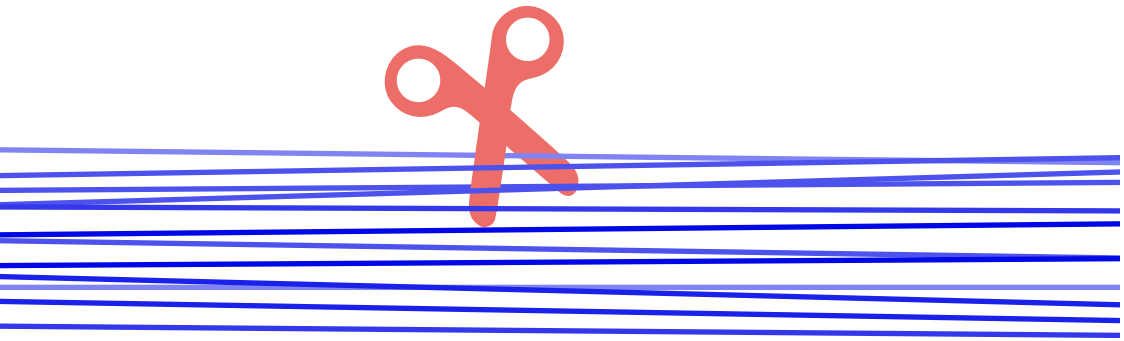
À titre d'exemple, la police pourra disposer d'assez de temps et d'effectifs pour examiner un résultat généré par IA, comme l'identification d'un suspect à partir d'une analyse et d'une évaluation de données biométriques. Les résultats d'un «grand modèle de langage», entre autres, peuvent aussi être évalués au moyen d'une recherche. À noter, toutefois, que ces deux catégories d'IA, «chaudes» et «froides», sont avant tout employées à des fins didactiques. En pratique, les frontières entre elles sont souvent floues. L'étude décrite dans la présente brochure portait principalement sur le contrôle des IA «chaudes».

Atterrissage par intelligence artificielle

Les résultats de l'étude ont été obtenus à l'aide d'expériences en sciences sociales. Plusieurs participants ont dû faire atterrir en toute sécurité un objet volant sur l'ordinateur du laboratoire au moyen d'une IA, l'assistance étant calibrée en fonction de leurs compétences. Le processus était «chaud» dans la mesure où un temps limité était imparti aux participants, qui étaient livrés à eux-mêmes.

Pour que l'expérience soit aussi réaliste que possible, une IA courante a été entraînée et adaptée aux circonstances. La situation créée reproduisait ainsi parfaitement des processus réels, comme l'atterrissage d'une capsule spatiale sur la lune, d'un drone sur un porte-avions, ou encore la conduite d'une voiture. La validité des résultats de l'étude est limitée à ces cas spécifiques.





Recommandations:

La mise en place d'obligations en matière de contrôle et de surveillance de l'IA doit tenir compte des réalités de chaque utilisation

Il est préférable d'exiger la surveillance et le contrôle humains de l'IA pour des cas concrets. Se borner à imposer des normes générales pour contraindre les exploitants à assurer une surveillance humaine de l'IA risque de s'avérer vain ou de ne pas donner le résultat escompté.

Premièrement, les systèmes d'IA sont non seulement très complexes, la plupart du temps, mais leur configuration et leurs fonctionnalités varient aussi très largement. Des exigences générales qui ne tiendraient pas compte des spécificités techniques et du contexte d'utilisation ont peu de chance d'être pertinentes. Deuxièmement, une réglementation trop vague laisserait une marge de manœuvre excessive aux exploitants.

Le contrôle et la surveillance de l'IA ne pourraient donc être garantis, de sorte que la responsabilité finale et la responsabilité juridique d'un éventuel résultat erroné ne pourraient que difficilement être attribuées à la personne interagissant avec l'IA qui prend la décision. Vue sous l'angle juridique et pratique, une obligation de contrôle ou de surveillance trop générale serait sans effet. Il convient donc d'exiger qu'elle soit conçue spécifiquement pour des champs d'application concrets, comme la médecine, la circulation routière, la finance ou la justice.

Sources (voir les références à la fin): (1), (2), (4), (5), (6).

