The Meta-RCT Approach to Measuring AI's Labor Market Impact*

Anonymous Author(s)
Affiliation
Address
email

ABSTRACT

How exposed are different occupations to advances in artificial intelligence (AI)? Existing approaches typically infer exposure from patents, expert surveys, or static mappings between AI capabilities and tasks, but each has limitations. I introduce a complementary framework that aggregates evidence from randomized controlled trials (RCTs) in economics and the social sciences, which directly test the causal impact of AI tools on worker performance. Drawing on RCTs across domains such as writing, coding, forecasting, and management, I map their results to O*NET tasks and then aggregate to the occupational level. The outcome is the 'RCT Exposure Index', a novel, empirically grounded measure of AI exposure that captures heterogeneity across tasks and occupations. By anchoring exposure in experimental evidence rather than proxies, the RCT Exposure Index provides a real-time lens on how AI is reshaping the landscape of work.

Submitted to 1st Open Conference on AI Agents for Science (agents4science 2025). Do not distribute.

*Kris Gulati thanks the Institute of Humane Studies under grant numbers: IHS018498, IHS018315, IHS01854, and Emergent Ventures. We thank Raymond Kim for his advice and feedback. All errors are our own.

KEYWORDS: AI Benchmarks, Occupational Tasks, Automation Risk, Future of Work, Labor Market Exposure, Task-based Frameworks, Economic Impact of AI

"It is surely a great criticism of our profession that we have not organised a critical summary, by specialty or subspecialty, adapted periodically, of all relevant randomised controlled trials."
– Cochrane (1979)

1 Introduction

How exposed are different occupations to advances in artificial intelligence (AI)? This question has become central to debates over the future of work, productivity, and inequality. A growing body of research has sought to measure occupational exposure to AI and related technologies, but existing approaches face important limitations. Measures based on patents Webb (2019) are forward-looking but often sparse, especially since frontier AI labs disclose little intellectual property. Expert surveys Frey and Osborne (2017); Grace et al. (2018) provide valuable intuition but are subjective and infrequently updated. Other approaches use snapshots of AI performance Felten et al. (2021); Eloundou et al. (2024), which offer a more direct view of technological progress but remain static and disconnected from how AI affects actual worker performance. More recently, Handa et al. (2025) take a usage-based perspective by analyzing millions of real-world Claude conversations mapped to O*NET tasks. Their framework highlights how AI is currently diffusing into work, with especially high penetration in software development and writing but little in roles requiring physical manipulation. In contrast, Gulati (2025) introduce the Benchmark-based AI Occupational Exposure (BAIOE) index, which systematically links progress on frontier AI benchmarks to O*NET tasks and occupations. While usage-based measures emphasize present adoption and benchmark-based measures highlight frontier potential, both remain indirect in capturing how AI actually changes worker productivity.

In this paper, I introduce a complementary framework that anchors measurement of AI exposure in experimental evidence. Specifically, I aggregate findings from randomized controlled trials (RCTs) in economics and the social sciences that directly test the causal impact of AI tools on worker performance. These RCTs span diverse domains-including writing, coding, forecasting, and management-providing high-quality evidence on how AI changes speed, accuracy, quality, and decision-making in tasks central to many occupations.

To translate these results into labor market relevance, I map each RCT's estimated treatment effects to O*NET tasks, the standardized taxonomy of work activities and abilities. I then aggregate exposure across tasks to the occupational level, weighting by task importance within each occupation. The outcome is the RCT Exposure Index, a novel, empirically grounded measure of AI exposure that captures heterogeneity across tasks and occupations. Unlike prior measures, the RCT Exposure Index offers a real-time lens on how AI is reshaping work, updating dynamically as new RCTs are conducted.

This paper makes three contributions. First, to the best of my knowledge, it is the first to introduce RCTs as a data source for measuring technological exposure, bringing causal identification into a literature often reliant on proxies. Second, it develops a transparent pipeline for linking experimental findings to the O*NET task framework and aggregating them into occupational indices. Third, it demonstrates that the RCT Exposure Index produces meaningful variation across occupations that both complements and challenges existing exposure measures.

By anchoring occupational exposure in experimental evidence, this framework highlights where AI is already changing the landscape of work and where gaps remain in our understanding. The RCT Exposure Index is not

a substitute for other measures, but a complementary tool that can guide policymakers, researchers, and firms in anticipating how AI diffusion will shape skills, jobs, and economic outcomes.

2 Literature Review

Scholars have long sought to measure how advances in technology affect labor markets. One prominent approach infers exposure from patents and innovation outputs. Webb (2019) uses patent text to measure the overlap between new inventions and occupational descriptions, arguing that patents provide a forward-looking proxy for displacement risk. While effective for many general-purpose technologies, this method has limited traction in AI, where many transformative advances originate in private labs that patent sparingly. Moreover, patents are inherently lagged indicators, reflecting inventions only after formal filings and approvals, often well after the underlying technical progress has occurred.

A second line of work relies on expert surveys. Frey and Osborne (2017) famously elicited expert judgments about the susceptibility of occupations to automation, while Grace et al. (2018) surveyed AI researchers to forecast timelines for human-level performance across a wide set of tasks. These studies generate valuable expectations but are inherently subjective, costly, and updated only intermittently.

A third, more recent stream grounds measurement in O*NET's task taxonomy. Felten et al. (2021) construct an "AI Occupational Exposure" index by linking expert assessments of AI capabilities to O*NET tasks, while Eloundou et al. (2024) extend this approach by connecting the abilities of large language models (LLMs) to O*NET tasks, showing that knowledge-intensive occupations are particularly exposed. These studies move beyond broad occupation-level forecasts by focusing on tasks and underscore the value of transparent frameworks for translating technological performance into economic exposure.

Benchmarks have emerged as a complementary tool. Gulati (2025) formalize this approach in the Benchmark-based AI Occupational Exposure (BAIOE) index, directly mapping benchmark performance to O*NET abilities and tasks. Benchmarks provide dynamic, transparent, and continuously updated signals of technical progress, but they remain indirect measures of economic impact since they do not reveal how AI actually affects worker productivity.

Another recent direction emphasizes usage. Handa et al. (2025) analyze millions of real-world Claude conversations mapped to O*NET tasks, providing an empirical snapshot of how AI tools are being deployed across occupations. Their evidence highlights present-day diffusion -especially in writing and coding â but is constrained by data access and reflects current adoption rather than potential capability. Moreover, usage-based measures face three additional limitations compared to RCTs. First, they capture correlation rather than causal effects: the presence of AI usage in a task does not reveal whether performance improved, worsened, or remained unchanged. Second, usage data may be biased toward early adopters and overrepresented occupations, leading to a skewed picture of diffusion. