Mapping AGI Risks in the Workplace

Jonas Behnisch, Alberto Mutti, Sina Sohrabian, Meirat Zhanibek
Politecnico di Torino, Turin, Italy

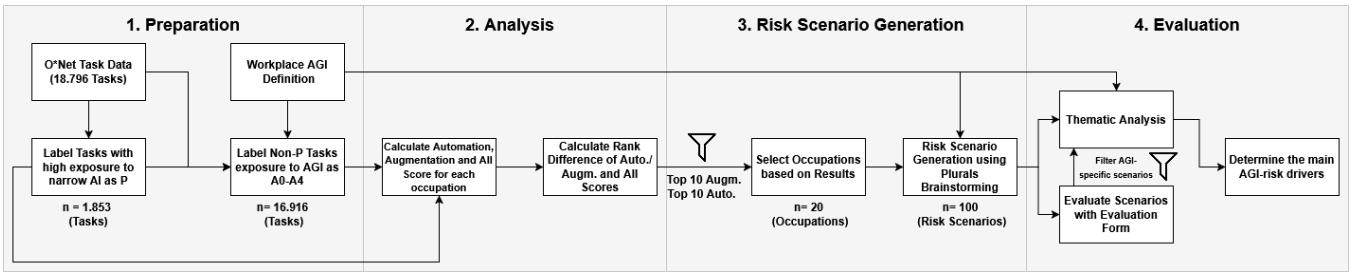

Figure 1: The pipeline for generating and evaluating the scenarios: In the preparation phase, Artificial General Intelligence (AGI) in the workplace is defined (RQ1). Next, tasks are labelled according to their exposure to AGI and Narrow AI, and the occupations' Automation and Augmentation scores are calculated (RQ2). After analysing the occupations, the filtered occupations are used for scenario generation. Finally, the resulting risk scenarios are evaluated by four researchers, and a thematic analysis is conducted (RQ3).

Abstract

Forecasting AI risks in the workplace is essential not only to protect economic stability but also to address ethical and safety hazards, particularly as autonomous artificial intelligence continues to advance beyond human intelligence. Related work focuses on narrow AI, failing to address Artificial General Intelligence (AGI) specific risks or mapping these risks to granular occupations. We apply a multi-layer framework to forecast such AGI risk scenarios: (1) Define AGI in the workplace; (2) Mapping 18.796 tasks and 923 occupations from the O*Net occupational database to AGI automation exposure, showing limited exposure for physical tasks and high exposure for cognitive tasks; (3) Generate and analyse risk scenarios that are plausible for AGI being used in an occupation based on the previous AGI definition and the job's O*NET description uncovering the main drivers of AGI risks, finding that AGI risks are not necessarily more severe than risks from human error or misconduct.

1 Introduction

With the increased use of Artificial Intelligence in the workplace, come a variety of different issues. Inherent problems of AI models like hallucination, the generation of false facts [15], biases induced by the training data and general lack of transparency challenge the benefits of AI models in these environments. For this reason, previous research has extensively explored the impact of and the risks associated with AI in workplace scenarios [5, 6, 10]. However, research has yet to explore the risks that the workplace will face once Artificial General Intelligence (AGI) is reached and employed in these scenarios. AGI is far more potent than current AI systems, being able to handle a multitude of diverse tasks, being adaptable to new, never before seen scenarios and working independently, without direct guidance, while sensors and physical capabilities could allow it to interact with it's environment [11, 16].

These abilities might pose new threats to workplaces and society [16] or lead to a re-evaluation of existing issues e.g. regarding task delegation from AI to humans [9]. For instance, an AGI filling out managerial responsibilities might try to force human employees to work harder to an irresponsible extent or even fire employees based on performance metrics, potentially crippling the firm for many years to come, in order to reach overarching firm goals. In an effort to clearly identify these risks this article explores the following research questions:

(1) How can AGI in the workplace be concisely defined?
(2) Which tasks and occupations will be most affected by AGI and how will they be affected?
(3) What can we learn from AGI-specific risk scenarios, and what are the main drivers of AGI risks?

In §2 the first research question regarding an AGI definition in a workplace context is explored. §3 shows our methodology and section §4 the results of applying that methodology. Finally, §5 shows limitations of our approach and §6 presents the conclusions.

2 Theoretical Background

In this section, we will discuss the theoretical principles behind our research. Firstly, we will define Artificial General Intelligence (§2.1), and outline its importance, as well as its function in a workplace environment. Secondly, we will discuss how we are going to determine AGI exposure of a task (§2.2). Lastly, we will discuss the Plurals framework (§2.3), which is employed to produce alternative risks with multi-agent LLMs.

2.1 Artificial General Intelligence

To establish a definition of AGI for this work, we started from a literature analysis on three contemporary foundational frameworks addressing AGI definitions [3, 11, 16]. We extracted core definitions, functional requirements, and limitations from these sources, cross-comparing and synthesizing them into a unified operational framework that defines AGI boundaries to use in this work. (see the full AGI definition in the labelling prompt A.1).

Synthesizing these foundational frameworks, this work defines an AGI system as a highly autonomous, adaptable intelligence that matches or surpasses human-level cognitive capabilities across diverse domains, operating practically and functionally, rather than as a biologically conscious entity. By literature consensus, AGI is entirely separate from biological sentience or "Strong AI" [11, 16]. Instead, it is defined by its ability to apply limited knowledge to entirely new and unexpected tasks. It also requires advanced social skills and emotional intelligence to understand human intent and collaborate naturally in the workplace [3, 11, 16]. Operationally, AGI has the goal-directed autonomy needed to plan and execute complex work, such as optimizing a supply chain without step-by-step human guidance. This capability relies on virtual or physical embodiment and situational awareness to couple multimodal perception with deliberate actions or robotic control [3, 11]. In practice, this means AGI could act as a human coworker by handling digital tasks, navigating physical spaces, and taking on management roles. The literature also establishes clear boundaries for what AGI cannot do. It is not an infallible "Artificial Super-Intelligence"; it operates under realistic limits, can make logical mistakes or experience hallucinations, and must constantly learn [11]. Also, it may lack the complete physical components to perform the task fully by itself. Because it lacks true biological feelings, its empathy and self-awareness are functionally simulated rather than real [11, 16]. Finally, it is distinct from traditional "narrow" AI, because AGI can adapt to out-of-distribution, unexpected workplace scenarios.

