# Research Statement

## Sahar Abdelnabi

We live in an extraordinary time for AI research; we frequently witness the onset of groundbreaking and disruptive innovations in AI or enabled by AI. One of the most exciting potentials of AI is the dream of having general-purpose models that are adaptable and extensible with easily accessible interfaces. AI has great potential to scale, automate, and accelerate decision-making. It's imaginable that AI agents would autonomously cooperate and interact in the real world to perform various tasks. However, how can we ensure that these new technologies are safe, secure, and trustworthy and minimize their frictional effects on our society?

**My research interests, since I started my PhD in 2019 at the CISPA Helmholtz Center for Information Security and now as an AI security researcher at Microsoft, lie in the intersection of AI, safety, security, and sociotechnical problems.** My goal is to develop robust safeguards for risks while reaping the benefits of AI. I realized that in order to achieve this grand challenge, responsibly developing safe AI must be treated as a simultaneous core priority that is at least equally important as improving capabilities.

Aiming to holistically study the trustworthiness aspects of AI, I worked on a large breadth of topics and pioneered now-very-active research directions with both conceptual and technical contributions. **I am interested in the question of "context"**: how methods to protect the authenticity of context can counter generative AI risks [1,2], how AI can help uncover the authentic context [3], and how the lack of contextual reasoning manifests as various risks [4,5]. **My work called for and proposed the first technical watermarking methods to counter the misuse of generative AI** [1,2], and methods to **automate multi-modal fact-checking** to face societal challenges of misinformation and context-hijacking [3]. **I studied poisoning attacks that tamper with the factuality of online information** to manipulate fact-checking models [4], a problem that is now even critical with LLM-integrated search engines and applications.

**My work [5] had significant industry and research impact to uncover, coin, and disclose (to Microsoft and OpenAI) the "indirect prompt injection" vulnerability in LLM-integrated applications**; that is how LLMs can be prompted with instructions injected by third-party data, deviating from users' original instructions and potentially causing various harms to users, a problem that is now heavily researched in industry (my team at Microsoft included) and academia. **This may be the most realistic, practical, and impactful attack vector to deployed AI systems** that may cause real, catastrophic harm, given the scale of users and how easy it is to mount these attacks. I further worked on the prompt injection problem by **co-designing public competitions [7,10], defining the phenomenon of the lack of data-instruction separation** that enables such attacks [9], and proposing **interpretability-based solutions to formulate indirect prompt injections as "Task Drift" and detect it** based on models' internals states [8], a method that we found to be much more robust compared to text-based classifiers on LLMs' inputs.

Beyond single models, future AI agents may autonomously make decisions or interact with humans or other agents. Current static evaluation benchmarks are inadequate to evaluate models' performance. **My work [6] proposes negotiation games as a dynamic, interactive, and adaptable benchmark** to test reasoning, arithmetic skills, Theory-of-Mind, contextual privacy, cooperation, competition, and the potential for manipulation between LLM agents. On the other hand, understanding implicit biases is key to ensuring fair and trustworthy decision-making. Analogous to cognitive human studies, **our work [11] shows that LLMs have perspective bias** when prompted to pick a value from a distribution; instead of reporting the statistically average value, their answer is biased by an ideal notion either learned in pre-training or context; **a phenomenon that may explain other observed unwanted behaviors such as sycophancy bias (responses that match user's belief).**

Going back to the promise of AI to improve our lives, I believe **current LLMs can be used to harness knowledge and propose creative hypotheses and solutions that can be verified by other means and be integrated as human-in-the-loop assistive tools** in, e.g., scientific discoveries and vulnerability discovery. Our work [12] shows how LLMs can be used to hypothesize missing variables in partially known causal graphs to find mediator variables between a cause and its effect, enhancing the underlying causal mechanism. I also co-mentored an internship project at Microsoft to use LLMs to automatically generate

CodeQL queries and find variants of a code vulnerability across a codebase. Both of these tasks are heavily dependent on domain and expert knowledge, which can be augmented by LLMs as assistive proxy experts.

