Fighting deskilling in Human–AI systems: Behavioral experiments and interventions in groups for accurate evaluation of information

Name(s) of PhD Advisor(s): Jean Claude Dreher
Host Laboratory: Institut des Sciences Cognitives, CNRS

Abstract

Concerns have been raised about AI over-reliance, cognitive offloading and the erosion of critical cognitive skills. Deskilling describes the loss of skills due to technological changes, including the use of AI system. Here we focus on skills required to make decisions in networked systems in which AI overreliance may reduce the ability to critically evaluate the veracity of news headlines. First, we will test AI-assisted groups (15 teams, 150 participants) and Human only groups (15 teams, 150 participants). We hypothesize that : (a) the AI-assisted groups, who receive AI-generated credibility scores for each news headline, will lead to deskilling (reduced critical thinking) over time (4 weeks period); (b) further AI removal leads to reduced critical thinking engagement compared to the Humans-only Group ; (c) The Human-Only Group will demonstrate better skill retention and more effective discussion strategies in the final phase of the task. This first experiment should demonstrate experimentally that assessment of news veracity with AI can lead to reduced critical thinking. Second, we propose a behavioral experiment to mitigate the deskilling effects of AI-assisted news evaluation. To do this, we will implement 3 psychological inoculation interventions, consisting in preemptive cognitive training that helps individuals to resist misinformation and over-reliance on AI. Together, these large experiments and novel inoculation strategies should show how to strengthen critical thinking in networked groups with or without AI assistance.

Short Description of the Hosting Research Group / Lab

The **Dreher lab** is an interdisciplinary research team at **the Institute of Cognitive Science (CNRS)** dedicated to advancing knowledge in **computational social neuroscience**, and **human–AI collaboration** (https://dreherteam.wixsite.com/neuroeconomics). Our group brings together expertise in: **Multi-Agent and Reinforcement Learning.** Developing, training, and analyzing sophisticated AI agents that learn to interact with humans or other agents in complex social environments. Designing new reinforcement learning frameworks that address ethical and cognitive concerns within AI-assisted collaboration. **Large-Scale Human Experiments.** Conducting rigorous online and in-lab user studies to explore emergent behaviors in social dilemmas, crowdsourcing tasks, and collaborative problem-solving. Possessing an established infrastructure (web-based platforms, participant databases, IRB/ethics approvals) for running repeated multi-user experiments in real-time decision-making. **Human–Computer Interaction and Ethics.** Investigating the long-term impact of AI on user skill retention and trust. Designing transparent interfaces, interpretability mechanisms, and guidelines for responsibly deploying AI in real-world collaborative settings.

Why This Lab is a Suitable Environment

Methodological Support: The lab's prior experience in building multi-agent simulators and running large-scale behavioral experiments ensures a robust infrastructure for the proposed PhD work.

Interdisciplinary Expertise: The diverse backgrounds of the lab's members—including computational social science, cognitive psychology, social neuroscience and machine learning—will support a holistic investigation into both the technical and human-centric aspects of AI-driven collaboration.

Active Research Community: The lab maintains collaborations with other academic groups and international networks focusing on societal impacts of AI. This environment offers ample opportunities for cross-disciplinary seminars, workshops, and mentorship that will enrich the PhD candidate's academic experience.

Relevance to Theme 3: JC Dreher has a proven track record of addressing how advanced technological systems affect human behavior and cognition in the field of social networks. He is the coordinator of the project PEPR eNSEMBLE 'COMCOMBrr' that aligns strongly with Theme 3's emphasis on critical thinking, expertise retention, and socio-technical resilience. By joining this lab, the PhD researcher will complement the planned post-doc funded by COMCOMBrr to work on a related and complementary project. He will gain access to cutting-edge resources and a supportive intellectual environment, equipping them to investigate how AI systems can both enhance cooperation and preserve (or even bolster) fundamental human skills.