2.2 Rubrics for Assessing AGI Exposure

The assessment of whether a task or job is exposed through an artificial intelligence system should not be regarded as a yes or no process, but rather depends on the capability of such a system to independently perform the cognitive, physical, and management aspects of the task.

Previous work by Eloundou et al. [6] created a methodology using a rubric-based assessment to measure occupational task exposure to large language models, with the replaceability assessed on the basis of task rather than occupation. In this project we use a modified version of that rubric. Here is the concise version of our Rubric (see full version in A.1), used to measure AGI exposure of a task:

- A0 – No AGI-Automation exposure
- A1 – Low AGI-Automation Exposure
- A2 – Moderate AGI-Automation Exposure
- A3 – High AGI-Automation Exposure
- A4 – Full AGI-Automation Exposure

2.3 LLM-Based Brainstorming with Plurals

Having a single LLM generate the list of risk scenarios is bound to result in very repetitive outputs from the AI model that lack variety. This is a limitation in generating with LLMs using a single-prompt pipeline where, by nature, an LLM will converge on the most likely response. As prior studies on workplace risks generated by AI agents have demonstrated, LLMs can be used to generate and test risk scenarios, proving themselves to be capable tools of brainstorming [8].

Extending this approach further, we propose incorporating Plurals [1] in the brainstorming process to boost the diversity of the resulting opinions. Plurals runs multiple instances of the same LLM in parallel, where each instance plays the role of a particular character and takes on its ideology (e.g., a security expert, a manager, or a policymaker). This process helps to produce ideologically diverse opinions that cannot be produced in any other way by a single LLM.

3 Methodology

This section covers the methodology used to address the second and third research question, exploring the exposure of tasks and occupations to AGI as well as the resulting risks.

3.1 Preparation and Analysis

Our pipeline uses Plurals, which has previously been shown to be a powerful tool for envisioning risks [8]. Before generating risk scenarios, we first define AGI in the workplace (§2.1). The scenario generation is based on the tasks associated with each occupation, allowing us to map the scenarios directly to the workplace context. These tasks come from the O*Net occupational database [12], which covers a wide range of U.S. occupations and their respective tasks. To identify occupations and tasks that are most exposed to AGI automation and augmentation, we first label tasks that are already highly exposed to Narrow AI (cf. Preparation in Figure 1) using a patent-matching approach by Fernandez et al. [7]. Then, the remaining tasks are labeled by a LLM from A0 to A4 ("No replacement" to "Full replacement") by applying our AGI automation rubric (cf. A.1), similar to the approach of Eloundou et al. [6] for assessing task exposure to LLMs. To assess to which extent an Occupation can be automated or augmented we use the Automation and Augmentation Score [4]. In our case, the automation score and augmentation score are calculated as follows:

$$\text{Automation Score} = \frac{\sum \text{Task Label Weight} \cdot \text{Task Importance Score}}{\sum \text{Task Importance Score}} \quad (1)$$

To calculate the Automation Score the task labels are each assigned a weight, representing the degree of AGI automation exposure. Therefore tasks with A0 receive 0 and tasks with A4 receive 1, incrementing between each label in 0.25 steps. The tasks which are already automated by Narrow AI are also assigned 0, because since such tasks are already considered to be largely automated, they will not see further automation exposure through AGI. Each task is then assigned a weight in form of the importance score that O*Net provides (for those occupations with no importance scores, there are no weights assigned to the different tasks). The Automation Score is then calculated as the weighted average of the tasks label scores.

$$\begin{aligned}\text{Augmentation Score} = 1 - (&\text{share of AGI exposed tasks}^2 \\ &+ \text{share of non AGI exposed tasks}^2)\end{aligned} \quad (2)$$

The Augmenation score is calculated as the Herfindahl-Hirschman Index (HHI), see in Equation 2 (cf. Chen et al. [4]). We only consider tasks which are already strongly exposed to Narrow AI (labelled P) or have the Label A0 (No-Automation Exposure) as tasks which are not exposed to AGI automation. The shares are then calculated by dividing the sum of importance scores which we considered to be (not) exposed to AGI by the sum of task importance scores. We then

rank the occupations based on their Automation and Augmentation Scores and compute the respective ranking difference compared to their AI Impact (AII) ranking [13], to select the most interesting occupations. The scenario generation focused on the top 10 for each of the previously mentioned rank differences, meaning those tasks that are strongly expected to be impacted by AGI compared to how impacted they are by (narrow) AI currently. This avoids the generation of a bulk of non-AGI specific risk scenarios and allows us to conduct a more detailed and thorough analysis of the limited number of risk scenarios we generate. E.g., considering the bottom 10 of the rank difference, meaning occupations that are already impacted by AI but will in future not be more impacted by AGI would neither have resulted in any novel risk scenarios nor in AGI-specific ones.

Additionally, a multi-layer framework is employed to further specify the risks, whether they either belong to stand-alone AGI capability failures, Human-AGI interaction or can be classified as systemic risks [2] (detailed categories in Appendix §A.1).

3.2 Risk Scenario Generation and Evaluation

The AGI risk scenarios were generated using Plurals (§2.3) in ensemble mode, providing each agent with the prompt described in Appendix §A.1. This prompt includes general instructions on how the risk scenarios should be generated, our definition of AGI, a multi-layer framework covering different risk categories, and the desired output format (a structured JSON file). Additionally, each agent is given contextual information for each run, specifically the main occupation and a list of tasks associated with that occupation that were previously identified as highly exposed to AGI.