Les possibilités d'intervention humaine dans les processus en cours basés sur l'IA ne sont pas une panacée

À première vue, les possibilités d'intervention humaine dans les processus continus (en cours), autrement dit «chauds», basés sur l'IA, semblent tout à fait logiques.

En pratique, de telles approches s'avèrent souvent inefficaces. Il est généralement difficile, pour une personne, de jauger les chances de succès et la fiabilité de systèmes d'IA. Elle aura souvent tendance à surestimer ou à sous-estimer ses propres capacités, à donner un crédit immérité aux résultats générés par l'IA (on parle d'automation bias) ou, à l'inverse, à les négliger sans raison valable (algorithm aversion).

Le niveau d'attention des êtres humains est, le plus souvent, très réduit. Comme les décideurs humains sont fréquemment le dernier maillon de la chaîne décisionnelle des processus basés sur l'IA (quand bien même ils ne rempliraient qu'une simple fonction de supervision), c'est à eux qu'incombent la responsabilité et l'obligation de rendre des comptes si les résultats sont erronés ou défectueux.

En d'autres termes, les décideurs humains font office de boucs émissaires en cas de dysfonctionnement du système d'IA.

Notons par ailleurs que les formations à l'usage de systèmes d'IA spécifiques n'augmentent pas nécessairement ni suffisamment les chances de succès du contrôle humain, et en particulier de l'intervention. Compte tenu des circonstances, de nouvelles approches doivent être mises au point au sein du régime de responsabilité civile.

Sources: (2), (4), (6).

L'explicabilité des systèmes d'IA hautement performants ne peut pas être acceptée telle quelle

Pour mieux contrôler les systèmes d'IA, il serait envisageable d'améliorer leur explicabilité et, pour ce faire, d'inscrire dans la loi une obligation d'explication.

Bien que le thème de l'explicabilité des systèmes d'IA (explainable AI, abrégé XAI) fasse actuellement l'objet d'intenses recherches, de nombreux systèmes très performants – notamment ceux de deep learning, qui reposent sur des réseaux neuronaux extrêmement complexes – restent inexplicables. Même si les données saisies (l'input) sont connues, les mécanismes qui permettent de générer un résultat concret demeurent inexplicables du fait de leur subtilité. Le système d'IA concret apparaît donc comme ce que l'on appelle une black box.

En conséquence, les obligations d'explication ou de transparence ne peuvent porter que sur les données saisies et/ou la conception fondamentale du modèle. Elles doivent également être modulées en fonction des circonstances d'utilisation et du public qu'elles visent.

En ce qui concerne les mécanismes existants qui visent à améliorer la prévisibilité des résultats de l'IA (les confidence scores ou heat maps), il est conseillé de vérifier leur efficacité dans un champ d'application concret. Enfin, il faut garder à l'esprit qu'une obligation d'explication approfondie et inscrite dans la loi constituerait un obstacle pour les systèmes d'IA hautement performants, en particulier les large language models.

Sources: (4), (7).

Une réglementation complète de l'intelligence artificielle aura des conséquences sur la capacité et la volonté d'innover

Les législateurs suisses pourraient être tentés de réglementer l'IA de manière globale («horizontale»), c'est-à-dire dans tous les champs d'application. C'est en tout cas la voie que l'UE a choisie. Une réglementation horizontale de l'IA charriera néanmoins son lot d'incertitudes dans un futur proche, surtout pour les petites et moyennes entreprises.

L'IA n'a, en effet, encore jamais été réglementée, et seul l'avenir montrera si la législation de l'UE fait ses preuves.

Toujours est-il que les doutes actuels pèseront rapidement sur la volonté et la capacité d'innover des entreprises.

Cette tendance négative sera difficile à infléchir, même lorsque la situation juridique se sera stabilisée. Une adaptation prudente des normes juridiques en vigueur à certains domaines d'application serait moins risquée. Il ne faut pas non plus oublier que les entreprises sises en Suisse et en Europe qui développent et utilisent l'IA à des fins commerciales (principalement les start-up) sont en général bien intentionnées.

La législation et l'administration devraient donc cibler en priorité les utilisations potentiellement suspectes de l'IA: par exemple, celles dont le fonctionnement technique est mis en cause, ou qui sont susceptibles de reposer sur des données acquises de manière illégale.