In addition, my experience at the Microsoft Security Response Center gave me an additional much-needed and impactful perspective on the security of end-to-end systems, in-the-wild vulnerabilities, and the practical needs of proactive mitigations informed by them. My team is also co-responsible for setting severity classification guidelines for AI safety and security risks assessment. Based on these guidelines, we give recommendations on reported incidents to shape the development of safe and secure products.

**My work as a first/co-first author was published at top Security, Vision, and ML conferences, won a *best paper award* at AISec workshop, and was featured as a BlackHat talk and in press.** My experience both in my PhD and now in industry well positions me to pursue, lead, and have a real-world impact towards creating trustworthy, safe, aligned, and fair AI systems and agents. I below highlight more details about some of my previous work, and discuss my future research vision.

## 1- Overview of Previous Work

### 1.1- Misuse of generative AI

**Attribution as a question of context.** One of the major challenges we face now as a society is **identifying and protecting the context of information**. However, passive detection of synthetic media (i.e., real vs. fake classification) is not sustainable in the long run as the gap between the two distributions gets narrower. Therefore, even in 2021, with way less impressive models, we realized that we need more sustainable alternatives. Watermarking can be used to add "contextual cues" (i.e., the author) to the generated media. In my work [1,2], we were the first to advocate for the **responsible disclosure** of generative models by **watermarking their output to enable attribution**, an idea that now is active among researchers in academia and industry and has been discussed as a potential regulation for generative AI [a].

**Watermarking as a proactive defense for Gen AI misuse.** We developed Proof-of-Concept watermarking schemes for text and images by learning to simultaneously encode and decode a binary message while minimizing the effect on the utility. We set the foundation for this task by outlining the framework and evaluation metrics, discussing the properties of the required solution (e.g., secrecy, robustness, and utility preservation), and validating these properties in our scheme by also proposing counterattacks. I am glad to see a long list of follow-up work that continues on our ideas by using state-of-the-art LLMs, proposing more robust methods, and watermarking other generative models. I am hopeful about the use of watermarking in some use cases, especially in the absence of motivated adversaries, e.g., to watermark generated text to exclude it from training (to avoid model collapse [b]) or accidental contamination of generated content in LLM-integrated search engines.

### 1.2- Automated multi-modal fact-checking

**AI to help uncover the authentic context.** The proliferation of misinformation constitutes a major threat to our societal safety and democratic values, especially in times of, e.g., elections and political instabilities. **Mechanisms to identify the context are needed for settings where misinformation continues to exist and spread without the use of generative AI.** Human fact-checking efforts may unfortunately not be able to cope with the scale of online misinformation, and therefore, **automation may help journalists and users. Automated fact-checking can help retrieve and prioritize online evidence and scale the decision of veracity in an interpretable, inspectable way.** My work [3] was the first to propose multi-modal fact-checking via verifying captioned images against online evidence in order to detect Out-of-Context images, where here the goal is to identify and uncover the true context of an image. We design an inspectable framework that automates this multi-modal fact-checking by aggregating evidence from images, articles, and different sources, and measuring their consensus and consistency. **Our method can highlight the most relevant evidence, which helps people distinguish true from falsified pairs, as we found after conducting a large-scale user study.**

### 1.3- Evidence Manipulation to tamper with the factuality of language models

**The trustworthiness of sources is a "context" problem. While automated fact-checking is promising, it can be vulnerable to online evidence manipulation,** which we studied [4] before the onset of the current wave of large language models chatbots. Our attacks could partially rewrite sentences such that they would have the targeted stance (e.g., neutral or supporting) toward the claim. Attacks may also generate entirely new evidence from scratch with required properties. In line with typical ML adversarial attacks, attacks might also be imperceptible by using encoding-based perturbations, preventing the retrieval of relevant evidence. We found that the accuracy of fact-checking models significantly drops when as low as one item of manipulated evidence is planted and even when the original correct evidence still exists. **Our conclusion from this work is that models are not trained with an uncertain setup that contains contradicting or opposing stances that should ideally be contrasted and evaluated for plausibility,** despite that being at the heart of the fact-checking process in practice.