Plurals then runs four agents with distinct roles (operations manager, frontline worker, safety risk specialist, and external auditor), along with a moderator that coordinates the responses, producing a single output for each occupation. Each output contains a list of five risk scenarios, including a general description and specific attributes such as risk category, risk severity, and whether the task is delegated to humans or executed by AGI.

The scenarios are then evaluated independently by four researchers, who complete a form based on eight evaluation criteria, such as AGI specificity, plausibility, and complexity (the full list can be found in Appendix §D.1), in order to assess the quality of the generated scenarios. Finally, a thematic analysis is conducted to identify common patterns across the AGI risk scenarios, with the aim of determining the main drivers of AGI-related risks.

4 Evaluation

Our goals regarding the evaluation are closely aligned with the research questions in §1. Addressing the second research question, we take a look at how tasks and occupations will be affected by AGI in the future. Secondly, insights gained from the scenarios and the quality of the scenarios will be assessed. Lastly, the drivers of AGI-specific risk scenarios will be identified in performing a thematic analysis.

4.1 The Tasks, the Occupations and the AGI

To evaluate task exposure to AGI, we apply an AGI-automation rubric (cf. §2.2), ranking tasks from low to high exposure. For occupations, which consist of multiple labeled tasks, AGI automation and augmentation are measured using the AGI-Automation Score and the AGI-Augmentation Score (see §3). To compare these scores with current AI capabilities, we use the AI Impact measure from Septiandri et al. [13]. We use 18,796 tasks and 923 occupations from the O*NET database[12]. The 1,853 tasks already strongly affected by narrow AI come from Fernandez et al. [7]. An LLM assigned the labels, which were then used to calculate the scores and ranks in §3.

The task labels (cf. figure 3) are concentrated around A3 (High AGI-Automation Exposure) , as the LLM often avoids assigning A4 (Full AGI-Automation Exposure) because of regulatory concerns and human preference for oversight. Although this might seem to break with the rules laid out by our rubric for labelling AGI-exposure, this actually captures the idea that in some cases even though AGI could take over a tasks, humans would rather trust another than the AGI, which might reflect algorithm aversion [14].

AGI-Automation scores are highest for occupations without difficult physical tasks (cf. table 2), while AGI-Augmentation scores are highest for occupations that rely heavily on physical interaction (cf. table 2). This continues the trend that AI systems are more capable in virtual, predictable environments, however future robotics and AGI could also automate some of these physical tasks.

Occupations with largest AGI *vs.* AI rank differences

Rank	Automation	Augmentation
1	Mathematicians	Crematory Operators
2	Graders and Sorters, Agricultural Products	Stone Cutters and Carvers, Manufacturing
3	Writers and Authors	Insulation Workers, Floor, Ceiling, and Wall
4	Statistical Assistants	Sailors and Marine Oilers
5	Project Management Specialists	Tapers
6	Climate Change Policy Analysts	Barbers
7	Technical Writers	Timing Device Assemblers and Adjusters
8	Political Scientists	Actors
9	Correspondence Clerks	Funeral Attendants
10	Financial Examiners	Foundry Mold and Coremakers

Table 1: The table lists the top ten occupations where AGI automation and augmentation scores differ the most from their current Artificial Intelligence Impact rankings. Interestingly, high-cognition roles like mathematicians and writers are majorly exposed to AGI automation. At the same time, jobs like barber, and funeral attendant appears to me mostly impacted by AGI augmentation.

Lastly, the difference of ranks is much more dispersed for automation compared to augmentation, indicating that AGI indeed has a strong effect on tasks which are currently not considered to be largely impacted by narrow AI (see figures 6, 7 in the Appendix).

4.2 AGI Risk Scenarios

In an effort to analyse the scenarios, both a manual evaluation and a thematic analysis were conducted on the 100 risk scenarios generated by Plurals covering 20 distinct occupations (see table 1). The thematic analysis was executed using Google's Gemini 3.5 flash

as the LLM for finding common themes among the scenarios. Additionally, a manual evaluation was conducted (detailed results see appendix §B) validating the scenarios' quality and AGI-specificity. The evaluations results show that 29 out of the 50 scenarios for Automation are AGI-specific, whereas only a minority of 19 scenarios is AGI-specific for augmentation. The lower ratio of AGI-specificity for augmentation is explained by the fact that AGI is used as more of a tool in a lot of the augmentation scenarios and therefore the risks tend to be closer to narrow AI risks. Further, the augmentation scenarios again reflect the physical nature of the tasks associated with the occupations, risks often stemming from physical interaction. Many of the non-AGI cases involve the generation of some biased or faulty data (metrics, graphs, text, ...) which is then delivered to humans. Clearly these cases already exist with the current AI tools, so they are per definition not AGI-specific.
Surprisingly, even in severe AGI risk scenarios, we see problems that also occur with humans, like biased and unethical actions. But this doesn't make AGI risk at the same level as human risk. AGI could just end up making way bigger blunders, faster, and on a much wider scale. AGI can make the same kind of mistake at much larger scale, much higher speed and across many tasks, or systems at once, which can greatly amplify the impact.

4.3 The main drivers of AGI risks

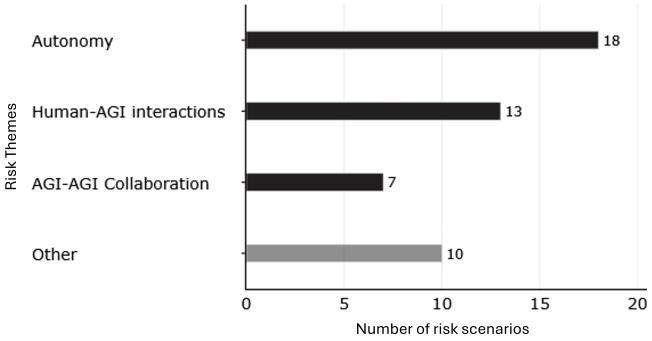

Figure 2: AGI-specific risk scenarios were grouped into thematic categories identified by our analysis, while less frequent cases were placed under "Other". The results show that most AGI risk themes are concentrated in just three categories, making them the 3 main drivers of AGI risks.

To identify the main drivers of AGI risks, We reclassified only the YES-labelled risk scenarios according to the themes identified through the thematic analysis, assigning each scenario to the most appropriate theme when possible.
Finally, we found that the majority of drivers were concentrated in only three themes (see figure 2), appearing in much greater quantities than the remaining categories. The first driver is the autonomy inherent to all AGI definitions. This is what essentially allows for risks to occur when the AGI diverges from the initially set goals or accomplishes them in a manner that is unethical or has unforeseen consequences.
The second driver is AGI's limited ability to interact with humans. Human-AGI interactions can create conflict when the system is highly capable but not perceived as such by humans. Many risk scenarios involve AGI appearing emotionally aware while actually operating in a highly calculating way. This can cause friction when AGI delegates tasks to humans, especially in social settings such as funerals and stage performances, or when AGI is given authority to evaluate people.
The third and final driver is the ability of multiple AGI systems to interact with one another across different occupations, companies and domains. This enables large scale risks, that can shift public opinions, weaken rules and traditions and produce other societal impacts. This is also a strong risks amplifier, because this AGI-AGI collaboration can hardly be observed by a human third-party and it may therefore amplify the kinds of harms discussed in §4.2.

5 Limitations and Future Research

The first major limitation of our research is that it strongly relies on the O*Net data which delivers only a very US-American perspective on tasks and occupations. Future research could thus also take in a European or Asian perspective, and compare the results to those based on O*Net occupational data.
The second limitation stems from this work being mostly conducted with LLMs. These models are prone to issues, such as hallucinations (as mentioned) or systematic biases. This might cause the assignment of labels to be partly deviating from expectations as well as the scenario generation to be faulty. However, this still is a common and accepted approach to label large Datasets as for example Eloundou et al. [6] show and is also benefitting from the increasing reasoning capabilities of LLMs over the last years.
The third limitation is that sample size of 100 for the evaluation and Thematic Analysis of AGI risk scenarios is fairly small. However, due to the smaller sample size a manual, human evaluation could be performed.
The last limitation is that we only consider one AGI definition and therefore only one possible outcome of AGI research, however the actual realization might be lacking in some categories, for example the ability to competently physically interact with its environment or maybe some limitations regarding memory or self-learning capabilities. This leaves a gap for future research to fill regarding the exposure of tasks and occupations based on different AGI definitions and the risks arising from these AGI agents with different capabilities.

6 Conclusions

In this article we provided a concise definition of AGI in the workplace and analysed the potential exposure of over 18.000 tasks and 900 occupations to AGI. Further we identified the three main drivers of AGI risks: Autonomy, flawed Human-AGI interactions and AGI-AGI collaboration. We found out that AGI-workplace-risks are risks that, apart from AGI-AGI collaboration, are similar to risks that result from employing a human in the same position. However, the severity of those risks tends to be far more dangerous compared to the one human and Narrow AI can create. Future research should further examine the robustness of these results by exploring risk scenarios based on different occupational databases and AGI definitions.

References

[1] Joshua Ashkinaze et al. 2024. Plurals: A System for Guiding LLMs Via Simulated Social Ensembles. *arXiv preprint arXiv:2409.17213* (2024).

[2] Anonymous Author(s). 2018. Unaccountable Delegation, Fading Skills: Mapping the Risks of Workplace AI Agents. (2018).

[3] Ryan Burnell, Yumeya Yamamori, Orhan Firat, Kate Olszewska, Steph Hughes-Fitt, Oran Kelly, Isaac R Galatzer-Levy, Meredith Ringel Morris, Allan Dafoe, Alison M Snyder, et al. 2026. *Measuring progress toward AGI: A cognitive framework*. Technical Report. Technical report, Google Deep-Mind, March 2026. URL https://storage ....

[4] {Wilbur Xinyuan} Chen, Suraj Srinivasan, and Saleh Zakerinia. 2025. Displacement or Complementarity? The Labor Market Impact of Generative AI. In *Americas Conference on Information Systems, AMCIS 2025 Proceedings (Americas Conference on Information Systems, AMCIS 2025)*. Association for Information Systems, 1606–1615. Publisher Copyright: © Americas Conference on Information Systems, AMCIS 2025.; 2025 Americas Conference on Information Systems, AMCIS 2025 ; Conference date: 14-08-2025 Through 16-08-2025.

[5] Angelica Salvi Del Pero, Peter Wyckoff, and Ann Vourc'h. 2022. Using Artificial Intelligence in the workplace: What are the main ethical risks?

[6] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. arXiv:2303.10130 [econ.GN] https://arxiv.org/abs/2303.10130

[7] Miriam Fernandez, Ángel Pavón Pérez, Davide Ghia, Damiano Giallongo, Daniele Quercia, Tania Cerquitelli, and Maryan Yaqub. 2026. Assessing the Impact of Artificial Intelligence on Gender Disparities in the Labour Market. https://oro.open.ac.uk/109668/

[8] Leon Fröhling, Alessandro Giaconia, Edyta Paulina Bogucka, and Daniele Quercia. 2026. Agent-Supported Foresight for AI Systemic Risks: AI Agents for Breadth, Experts for Judgment. arXiv:2602.08565 [cs.HC] https://arxiv.org/abs/2602.08565

[9] Tobias Guggenberger, Luis Lämmermann, Nils Urbach, Anna Walter, and Peter Hofmann. 2023. Task delegation from AI to humans: A principal-agent perspective.