Sources: (1), (3).

L'intelligence artificielle est déjà largement abordée par le droit en vigueur

À de nombreux égards, l'IA nous met face à des difficultés inédites. Pourtant, les systèmes d'IA sont déjà largement encadrés par le droit en vigueur. En particulier, les droits fondamentaux, le droit de la protection de la personnalité et des données, le droit d'auteur, le droit du travail et le droit pénal s'appliquent tous aux systèmes d'IA et les encadrent au moins en partie.

Les autorités administratives et les tribunaux, mais aussi les particuliers, pourront, en outre, résoudre nombre des problèmes et des incertitudes qui se posent lors de l'application du droit en vigueur à l'IA en s'appuyant sur ce même droit.

Pour résumer, l'IA n'est pas complètement dérégulée et n'évolue pas dans un vide juridique. Elle présente néanmoins de nombreuses difficultés spécifiques auxquelles les législateurs devront s'attaquer. Ils devront en revanche résister à la tentation de régler la question par une loi trop générale.

Sources: (2), (3), (5).

Les répercussions sociales de l'intelligence artificielle doivent être traitées séparément de ses risques opérationnels

Les répercussions de l'IA se font sentir dans le monde du travail, l'éducation, la sécurité sociale, le partage des richesses et le modèle libéral. Dans ces circonstances, les organes politiques se retrouvent confrontés à plusieurs sortes de décisions.

1. Dans un premier temps, ils doivent déterminer si l'IA peut être indésirable selon le type d'utilisation, et donc à prohiber dans certains domaines.
2. Ensuite, il leur faut décider comment accompagner les changements induits par l'IA, notamment sur le plan social.
3. Enfin, ils doivent prendre en compte les risques d'exploitation concrets inhérents à l'IA (comme les discriminations cachées ou les problèmes d'influençabilité de l'IA, p. ex. de nouvelles formes de cyberattaques visant les données d'entraînement) et mettre au point des solutions réglementaires tangibles.

Sources: (1), (3), (5).

L'approche réglementaire la plus prometteuse: combiner les principaux généraux de base aux normes juridiques spéciales

Les changements nécessaires à apporter au droit positif pour mieux appréhender et réguler l'IA ne devraient pas être promulgués via un acte législatif général, une «loi sur l'IA».

Comme déjà expliqué, ce type d'approche est beaucoup trop abstrait et manque cruellement de précision. Qui plus est, le processus législatif qu'il demanderait serait beaucoup trop long.

En conséquent, dans un premier temps, des principes abstraits ne devraient être adoptés que pour servir de lignes directrices qui délimiteront les attentes générales en matière d'usage des systèmes d'IA, afin d'orienter l'économie, en particulier.

Parmi ces attentes figurent notamment la protection contre les discriminations (les «biais»), la robustesse des systèmes d'IA et un régime de responsabilité clair en cas de dommage. Les lignes directrices du Conseil fédéral de 2020 pointent dans la bonne direction et peuvent être adaptées à l'économie.

La prochaine étape consistera, pour les législateurs, à adapter graduellement les lois spéciales en fonction de l'urgence.

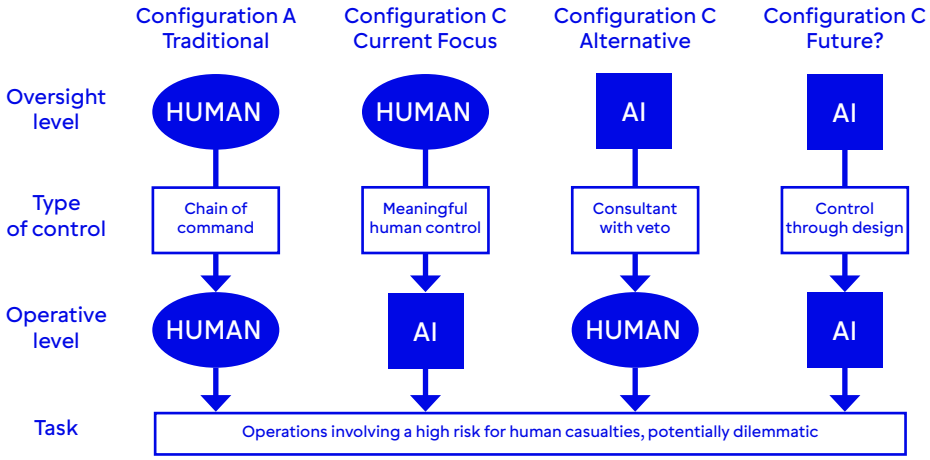
Sources: (1), (6).