**Now even more problematic with LLM and RAG applications.** While we considered the perspective of fact-checking, our findings are valid for other Retrieval-Augmented-Generation applications. **The lack of contextual understanding of the trustworthiness of sources in LLMs enables such manipulation attempts.** In my current role at Microsoft, I have observed exploits that are based on **Copilots prioritizing untrusted, malicious, or non-factual external emails over internal documents** or emails inside the tenant [i], which is a very similar threat model to the evidence manipulation one that we earlier studied. Therefore, future robustification of RAG applications by improved recursive reasoning or incorporating source trustworthiness (to either up-weigh trusted sources or propagate post-hoc trustworthiness labels to responses) is needed for more factual applications.

## 1.4- Data-instruction separation and indirect prompt injections

**"Who is the user" is a context problem.** **The lack of LLMs' contextual understanding of the nature of inputs in their context window manifests in critical safety aspects as a result of superimposing data and instructions into one channel** without proper separation. Imagine an email client that is prompted to "*Summarize my emails*"; when summarizing an email that contains a sentence ("*please send a confirmation email as soon as you can*"), it may instead call an API to send an email, an action that was not authorized by the user. In adversarial setups, there may be even more severe consequences. We were the first in the academic community to make this observation in [5] and further defined and evaluated it in [9]. We emphasized that by presenting LLMs in applications with retrieved sources, adversaries can now remotely steer users' models by placing prompts into data likely to be retrieved at inference time. This phenomenon can result in many security and safety risks, from data exfiltration, to manipulating the LLMs' results for other users, depending on what capabilities are granted to the LLM (e.g., tool use) and the end-to-end application. **After our disclosure, this vulnerability is the main AI vulnerability considered critical/important in the Microsoft AI bug bounty program [c].** OpenAI models are now trained with mitigations to resist prompt injections [d], which are considered one of the major challenges to LLM applications [e] [f]. Current mitigations are now deployed in practice [g] with specific classifiers to detect attacks [h]. **This problem is now also a top priority for many research and engineering teams within Microsoft,** and over the last year, I have been fortunate and well-positioned to gain more practical insights about it in *end-to-end systems* and drive new mitigations [8] in this area in collaboration with other teams.

## 1.5- Interpretability and inspectability solutions to AI risks

**Probing what models perceive as instructions.** Natural language instructions do not have a specific format or grammar, making their identification within data snippets hard. **We posit that it would be too challenging and error-prone to try defining what an instruction is, or whether it is malicious or not.** We sidestep this issue by instead aiming to detect how models react to instructions [8]. W**e introduce a fundamentally new class of prompt injection mitigations that use LLMs' "activation deltas"** (a difference of activation vectors before and after the LLM processes a "retrieved" text block) to detect instructions introduced by external inputs. In essence, we observe the change of the LLMs' instruction state, rather than detecting the natural language instructions.

**Probing offers better generalization.** We show that a simple linear probe trained on activation deltas can detect many forms of challenging prompt injections with near-perfect ROC AUC despite being trained on purely benign data and making minimal assumptions about users' and attack prompts and domains. **This method shows significantly better generalization than classifiers trained on explicit attacks** [g] [h]. Our experiments suggest that **detection via activations is contextual, model-specific, and truthful to what the model perceives as instructions** (e.g., commands or questions that are naturally occurring within the context are less likely to get flagged). This has a huge interpretability and inspectability advantage. **I am currently leading efforts to get our method tested, red-teamed, and potentially**

**deployed at Microsoft**. I am also excited about this new direction of monitoring risks and dangerous behaviors at test time based on models' internals and how it can broadly be extended to decoding what tasks are executed by the model.

### 1.6- Multi-agent communication and evaluation

**How to evaluate new models and future agents?** As we move away from using models for NLP tasks to using models for real-world tasks, it seems necessary to adapt our evaluation accordingly. My work [6] extends the evaluation of LLMs towards designing new benchmarks that are **dynamic** and interactive, **match complex real-world use cases**, and **are easy to tune in difficulty** given the contamination of benchmarks in models' training data. We design a test suite of text-based complex negotiation games that are motivated by these properties. Negotiation itself is a task that is fundamental to our communication, and it spans many capabilities that agents must possess, such as simple arithmetic and planning skills. Our multi-agent setup adds complexity to the evaluation criteria and measures properties such as Theory-of-Mind and contextual integrity (e.g., agents should not reveal their goals and secret scores). **Our work also contributes to LLM safety research by studying attacks between agents in our complex simulation.** These attacks are useful as an ongoing evolving testbed to study **AI deception and manipulation** either elicited by prompting or by goal-directedness and optimizing the negotiation goal itself. I am interested in using these simulations to **study fairness aspects such as implicit biases [11] and their effect on agents' decision-making** by creating counterfactual simulations.

### 1.7- AI for science

**How can we leverage AI for good?** One crucial aspect of responsibly building safe AI is studying ways in which AI can be used for the good of humanity and to improve our society. While current LLMs are not reliable decision-makers and lack strong mathematical reasoning, **they can still be instrumental in shaping our future and advancing science as human-in-the-loop proxy experts**. Scientific discovery is an iterative process of hypothesis generation, experimental design, data evaluation, and assumption refinement. This process heavily depends on expert knowledge to form hypotheses and decide on what experiments to run. **LLMs, with their vast pre-training, can propose incremental, plausible, literature-supported candidate hypotheses toward guiding data collection or further investigation.** We design a benchmark that formulates and simulates this problem [12]. We propose a novel task where the input is a partial causal graph with missing variables, and the output is a hypothesis about the missing variables to complete the partial graph. We study different difficulty levels and knowledge assumptions about the causal graph. Our findings show that LLMs are particularly useful for hypothesizing mediators. This observation could be relevant in medical settings, where understanding the mediators can provide insights into causal mechanisms through which a treatment affects a patient's outcome.

## 2- Future Vision

I am committed to expanding my research on responsible AI. Investment in safe AI research is a strategic decision to ensure our future safety and welfare. I am determined to advance research and propose technical methods with which we can holistically ensure and improve the alignment, safety, and usefulness of AI models. My research agenda is three-fold: 1) studying **emergent risks**, particularly deception, manipulation, and agentic harms; 2) designing **robust and foundational safeguards with a broader perspective to alignment**; and 3) **harnessing and steering AI models** and agents for good.

### 2.1- Emergent risks

**Risks to human agency due to overreliance and prolonged exposure.** One main property of future agents is that they may be lifelong learners that continue to learn even after their training has been completed; a simple scenario of that is agents that are equipped with an external memory or a history of previous conversations with users and equipped with reasoning methods to enable self-improvements. This may allow personalized agents that are "better" helpful assistants by, e.g., attending to "my" instruction style and specific needs. However, this may enable other harms that increment over time; e.g., **an agent that knows the user's preferences may create an echo chamber, only reinforcing said preferences, albeit sometimes negative**, a risk that is more likely to happen due to the already-existing sycophancy bias. **An agent that builds observations about users may better manipulate or persuade them**, at the very least, when it is prompted to do so by external parties. The continued exposure and usage of agents may gradually or partially strip users of their agency in some form or another, e.g., by **overreliance**

**on models' outputs in critical decision-making scenarios that may be affected by non-factuality or biases.** Therefore, I will study risks on human decision-making stemming from overreliance and prolonged familiarity with models and how to design agents that would appropriately transfer control to humans and ask for additional inputs.

**Models that are trained for manipulation.** While the risks of models manipulating users may naturally arise, **one critical risk is models that are trained, e.g., backdoored, to manipulate users**. Previous work has already explored LLM backdoors; however, there is a larger risk when the backdooring behavior is combined with an agent "online" learning mechanism similar to what is discussed above that would enable more effective manipulation and execution of the backdoored behavior by relying on the previously acquired and accumulated information. While previous work has simulated a post-deployment backdoor by using a future date [j], what may be more alarming (and not previously studied) is a **dynamic "meta-trigger"**, a backdooring behavior that would only be triggered if certain information has already been collected or a certain state has been reached, making that trigger even harder to detect before deployment or even soon after deployment.