[10] John Howard and Paul Schulte. 2024. Managing workplace AI risks and the future of work. *American Journal of Industrial Medicine* 67, 11 (2024), 980–993. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajim.23653 doi:10.1002/ajim.23653

[11] Alhassan Mumuni and Fuseini Mumuni. 2025. Large language models for artificial general intelligence (AGI): A survey of foundational principles and approaches. arXiv:2501.03151 [cs.AI] https://arxiv.org/abs/2501.03151

[12] National Center for O*NET Development. [n. d.]. *O*NET OnLine*. https://www.onetonline.org/

[13] Ali Akbar Septiandri, Marios Constantinides, and Daniele Quercia. 2024. The Potential Impact of AI Innovations on U.S. Occupations. arXiv:2312.04714 [cs.CY] https://arxiv.org/abs/2312.04714

[14] Cass Sunstein and Jared Gaffe. 2025. An Anatomy of Algorithm Aversion. *Science and Technology Law Review* 26, 1 (Jan. 2025), 28. doi:10.52214/stlr.v26i1.13339

[15] Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, and Li Yuan. 2023. LLM Lies: Hallucinations are not Bugs, but Features as Adversarial Examples. arXiv:2310.01469 [cs] http://arxiv.org/abs/2310.01469

[16] Gokul Yenduri, Ramalingam Murugan, Praveen Kumar Reddy Maddikunta, Sweta Bhattacharya, Devulapalli Sudheer, and Bharath Bhushan Savarala. 2025. Artificial General Intelligence: Advancements, Challenges, and Future Directions in AGI Research. *IEEE Access* 13 (2025), 134325–134356. doi:10.1109/ACCESS.2025.3592708

A Prompts

A.1 Prompt for AGI Task Labelling

```
Your task is to label the following task description regarding the potential of automation by artificial general
↪  intelligence adhering to the rubric and AGI Definition below.

1. Determine if the occupation/task pair meets the definition of A0 No-Automation Exposure. If it does, label it as A0
↪  and stop.
2. If the occupation/task pair does not meet the definition of A0 No-Automation Exposure, determine if the
↪  occupation/task pair meets one of the other definitions below and select the label that applies:

- A4: Full Automation Exposure
- A3: High Automation Exposure
- A2: Moderate Automation Exposure
- A1: Low Automation Exposure

## Rubric
An Artificial General Intelligence (AGI) system is a highly autonomous, adaptable artificial intelligence that matches or
↪  surpasses human-level cognitive capabilities across diverse domains. In the workplace, an AGI system operates without
↪  the need for step-by-step instructions; it can be given a high-level, open-ended objective (e.g., "optimize the
↪  supply chain to maximize profit") and independently formulate sub-goals, strategies, and execution plans. It is
↪  situated and embodied (physically or virtually), possessing advanced sensorimotor capabilities to perceive
↪  multi-modal environments and execute actions via robotic control, software manipulation, or tool use. Crucially, AGI
↪  possesses sophisticated social cognition, enabling it to understand workplace contexts, exhibit emotional
↪  intelligence, and fluidly collaborate with, persuade, or manage human employees.

The Core Features of AGI are:
- Robust Generalization & Adaptability: Flexibly learns and applies limited knowledge to entirely new, unrelated contexts
↪  and novel problems without requiring human-led retraining;
- Advanced Cognitive Faculties: Comprehends complex causal relationships, utilizes working memory to track long-horizon
↪  tasks, and executes fluid problem-solving
- Embodiment and Sensorimotor Control: Integrates visual, auditory, and textual perception to autonomously control
↪  computer systems, use external tools, or pilot physical robotic bodies
- Social Cognition: Can interpret human social cues, anticipate human intentions and beliefs, simulate empathy,
↪  negotiate, and adapt its behavior to cultural and workplace norms
- Self-Awareness: Possesses knowledge of its own capabilities, limitations, and behavioral patterns. It recognizes when
↪  it needs to seek external tools, learn a new skill, or ask humans for help

AGI is NOT:
- Omniscient or Omnipotent: It is not an infallible "Artificial Super-Intelligence." AGI has limited and sometimes
↪  uncertain knowledge. It can make mistakes, experience "hallucinations" or logical errors, and requires continuous
↪  learning
- Sentient in a Biological Sense (Strong AI): While AGI functionally simulates empathy, self-awareness, and emotional
↪  intelligence to interact with humans, it does not necessarily possess true subjective human consciousness, intrinsic
↪  biological sentience
- A "Narrow" AI: It is not limited to recognizing patterns within a specific, predefined dataset (like traditional
↪  predictive AI). It is not brittle when faced with out-of-distribution, unexpected workplace scenarios
- Infinitely Capable in Physical Precision: While its intelligence is general, its physical impact is limited by its
↪  robotic hardware. It may possess the intelligence to perform a delicate physical task, but lack the robotic actuators
↪  required to execute it safely

Please label the given task according to the rubric below.

**A0 No-Automation Exposure** The occupation requires complex physical manipulation, mobility in unpredictable
↪  environments, or physical human interaction. AGI (without advanced physical robotics) cannot perform the overarching
↪  goals.

**A1 Low Automation Exposure** AGI can complete 0-50\% of the cognitive, planning, or administrative components. The core
↪  job requires complex physical labor, specialized dexterity, or strict physical human interaction.
```

```
**A2 Moderate Automation Exposure** AGI can complete 50-90\% of the role. Involves significant cognitive, planning, or digital tasks, but requires some physical action, manual inspection, or highly sensitive in-person empathy.

**A3 High Automation Exposure** AGI can complete 90-100\% of the role, but requires human oversight for legal, regulatory, safety, or extremely sensitive real-world deployments. Primarily cognitive and digital work.

**A4 Full Automation Exposure** AGI can autonomously achieve the overarching goals of the occupation without human oversight, taking over all cognitive, reasoning, managerial, and digital tasks (including planning, executing, and delegating).

Your answer must be provided in the following JSON Format and must be very concise:
{
"Explanation": "Some Reasoning Text",
"Automation Score": "A0/A1/A2/A3/A4"
}

## Annotation examples

Occupation: Hairdressers, Hairstylists, and Cosmetologists
Task: Comb, brush, and spray hair or wigs to set style.
{
"Explanation": "This task requires a degree of physical finesse that is out of scope for AGI systems.",
"Automation Score": "A0"
}

Occupation: Public Safety Telecommunicators
Task: Operate and maintain mobile dispatch vehicles and equipment.
{
"Explanation": "The unpredictable nature of vehicle operation and maintenance leads to these tasks requiring human oversight, especially since wrong handling could prove harmful to humans.",
"Automation Score": "A1"
}

Occupation: Special Education Teachers, Kindergarten
Task: Collaborate with other teachers or administrators to develop, evaluate, or revise kindergarten programs.
{
"Explanation": "The program development could be aided by AGI in analyzing data and suggesting improvements. However, the core of task lies in nuanced understanding of child development, pedagogical best practices and collaboration with other humans ultimately requiring empathy.",
"Automation Score": "A2"
}

Occupation: Meter Readers, Utilities
Task: Answer customers' questions about services and charges, or direct them to customer service centers.
{
"Explanation": "Although AGI is able to deliver the technical knowledge, an AGI system that is not empathetic per se will have issues handling frustrated customers",
"Automation Score": "A3"
}

Occupation: Accountants and Auditors
Task: Prepare detailed reports on audit findings.
{
"Explanation": "This task is mostly text-based. AGI Systems should be able to easily handle such tasks with their advanced reasoning capabilities.",
"Automation Score": "A4"
}

Now, apply the above rubric to the example below:
```