La loi de l'UE sur l'intelligence artificielle ne devrait pas être reprise explicitement en Suisse pour le moment

La Suisse devrait s'abstenir, dans un premier temps, d'adopter explicitement le nouveau règlement de l'UE sur l'encadrement de l'intelligence artificielle. À la place, les politiques et les législateurs suisses devraient laisser aux entreprises la liberté de décider quand et dans quelle mesure se conformer à ce règlement. La Suisse pourrait ainsi conserver une certaine marge de manœuvre pour innover. Les législateurs devraient toutefois éviter les frictions avec le droit européen lors de la modification de lois spécifiques suisses.

Les autorités devraient informer sur les obligations découlant du règlement de l'UE et aider les entreprises (notamment les start-up) à s'y conformer rapidement pour leur ouvrir l'accès au marché européen.

Sources: (1).



Source: Markus Christen, Thomas Burri, Serhiy Kandul, Pascal Vöros, Who is controlling whom? Reframing "meaning human control" of AI systems in security, Ethics and Information Technology (2023) 25:10.

Références:

1. Thomas Burri, "A challenge for the law and artificial intelligence", Nature Machine Intelligence, 5, 1508–1509 (2023), https://doi.org/10.1038/s42256-023-00768-5 (paywall), prépublication: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4620764

2. Markus Christen, Thomas Burri, Serhiy Kandul, Pascal Vörös, "Who is controlling whom? Reframing 'meaningful human control' of AI systems in security", (2023) *25 Ethics and Information Technology* (10)

3. Emmie Hine, Yasaman Yousefi, Parisa Osivand, Dirk Brand, Kholofelo Kugler, Pier Giorgio Chiara, "The AI Act Grand Challenge shows how autonomous robots will be regulated", 8, *Science Robotics* 84 (2023), DOI: 10.1126/scirobotics.adk5632 (Paywall)

4. Serhiy Kandul, Micheli Vincent, Juliane Beck, Thomas Burri, François Fleuret, Markus Kneer, Markus Christen, "Human control redressed: Comparing AI and human predictability in a real-effort task", (2023) 10 *Computers in Human Behaviour Reports* (May 2023, 100290), https://doi.org/10.1016/j.chbr.2023.100290

5. Thomas Burri, Markus Christen, Juliane Beck, Daniel Trusilo, "Eight Recommendations for Ethical and Legal Assessments of Robotic Systems Interacting with Humans", in Woodrow Barfield, Ugo Pagallo, et Yueh-Hsuan Weng (éd.), Research Handbook on the Law, Regulation, and Policy of Human Robot Interaction, CUP 2024 (à paraître), prépublication: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4437806

6. Juliane Beck, Thomas Burri, "From 'Human Control' in International Law to 'Human Oversight' in the New EU Act on Artificial Intelligence, in Daniele Amoroso et Filippo Santoni De Sio (éd.), Research Handbook on Meaningful Human Control of Artificial Intelligence Systems, Elgar, 2024 (à paraître), prépublication: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4236554

7. Serhiy Kandul, Vincent Micheli, Juliane Beck, Markus Kneer, Thomas Burri, François Fleuret, Markus Christen, "Explainable AI: A Review of the Empirical Literature" (en préparation), version provisoire: https://ssrn.com/abstract=4325219

SNF Beitrag Nr. 407740_187494 / 1
Armasuisse W+T Vertrag Nr. 8003535283, 8003538711, 8003539311

Imprimer:
Mise en page et illustration: Malte Euler
Police: TOS Sans par ShowMeFonts
Papier: IGEPA Circle offset Premium White Recycle

Schweizerischer Nationalfonds
Universität St.Gallen
UNIVERSITÉ DE GENÈVE
Universität Zürich[UZH]