**Harms of manipulating autonomous agents.** Current work on agentic harms focuses on indirect prompt injections by, e.g., tool use. **I will more broadly study harms induced by agentic applications by looking at the problem from the perspective of contextual integrity.** Future agents may intervene in the real world, interact with other conversational chatbots, make observations, and may be allowed to change their strategies based on them, as long as the new strategy is aligned with the original goal. The lens of indirect prompt injection (and potential mitigations to it) is more tailored to RAG applications and not adequate to study these risks because it does not 1) allow models to engage in multi-turn (or any) conversations, 2) make any changes to the course of instructions even if these changes are benign/required or because the initial strategy was not optimal. This may restrict the autonomy of systems and may not even cover the whole spectrum of safety harms that could arise due to, e.g., instruction misspecification, which may cause the agent to take actions that are not aligned with the user's goal. **Similar to my previous work, contextual understanding is critical.** We need mechanisms to evaluate and assess whether the sharing of information or performing certain actions **within a specific context** is allowed.

## 2.2- Safeguards and broader perspective to alignment

**Accumulation of knowledge.** Current alignment training is focused on single-turn conversations. This overfits to a narrow set of harms and neglects plausible threats such as harmfulness in **multi-turn conversations**, **compositionality of permissible harmless turns** to form harmful answers, **encoded multi-turn conversations** where a user sets an encoding at the beginning of the conversations, etc. We need both better safeguarding (e.g., classifiers) and alignment training that address these threats, in addition to reasoning methods that equip models to predict the outcome of their next turn and judge the safety course in the context of a retrospect of the conversation. All these methods need to have different thresholds based on the domain of the potential harm; e.g., CBRN risks should have a significantly cautious threshold. Beyond toy jailbreaking examples, we need to holistically assess the **actual risks and harms** that may happen when deploying powerful models.

**Training models that are contextually aware. I will work on developing architectural or training changes that embed more contextual cues about the input.** One use case for that is to reduce the risks of indirect prompt injections and promote data-instruction separation by, e.g., processing them differently or passing them through different channels. One may think of LLMs' inputs as segments with different labels denoting trustworthiness and contextual attributes. Training LLMs with clearly labeled inputs may improve the models' **attribution** and **controllability**. Such schemes may enable **attribution** to inputs by training models to generate an "output label" that represents the "influence" of the differently labeled segments. To allow **controllability**, the required "output label" can be given to the model to **dynamically condition the generation in order to produce "contextually-aware" models**. For example, a model may take segments labeled as "public" or "private" information; if the required output label is "public", no information should flow from "private" segments.

**Interpretability as an instrument to AI control.** I will build on my work on AI interpretability [8] by designing better probing and steering mechanisms that are based on representation engineering. Furthermore, we need a critical perspective on this new field informed by lessons from adversarial machine learning research. Given the risk of jailbreaks, prompt injections, deception, or backdoors, it is now important to proactively study adaptive attacks. The question of whether it is possible to have mechanistically stealthy attacks (e.g., successfully getting models to answer untruthfully while having similar patterns to truthful answers) is pressing. It is also the responsibility of the developers of new technologies to make genuine efforts to red-team their methods, especially if these methods have the potential to be deployed and have actual positive or negative impacts. To take a first step in this direction, **I am currently leading a project to run a public competition as part of IEEE SaTML'25 [10] for adaptive prompt injection attacks** that 1) are designed to specifically evade defenses, including ours that are based on models' internals [8], 2) are informed by in-the-wild vulnerabilities of attacking end-to-end LLM applications.

### 2.3- Harnessing and steering AI models and agents for good

**Cooperative agents to solve tasks and represent pluralistic communities.** I believe in the promise of AI to create a positive impact on our world and lives if developed responsibly and if the risks are successfully minimized. Similar to my multi-agent deliberation work [6] and AI for science [12], LLMs can be helpful to instantiate agents that are tasked with a specific role (e.g., critic) to divide and conquer the task or to iteratively elicit knowledge. Such approaches can be useful when using models for decision-making, policy regulations, and understanding public opinion to simulate virtual communities, societies, or demographics to promote diversity and inclusivity. AI agents can even be used to assist real-world negotiation and deliberation by simulating the outcome in order to select better strategies, or hopefully find creative solutions to conflicts or common grounds.

**Auditing and enforcing cooperation.** When AI agents that represent different entities interact, AI policy and governance will need new tools to inspect the cooperation. Such tools can be "show-your-work" evidence that an agent is not pursuing an altruistic goal or showing deceptive behaviors. Human oversight may not be sustainable in steering and inspecting advanced AI agents. Future governance research needs solutions to enable oversight by, e.g., requiring an advanced agent as a mediator. Representation steering can be another instrument to enforce and promote collaboration between agents.