Description of the PhD Proposal

Context (and Scenarios)

To be informed citizens in today's information-rich environment, individuals must be able to evaluate the truthfulness of the information they are exposed to on the Internet. For example, in the context of education, school teachers currently have limited options if they are willing to assess the students' evaluation of digital content. Recent studies assessing adolescents' ability to effectively search for, evaluate, and verify social and political information available online found that students struggled to effectively evaluate online claims, sources and evidence (McGrew et al., 2018). Fake news develop and disseminate widely, particularly when the nature of the information is difficult to evaluate or ambiguous.

Problem and Objectives

Recent research has shown that while AI-based interventions can dramatically enhance collaborative outcomes—ranging from higher efficiency and stronger cooperation to increased fairness—they may unintentionally reduce people's capacity for independent reasoning and adaptability (Parasuraman & Riley 1997; Bainbridge 1983; Lee & See 2004; Frey & Osborne 2017; Daugherty & Wilson 2018). This phenomenon, often referred to as "deskilling," becomes especially problematic in organizational and educational contexts where users risk becoming passive executors of AI-generated directives rather than active contributors to strategic planning. If individuals grow reliant on AI for guidance, their ability to innovate, solve problems, and make nuanced decisions may diminish over time, jeopardizing both immediate and long-term performance at the personal and institutional levels.

The risk of deskilling is rea when letting AI lead decision-making roles: users may gradually lose their critical thinking abilities, problem-solving initiative and collaborative skills (Rafner et al., 2021; Parasuraman & Riley 1997; Endsley & Kiris 1995; Carr 2011; Crowston and Bolici, 2025). Yet, little is known about deskilling in humans-AI in networked groups assessing veracity of information. Here we aim to study experimentally the loss of the capacity to assess the veracity of information. Critical thinking might atrophy when an AI system takes over significant portions of the decision-making process. As humans become less involved in evaluating the veracity of the news they read with AI assistance, they may lose the very competencies needed to adapt in future scenarios without AI support. We will focus on the decline of human critical thinking skills and we will propose intervention methods to foster and encourages critical thinking.

Brief Overview of the State of the Art

A large meta-analysis has reported that on average, human–AI combinations perform significantly worse than the best of humans or AI alone (Vaccaro et al., 2024; Burton et al., 2024). Moreover, the effectiveness of synergy between humans-AI systems depends upon the type of tasks. Decisions tasks have been associated with performance losses and creation tasks were associated with performance gains, when compared to humans alone or the AI alone. In this PhD proposal, we aim to examine how affect not only short-term group performance but also the long-term expertise of human participants. Our plan involves running controlled experiments in collaborating groups to measure shifts in critical thinking over time. We will also design and evaluate inoculation interventions to mitigate the potential erosion of user skills in assessing news veracity. By integrating perspectives from wisdom of the crowd, group decision making, computational social science/economics and cognitive psychology, this research aims to test how hybrid AI systems supposed to help us to formed informed judgments can be detrimental to essential human competencies.

Research Questions

A central concern driving this PhD project is how prolonged reliance on AI-assisted collaboration may lead to a phenomenon of *deskilling*, in which human participants gradually lose their to develop critical thinking. We ask three specific questions on AI reliance to judge news veracity when making collaborative decisions in groups. **What mechanisms drive deskilling in AI-assisted collaboration?** On the surface, AI-generated recommendations can appear so compelling—or so convenient—that humans simply follow them without deeper reflection. This raises the possibility that people are outsourcing critical thinking to the AI, ultimately eroding their own capacity for strategic exploration and initiative. Understanding *why* this occurs—whether it stems from overtrust, cognitive offloading, or the AI's inherent efficiency—lies at the heart of this research. **How can we measure and assess loss of critical thinking when assessing veracity of information in networked groups?** To rigorously evaluate deskilling, developing or adopting indicators that track user-driven proposals, override their