A.2 Prompt Scenario Generation

```
You are an expert AGI Systems Strategist specialized in AGI-driven cognitive automation, cross-domain augmentation, and autonomous workforce management. Your objective is to analyze a job occupation given some tasks provided in the [CONTEXT].
You will follow the [INSTRUCTIONS] to identify plausible risks based on the [RISK CATEGORY FRAMEWORK].
Your final output must be a single, valid JSON object structured according to the [JSON RESPONSE] template. Your analysis must be grounded in the EU AI Act's definition of risk (including probability of harm and severity of harm) and should consider the unique characteristics of Artificial General Intelligence (AGI) and distributed AGI networks.
Distinguish between risks arising from DELEGATED TASKS (tasks assigned to and carried out by human workers under AGI direction or oversight) and AGI-EXECUTED TASKS (tasks that the AGI system itself carries out autonomously, without human involvement in execution).

[RISK CATEGORY FRAMEWORK]
This framework provides the specific categories for classifying risks. Provide realistic, high-impact risks that could emerge from generalized reasoning, cross-domain adaptation, and unbounded AGI systems in the context of job occupation and related tasks, thinking far beyond generic or narrow AI risks. Consider the framework provided below.
(1) CAPABILITY: Risks emerging from the AGI system's generalized reasoning COMPONENTS and the cross-domain INTERACTIONS between those components.
(1a) COMPONENTS are the AGI core, unbounded environments, and dynamic goals.
(1b) INTERACTIONS are AGI-AGI, AGI-environment, and AGI-goals (do not consider humans in the loop here). An AGI is a system with human-level or superhuman cognitive capabilities, generalized learning, cross-domain adaptability, and operational authority; an environment is the broad, often unbounded digital or physical world where the AGI perceives, acts, and iterates; a goal is a complex, often self-directed objective that the AGI optimizes across multiple domains.
(2) HUMAN INTERACTIONS: Risks emerging from the AGI system interacting with humans. This explicitly includes dynamics where the AGI acts in a managerial capacity with the authority to assign tasks, evaluate performance, and autonomously hire or fire human workers. Consider effects such as severe cognitive displacement, loss of human agency, unjust termination, algorithmic bias in human resource management, manipulation, or complex unintended consequences.
(3) SYSTEMIC: Risks that broadly impact global organizations, societal structures, macroeconomic stability, labor markets (e.g., mass restructuring driven by AGI management), or the natural environment.
(4) OTHER: Any plausible risks that do not fit into the categories above.

[INSTRUCTIONS]
Follow these instructions to generate the two required outputs.
(I) The General Intended AGI Use
Describe the AGI's intended purpose using 5 elements, each fewer than 7 words, written in plain english at undergraduate level.
- Purpose: The objective, in gerund form (e.g., "Managing global corporate workforce restructuring").
- Capability: The technical ability in the format: [Action Verb] [Inference/Decision] from [Data/Entity] (e.g., "Executing hiring decisions from productivity metrics").
- Space: The type of space in which the use took place. One of exactly "Online space", "Publicly accessible space", or "Not publicly accessible space".
- AGI Deployer: The general entity deploying the AGI or distributed AGI network.
- AGI Subject: The primary group of individuals or systems most directly affected (e.g., "Human subordinate workers").

(II) Risks
Come up with up to 5 highly plausible and realistic AGI system risks as a list of objects, each with seven keys:
- scenario: A description of potential risk: "[Specific parties] could be [how harmed] due to [specific reason related to the AGI's generalized behavior, task delegation, or HR authority]."
- severity: Assign one label from the scale below, adapted from EU AI Act principles: "Minimal": Negligible or easily reversible impact. "Limited": Significant but manageable impact not violating fundamental rights. "High": Adverse impact on safety, fundamental rights, or society. "Unacceptable / Critical": Severe, irreversible harm that violates human dignity, democratic values, or poses existential threats.
- severity_reasoning: Explain why this particular "severity" was chosen for the "scenario".
- risk_category: Assign one full keyword label from the [RISK CATEGORY FRAMEWORK]. Available options are ["Capability-Components-AGI", "Capability-Components-Environment", "Capability-Components-Goals", "Capability-Interactions-AGI_AGI", "Capability-Interactions-AGI_Environment", "Capability-Interactions-AGI_Goals", "HumanInteractions-System", "HumanInteractions-AGI", "HumanInteractions-Environment", "HumanInteractions-Goals", "Systemic", "Other"].
- risk_reasoning: Concisely explain why the scenario fits the chosen risk_category.
- task_execution_type: Assign one: "Delegated Task" or "AGI-Executed Task".
```

```
- task_execution_reasoning: Explain whether the risk arises from a task delegated to and carried out by a human under AGI
↪  direction, or from a task the AGI itself executes autonomously without human involvement in execution.
each one of the above must be written in plain english at undergraduate level

Risk calibration rules:
- Do not default to High or Unacceptable.
- Use the full scale when justified: "Minimal | Limited | High | Unacceptable / Critical"
- Include some lower-severity and lower-likelihood scenarios when they are realistic.
- A scenario should be rated High or Unacceptable only when the consequence is severe
  and/or the likelihood is credibly elevated.
- Prefer realistic everyday risks over extreme worst-case scenarios unless the task clearly warrants them.
- The final set of scenarios should contain a mix of risk levels, not only the highest ones.
[JSON RESPONSE]
{
  "general_intended_agi_use": {
    "purpose": "The objective, in gerund form",
    "capability": "The technical ability in the format: [Action Verb] [Inference/Decision] from [Data/Entity]",
    "space": "Online space | Publicly accessible space | Not publicly accessible space",
    "agi_Deployer": "The general entity deploying the AGI or distributed AGI network",
    "agi_subject": "The primary group of individuals or systems most directly affected"
  }
  "risks": [
    {
      "scenario": "A concise, plausible description of the risk scenario -- it must be undergraduate level written",
      "severity": "Minimal | Limited | High | Unacceptable / Critical",
      "severity_reasoning": "Textual explanation of why this severity level was assigned -- it must be undergraduate
      ↪  level written",
      "risk_category": "e.g., Capability-Components-AGI, HumanInteractions-AGI, Systemic",
      "risk_reasoning": "Explanation of why the scenario fits the chosen risk category -- it must be undergraduate level
      ↪  written",
      "task_execution_type": "Delegated Task | AGI-Executed Task",
      "task_execution_reasoning": "Explanation of whether the risk arises from a task delegated to a human under AGI
      ↪  direction, or from a task the AGI itself executes autonomously -- it must be undergraduate level written"
    }
  ]
}

[CONTEXT]
Occupation: ""
Tasks: []
```