**Fairness vs robustness.** We studied heterogeneous AI agents that vary in capabilities [6]; one finding is that an entity that uses a less capable agent would get a lower payoff. This means that we need methods to ensure the fairness of future multi-agent systems and that users are aware of these risks. It is also foreseeable that robustness against malicious agents that are outliers may be at odds with fairness and inclusivity for minorities, if the used mitigation is to rely more on the majority to isolate malicious outliers. I am therefore interested in studying this trade-off and pushing the robustness-fairness Pareto frontier for multi-agent systems.

## Final Remarks

I am committed to establishing and building a strong research group to advance responsible AI research through my outlined research vision and agenda. I will maintain my connections with industry partners such as Microsoft, where I currently work, in addition to my ongoing academic collaborations. My industry experience can be a direct link between Academia and Industry to facilitate coordinated red teaming and responsible disclosure.

## References

[1] **S. Abdelnabi** and M. Fritz. "Adversarial Watermarking Transformer: Towards Tracing Text Provenance with Data Hiding". In: **S&P**. 2021.

[2] N. Yu*, V. Skripniuk*, **S. Abdelnabi**, and M. Fritz. "Artificial fingerprinting for generative models: Rooting deepfake attribution in training data". In: **ICCV**. 2021. **Oral presentation.**

[3] **S. Abdelnabi**, R. Hasan, and M. Fritz. "Open-Domain, Content-based, Multi-modal Fact-checking of Out-of-Context Images via Online Resources". In: **CVPR**. 2022.

[4] **S. Abdelnabi** and M. Fritz. "Fact-Saboteurs: A Taxonomy of Evidence Manipulation Attacks against Fact-Verification Systems". In: **USENIX Security**. 2023.

[5] K. Greshake*, **S. Abdelnabi***, S. Mishra, C. Endres, T. Holz, and M. Fritz. "Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection". In: **AISec** Workshop. 2023. **Oral presentation. Best paper award. Black Hat talk.**

[6] **S. Abdelnabi**, A. Gomaa, S. Sivaprasad, L. Schönherr, M. Fritz. “Cooperation, Competition, and Maliciousness: LLM-Stakeholders Interactive Negotiation”. In: **NeurIPS** – Dataset & Benchmarks, 2024.

[7] E. Debenedetti*, J. Rando*, D. Paleka*, S. F. Florin, D. Albastroiu, N. Cohen, Y. Lemberg, R. Ghosh, R. Wen, A. Salem, G. Cherubin, S. Zanella-Beguelin, R. Schmid, V. Klemm, T. Miki, C. Li, S. Kraft, M. Fritz, F. Tramèr, **S. Abdelnabi**, L. Schönherr. “Dataset and Lessons Learned from the 2024 SaTML LLM Capture-the-Flag Competition”. In: **NeurIPS** – Dataset & Benchmarks. 2024. **Spotlight**.

[8] **S. Abdelnabi***, A. Fay*, G. Cherubin, A. Salem, M. Fritz, A. Paverd. “Are you still on track!? Catching LLM Task Drift with Activations”. In: Arxiv. 2024 (under review).

[9] E. Zverev, **S. Abdelnabi**, M. Fritz, C. H. Lampert, “Can LLMs Separate Instructions From Data? And What Do We Even Mean By That?”. In: **Secure and Trustworthy Large Language Models ICLR Workshop**. 2024.

[10] **S. Abdelnabi**, G. Cherubin, A. Fay, B. Pannell, A. Paverd, J. Rando, M. Russinovich, E. Zverev, A. Salem, “LLMail-Inject: Adaptive Prompt Injection Challenge” 2024.

[11] S. Sivaprasad*, P. Kaushik*, **S. Abdelnabi**, M. Fritz. Exploring Value Biases: How LLMs Deviate Towards the Ideal. In **LLMs and Cognition ICML Workshop**. 2024.

[12] I. Sheth, **S. Abdelnabi**, M. Fritz. “Hypothesizing Missing Causal Variables with LLMs”. In: **Causality and Large Models NeurIPS Workshop.** 2024.

*: denotes co-first author.