own beliefs in groups, and other facets of critical thinking becomes necessary. By designing consistent and reliable measures, the project aims to pinpoint exactly when and how user competence shifts over repeated interactions with AI. **Which design strategies preserve human expertise?** Given the potential risks of deskilling, the final question is one of *solution design*: Can we create interfaces and protocols that maintain or strengthen the human capacity of evaluating veracity of information in our world full of desinformation, even as AI becomes more adept at guiding group decisions? Proposals such as inoculation interventions might encourage users to remain cognitively involved, reducing the temptation to adopt AI outputs mindlessly. Determining which strategies effectively balance efficiency gains with sustained user agency will provide practical guidelines for AI deployment in collaborative contexts where multiple agents interact to assess news veracity.

Theoretical Foundations. We hypothesize that people make Bayesian inferences to make judgments about information. That is, they integrate their own beliefs and their peers' beliefs about the information value, then weigh both types of beliefs to decide whether they propagate the information in their social network. Moreover, individuals' decisions to disseminate information to their peers may rely on deception, such as aiming at modifying their peers' beliefs by sending undesired information. We thus expect participants to form different strategies of inferring beliefs, weighting beliefs and making decisions of information dissemination. We have recently developed theoretical models to implement such Bayesian inferences in group decision making during social dilemma (Park et al., 2019; Khalvati et al., 2019; Philippe et al., 2024). We plan to develop similar Bayesian approaches to the current experiments.

Approach and methods

Behavioral experiment in Networked Groups when evaluating information veracity: The objective of this experiment is to examine how AI-generated credibility scores influence collective decision-making in a networked group and whether reliance on AI reduces participants' ability to critically evaluate news articles independently over time. We will test 300 adult participants divided into 30 online groups (10 members per group). Each participant will interact within a social media-like network where they evaluate news articles. There will be 2 main conditions (Between-Subjects Design): (1) AI-Assisted Group (15 teams, 150 participants) who receive AI-generated credibility scores (e.g., a trust rating from 0 to 100) for each news article. To do this, AI will provide reasoning based on common fact-checking databases; (2) Human-Only Group (15 teams, 150 participants) who will have to evaluate news articles without AI assistance, relying only on their own critical thinking and discussion. The task will be built based on a variant of a task recently developed by our research team (Guigon et al., *Comm. Psychol*., 2024). This task has previously been used to show that metacognition biases information seeking in assessing ambiguous news. Now, for this PhD proposal, participants will be presented with the same ambiguous news already carefully selected to assess 3 themes: ecology, democracy, social justice. They will be asked if the news is true or false and to indicate their confidence in this news veracity assessment (allowing us to assess both veracity judgment and confidence in one's judgment). Over 4 weeks, participants will assess the credibility of 100 news articles (both real and fake) in a simulated news-sharing platform (using slack). **In Phase 1 (Training - Week 1)** all participants will receive an initial fact-checking tutorial on common misinformation tactics. The AI-Assisted Group will be introduced to the AI tool, while the Human-Only Group practices manual evaluation. **In Phase 2 (Evaluation Period - Weeks 2-3)** participants will review news articles, discuss their credibility in team chats, and vote on whether each article is real or fake. The AI-Assisted Group will see AI-generated credibility scores before making their decision. In contrast, the human-Only Group will make independent, discussion-based evaluations. Finally, in **Phase 3 (AI Removal Test - Week 4)** AI will be disabled for all teams. Participants will now have to assess a final set of 20 news articles without AI assistance. Their performance is compared to earlier decision accuracy. We will measure accuracy in news veracity (ie how well participants identify misinformation compared to fact-checking sources), cognitive Load (perceived difficulty of evaluating news). In addition, group discussion patterns will be analyzed using audio to text LLMs. This will allow us to test whether participants engage in high-level reasoning or default to AI recommendations. We will also test how frequently the AI-Assisted Group blindly follows AI suggestions without discussion (defining a AI Dependence Score) and will develop a post-experiment skill Retention Test in which, after a 2-week break, participants re-evaluate another batch of news without AI to measure long-term deskilling effects. We hypothesize that the AI-Assisted Group will initially make faster and more accurate veracity judgments. However, after AI removal, they will show higher error rates and reduced critical thinking engagement compared to the Human-Only Group. The

