Chers membres du réseau de l'ancienne FSFA, chers intéressés,

Les deep fakes – fichiers audio, vidéo et image falsifiés par IA – ont un potentiel de manipulation considérable. Les bases de la formation démocratique de l'opinion, notamment lors des élections, sont ainsi touchées. L'esprit critique et la remise en question des informations deviennent de plus en plus importants. Les institutions de formation de tous les niveaux sont sollicitées. Que faire si les deep fakes ne sont pas reconnus comme tels ? Et si l'incertitude quant à ce qui est fake et ce qui est authentique érode la confiance dans les processus politiques et, en fin de compte, dans l'État ? Voici quelques exemples à ce sujet.

**Débats avec des modèles d'IA – plus convaincants que les discussions avec des humains ?**
Une étude réalisée au Data Science Lab de l'EPFL sur la force de persuasion des humains par rapport aux modèles d'IA est arrivée à un résultat surprenant, ici. Sur un total de 850 participants, ceux qui ont débattu avec GPT-4 ont beaucoup plus facilement changé d'avis que ceux qui ont discuté avec des humains ; et ce d'autant plus nettement, avec une probabilité supérieure d'environ 82 pourcent, lorsque l'IA a pu accéder aux données personnelles des participants et argumenter sur mesure. Selon le professeur Robert West, directeur du laboratoire, il faut s'attendre à ce que de tels modèles linguistiques soient utilisés dans les campagnes électorales.

**Marqué comme vidéo générée par l'IA – mais grand effet malgré tout**
Une fausse vidéo bien réalisée (ici) a fait déclarer à la première ministre danoise Mette Frederiksen lors d'une conférence de presse que tous les jours fériés chrétiens seraient supprimés ; à l'avenir, il n'y aurait plus qu'un seul jour férié, la fête musulmane de la rupture du jeûne. La vidéo comportait en petit logo la mention « généré par l'IA » et l'auteur apparaissait à la fin sur l'image. C'était son rêve – sous la forme d'une satire. Il semble néanmoins que de nombreuses personnes soient tombées dans le panneau de la fausse vidéo, ce qui a déclenché un débat enflammé.

**L'IA dans les élections en Inde du 19 avril au 1er juin 2024**
Selon différents rapports médiatiques (p. ex. Reuters, swissinfo), il s'avère déjà que dans la plus grande démocratie du monde (environ 970 millions d'Indiens, plus de 800 millions d'utilisateurs d'Internet), la lutte contre la propagation des deep fakes est très difficile. Le gouvernement et l'opposition s'accusent mutuellement de manipulations dans les médias sociaux. De plus, de véritables bandes de faussaires s'en mêlent. L'Internet est surveillé de près, des contenus sont supprimés et des arrestations sont effectuées, mais il est difficile d'avoir une vue d'ensemble de ce qui se passe. Les deep fakes ont généralement déjà eu un impact, même s'ils sont rapidement découverts et retirés des plateformes. Il sera intéressant de procéder à des évaluations a posteriori.

**Il est possible de lutter contre les fake news – comme dans le swing state de l'Arizona**
Un petit studio de radio à Phoenix, Arizona, réfute en permanence les fake news qui sont importantes pour la campagne électorale américaine à venir. Radio Campesina s'adresse en espagnol aux Latinos, qui représentent environ un tiers de la population de l'Arizona. Les Latinos tirent principalement leurs informations des médias sociaux inondés de deep fakes. L'agence de presse américaine AP (Associated Press), basée à New York, a rapporté les faits, ici.

**Accord contre les manipulations électorales par l'IA**
Ce sont pourtant les géants de la technologie qui devraient être tenus d'agir. En février 2024 déjà, à l'occasion de la conférence de Munich sur la sécurité, plusieurs géants de la technologie, dont Google, Meta, Microsoft, OpenAI et TikTok, ont signé un accord contre la désinformation manipulatrice lors des élections, l'AI Elections Accord. L'accent a été mis sur les deep fakes. Il existe plusieurs déclarations d'intention des géants de la technologie sur les questions de sécurité de l'IA. Il y a peu, lors du sommet de l'IA des 21 et 22 mai 2024 à Séoul (Corée du Sud), 14 entreprises d'IA ont à nouveau signé un document correspondant, ici.

Mais où sont les effets de ces accords ? Qui doit assumer quelle responsabilité – les géants de la technologie, la communauté des États, la société civile, les institutions éducatives ?

Avec nos salutations les meilleures,
Pour le réseau de l'ancienne FSFA : Hanna Muralt Müller

***Nouveau droit de la protection des données : Si vous ne souhaitez plus recevoir cet e-mail, veuillez me contacter !***