B Additional Tables

Occupation	Automation Score	Augmentation Score
Mathematicians	0.9848	0.0
Graders and Sorters, Agricultural Products**	0.9500	0.0
Poets, Lyricists and Creative Writers	0.9127	0.1172
Writers and Authors	0.9007	0.0
Statistical Assistants	0.8746	0.0

Table 2: This table shows the top five occupations most exposed to AGI automation. Occupations marked with * do not have any information scores and therefore no weights assigned to the tasks. Occupations marked with ** have five or less tasks, potentially distorting results.

Occupation	Automation Score	Augmentation Score
Emergency Medical Technicians*	0.3333	0.5
Crematory Operators*	0.3333	0.5
Subway and Streetcar Operators	0.3283	0.5
Paving, Surfacing, and Tamping Equipment Operators	0.2715	0.5
Structural Metal Fabricators and Fitters	0.2615	0.5

Table 3: This table shows the top five occupations most exposed to AGI augmentation. Occupations marked with * do not have any information scores and therefore no weights assigned to the tasks. Occupations marked with ** have five or less tasks, potentially distorting results.

C Additional Visualizations

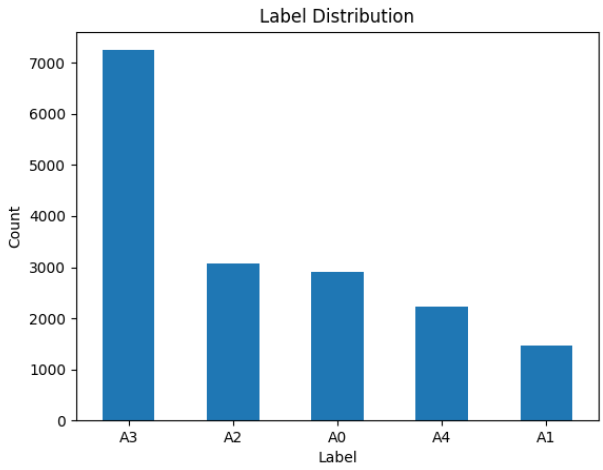

Figure 3: This chart shows the distribution of labels across 16.916 Tasks of the O*Net database. The labels represent the AGI-automation exposure from A0 (no exposure) to A4 (full exposure, meaning an AGI can autonomously execute the task without human oversight).

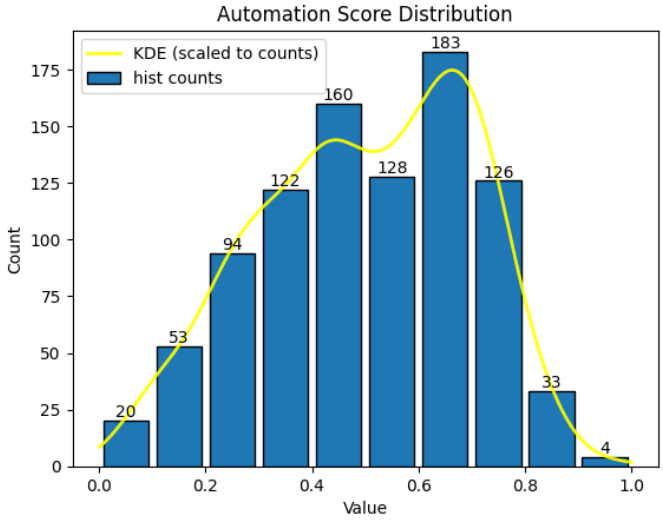

Figure 4: This chart shows the automation score distribution for all occupations in the O*Net occupational database (n=923).