Human-Only Group will demonstrate better skill retention and more effective discussion strategies in the final phase. We expect to show experimentally how AI weakens independent critical thinking skills when evaluating misinformation. This should highlight the risks of over-relying on AI-generated fact-checking rather than human reasoning. This should lead to the second part of the PhD to design AI systems that prompt user engagement instead of passive trust (e.g., requiring explanations for user decisions).

Intervention to reduce deskilling of AI-assistance in assessment of the information veracity. To mitigate the deskilling effects of AI-assisted news evaluation, we can implement psychological inoculation interventions—preemptive cognitive training that helps individuals resist misinformation and over-reliance on AI (Maertens et al., 2025). Below are three progressive inoculation strategies designed to strengthen critical thinking over time. They consist in 3 phases. **Phase 1: Cognitive Exposure Training (Pre-Task Inoculation).** The objective of phase 1 is to build awareness of AI limitations and train participants to detect misinformation before exposure to AI-generated credibility scores. The intervention approach consists in : (1) Interactive Debunking Workshop (Week 1). Participants engage in a "Spot the Fake" game where they analyze real vs. fake news without AI assistance. They receive real-time feedback explaining why certain articles are deceptive; (2) Misinformation Simulation. Participants will create their own misleading headlines and fake articles, learning the tactics used in misinformation campaigns; (3) AI Fallibility Demonstration. It will show how AI fact-checking systems can be biased or manipulated (e.g., examples of AI mistakenly flagging real news or missing false claims). The psychological mechanism of interest is the prebunking Effect which consist in exposing participants to weakened misinformation attempts, so that they build mental resistance to AI suggestions. **Phase 2: Active Critical Engagement (During AI-Assisted Task).** Here we will encourage participants to maintain critical reasoning skills while using AI, instead of blindly following credibility scores. The intervention Approach will consist in : (1) AI Transparency Prompts. AI does not directly provide credibility scores but instead asks: "What makes this article credible?, "What evidence supports or contradicts this claim?"; (2) Justification Requirement. Participants will have to explain their reasoning before seeing AI-generated scores. AI will only reveal its credibility rating after the user submits their independent assessment.; (3) Peer Discussion Encouragement. A system will highlight when group members have conflicting judgments, encouraging further discussion before reaching consensus. The psychological Mechanisms under study here will be cognitive Effort and accountability. This will force participants to engage in deliberate analysis, reducing passive AI dependence. **Phase 3: AI-Free Reflection & Skill Reinforcement (Post-Task Inoculation).** We will test and reinforce long-term independent evaluation skills after exposure to AI assistance. To do this, our intervention Approach will be : (1) AI-Free Challenge (Final Week). Participants evaluate news articles without AI support and compare their decisions to previous AI-assisted ones. Feedback highlights where they followed AI blindly vs. where they made independent judgments; (2) Misinformation Resistance Scoring. A "News Credibility Score" is assigned to participants based on their accuracy, reasoning depth, and resistance to misleading AI suggestions. Delayed Testing (One Month Later): participants will complete a follow-up misinformation evaluation task to assess long-term skill retention. Here the investigated psychological mechanisms will be self-reflection which encourages metacognition (Guigon et al., 2024)., ensuring participants internalize critical evaluation skills rather than offloading the task to AI.

Evaluation of the Contributions. By systematically observing how different networked environments with or without AI assistance affect user skill of critical thinking when facing misinformation, the PhD project aims to make several key contributions. Empirically, it will furnish quantitative evidence on whether certain AI-driven collaboration improvements correspond to unintended deskilling—or conversely, identify design features that help safeguard user expertise. The key methodological improvements of this PhD proposal include (1) a custom-made online collaboration platform designed as a social network that allows for detailed experimental control and data collection; (2) the use of modern machine learning tools for natural language processing and quantitative semantic analysis that were used in both experimental manipulation and resulting data analysis. Finally, our planned interventions should reverse deskilling of assessing veracity of information under AI assistance. Theoretically, the study will develop novel Bayesian models to account for how initial individual beliefs about news veracity develop over time under group and AI influences, thereby shedding light on how human-AI systems can preserve essential human competencies in the face of accelerating AI adoption.

Nature of Digital Collaboration

The function of collaboration in this PhD project revolves around communication, information sharing, coordination in groups, cooperative problem-solving and assessing veracity of information, where AI shape group judgments about news veracity. Conducted online in real time, the platform enables synchronous interaction, with the AI offering recommendations on news veracity, thereby potentially reducing critical thinking. Participants remain free to accept, reject, or override these suggestions, thus creating a human-in-the-loop workflow.

Each session, lasting roughly 45–90 minutes, will accommodate small to medium groups of participants (15 people on average) and may extend over multiple encounters to assess longitudinal aspects of skill retention. The spatial dimension is fully remote, allowing individuals from diverse locations to join via a browser interface or via slack. The platform itself is designed to be scalable (thus a few hundred of individuals could be tested simultaneously in future studies with the same approach), supporting dynamic updates and integrated real-time data collection.

The time scale of the experiments last a few weeks (repeated test-retest to assess loss of critical thinking over time).

Key features include transparency mechanisms for explaining AI recommendations or feedback loops that enable participants to defend their own beliefs.

Collectively, these elements reflect the real-world conditions of modern distributed teams and online communities. By embedding advanced AI recommendation systems with varying levels of autonomy, the project directly addresses Theme 3's core concern: whether users gradually lose vital skills through repetitive AI-led collaboration or find new ways to sustain and even enhance their capabilities in a digitally mediated, ever-evolving environment.

Contribution to Digital Collaboration: Expected Results and Impact

This PhD project expects to illuminate the interplay between AI-driven collaboration and human skill retention, offering nuanced findings for hybrid intelligence and digital teamwork. At the *experimental* level, we expect reduced AI Over-Reliance: participants will become less dependent on AI credibility scores and engage in more independent verification**.** In addition, we expect stronger Long-Term Misinformation Resistance. Through spaced inoculation and active reasoning, individuals will retain fact-checking skills over time**.** Finally, we expect to observe more effective Collective Decision-Making. That is, group discussions should shift from blind trust in AI to critical engagement and cross-verification**.**

From a *theoretical* standpoint, the project seeks to integrate computational social science perspectives on group decisions with cognitive psychology insights on skill retention. As a result, it will advance frameworks for cooperative multi-agent systems that explicitly model user cognition, thereby helping to bridge the gap between AI optimization and the preservation of human competencies. In doing so, it will also make *methodological* contributions by introducing metrics better suited for studying long-term critical thinking ability dynamics, potentially forming an open-source toolkit for future researchers.

These findings will yield *practical design guidelines* to clarify how AI influences both group performance and the skills to assess information veracity, which should help educators and policy-makers navigate the trade-offs between AI-assistance and preserving essential human capabilities. Aligned with Theme 3 of PC3, these insights will demonstrate how advanced AI interventions can be harnessed not merely help making group decisions.

Positioning in the eNSEMBLE Program

Within the eNSEMBLE Program, this project speaks directly to **Theme 3**, which focuses on ensuring that advanced AI systems do not inadvertently weaken human cognitive and collaborative skills over time. By empirically evaluating whether "explainable AI" or forced override strategies can preserve expertise, the research takes on the very question of deskilling in AI-assisted group tasks. This, in turn, ties into **Themes 1 and 2** on coordination and collaboration: the experimental setup related to misinformation and group decision tasks resonates with eNSEMBLE's wider mission of understanding how large groups leverage AI for collective action.