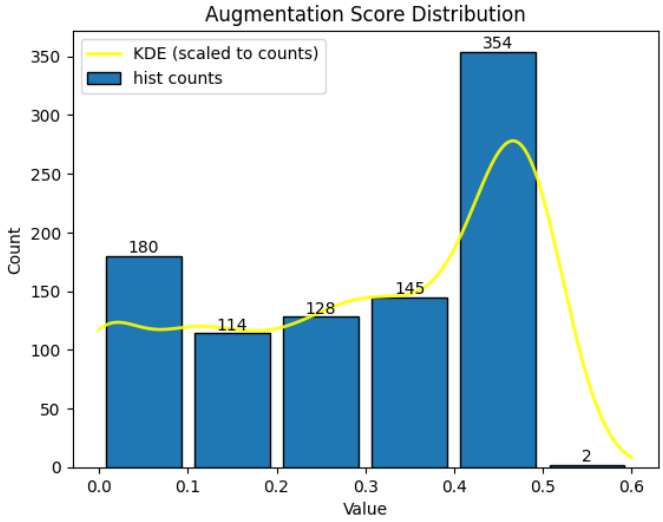

Figure 5: This chart shows the augmentation score distribution for all occupations in the O*Net occupational database (n=923).

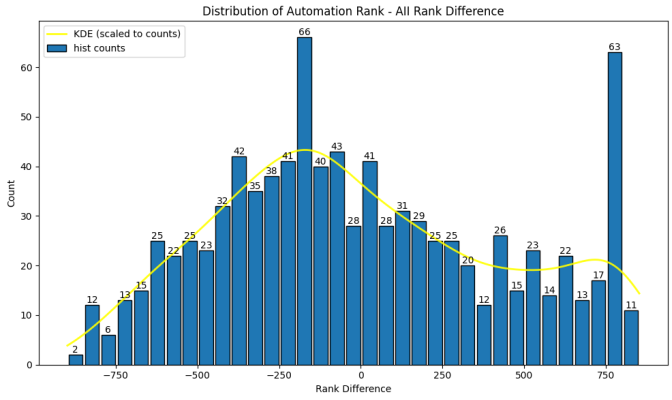

Figure 6: This chart shows the distribution of rank differences of the Artificial Intelligence Impact (AII) rank and automation score rank for all occupations in the O*Net occupational database (n=923). The rank of an occupation is higher, the higher the automation score/AII. The difference then shows how much more this occupation is considered to be exposed to AGI automation than general Artificial Intelligence Impact.

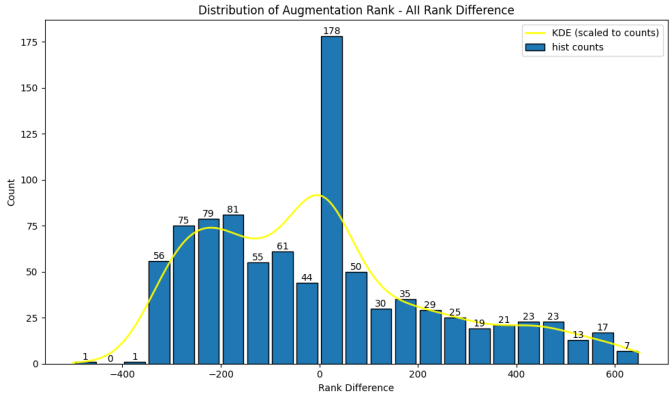

Figure 7: This chart shows the distribution of rank differences of the Artificial Intelligence Impact (AII) rank and augmentation score rank for all occupations in the O*Net occupational database (n=923). The difference then shows how much more this occupation is considered to be exposed to AGI augmentation than general Artificial Intelligence Impact.

D Risk Scenario Evaluation

D.1 Evaluation Criteria

Each scenario is evaluated along eight criteria (cf. [2]). The first criterion,*AGI-Specificity*, is treated as a binary filter rather than a scored metric, as it determines whether a scenario qualifies as genuinely AGI-specific. The other seven criteria are scored on a scale from 1 (lowest) to 5 (highest).

Table 4: Evaluation criteria, their dimensions, and definitions.

Criterion	Definition
AGI-Specificity	Does the scenario involve the use of AGI (not simple AI)? Is this risk ONLY possible with AGI?
Plausibility	Is the scenario a realistic risk outcome considering the task, industry, and job?
Connection	Is there a direct, coherent link between the task and the scenario?
Usefulness	Does the scenario provide a high-quality signal for policy-makers, auditors, and compliance people?
Actionability	Does the scenario provide a concrete starting point for a control or policy?
Detail	Does the scenario specify who is involved, how the risk is generated, and what the risk is?
Complexity	Does understanding the scenario require domain-specific job expertise?
Novelty	Does the scenario describe an emergent risk not typically found in current literature?

D.2 Evaluation Results

A sample of 100 risk scenarios derived from 20 occupations and their respective, AGI-exposed tasks from the O*NET Occupational Database, was evaluated. For the occupations with the highest Automation Score – AII rank difference 29 out of 50 were AGI-specific; For the occupations with the highest Augmentation Score – AII rank difference Only 19 out of 50 were AGI-specific. The tables 5 and 6 reports the mean and standard deviation for each scored criterion the respective 50 risk scenarios for the occupations with the highest positive difference in rank from Automation score/Augmentation score and AII .

Table 5: Average scores per evaluation criterion (scale 1-5) and standard deviation for 50 AGI automation risk scenarios.

Criterion	Mean	Std Dev
Connection	3.695	0.7254
Usefulness	3.465	0.6586
Plausibility	3.84	0.5687
Actionability	2.925	0.4410
Detail	3.42	0.6477
Complexity	3.01	0.7692
Novelty	3.23	0.7364

Table 6: Average scores per evaluation criterion (scale 1-5) and standard deviation for 50 AGI augmentation risk scenarios.

Criterion	Mean	Std Dev
Connection	3.7	0.7825
Usefulness	3.525	0.4691
Plausibility	3.71	0.4932
Actionability	3.28	0.4029
Detail	3.515	0.6296
Complexity	3.16	0.6522
Novelty	3.355	0.6249