The program's emphasis on **hybrid socio-technical solutions** finds a direct parallel in this project's exploration of Bayesian inferences. The research addresses eNSEMBLE's goal of adopting ethically and cognitively conscious AI strategies to reduce misinformation on-line and inform the public about skills to acquire for reducing susceptibility to AI-mediated misinformation. Methodologically, it aligns with the program's commitment to **large-scale online experiments**, harnessing repeated, digitally mediated

tasks that incorporate real-time rewiring options—thus giving the project a strong practical and ethical dimension.

Moreover, the PhD study fits eNSEMBLE's **interdisciplinary perspective**, drawing on computational social science, cognitive psychology, and AI design to explore autonomy and competence in human–AI interaction. The findings will also be relevant to parallel eNSEMBLE activities investigating trust, transparency, and governance, fostering collective insights on how best to harness AI in team-based and community-driven contexts.

By spotlighting the long-term societal impact of AI-driven collaboration, the project offers **concrete solutions** to the pressing concern of sustaining human skill such as critical thinking and judgments about veracity of information in a connected world. It thereby enriches eNSEMBLE's broader objective of building socio-technical systems that reinforce human capacities instead of eroding them, ensuring that even as AI takes on more significant roles in decision-making, people remain active and adept in the processes that shape their collective future.

Finally, this proposal nicely complements recent work by Dreher's team about assessing the veracity of ambiguous news, and the role of metacognition for guiding our decisions to seek further information (Guigon et al., 2024). By allowing participants to freely communicate and interact about news veracity, this proposal also complements more artificial settings developed by the PEPR COMCOMBr which aims to understand the influence of bots to experimentally controlled 'artificial' social networks.

References

- Bainbridge, L. (1983). Ironies of automation. In J. S. Annett (Ed.), Handbook of Man-Machine Systems (pp. 7–17). Wiley.
- Botvinick, M., et al. (2023). Scaffolding cooperation in human groups with deep reinforcement learning. *Nature Human Behaviour*.
- Burton, J.W., Lopez-Lopez, E., Hechtlinger, S. *et al.* How large language models can reshape collective intelligence. *Nat Hum Behav* **8**, 1643–1655 (2024).
- Carr, N. (2011). *The Shallows: What the Internet Is Doing to Our Brains*. W. W. Norton & Company.
- Crowston, K., & Bolici, F. (2025). Deskilling and upskilling with AI systems. Information Research an international electronic journal, 30(iConf), 1009-1023.
- Daugherty, P. R. & Wilson, H. J. (2018). Human + Machine: Reimagining Work in the Age of AI. Harvard Business Review Press.
- Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors*, 37(2), 381–394.
- Guigon, V., Villeval, M.C. & Dreher, J.C. (2024). Metacognition biases information seeking in assessing ambiguous news. *Communications Psychology*, 2, 122.
- Frey, C. B. & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? Technological Forecasting and Social Change, 114, 254–280.
- Lee, J. D. & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80.
- Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.
- Paiva, A., Santos, F. P., & Santos, F. C. (2018). Engineering pro-sociality with autonomous agents. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems* (AAMAS '18), 7994–7999.
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253.
- Rafner et al., Deskilling, Upskilling and Reskilling: a Case for Hybrid Intelligence, *Morals and Machines*, 2021
- Rand, D. G., Arbesman, S., & Christakis, N. A. (2011). Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences*, 108(48), 19193–19198.
- Sajid, N., Ball, P. J., Parr, T., & Friston, K. J. (2019). Active inference: Demystified and compared. *arXiv preprint*, arXiv:1909.10863.
- Shirado, H., & Christakis, N. A. (2020). Network engineering using autonomous agents increases cooperation in human groups. *iScience*, 23(6), 101438.
- Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 1-11.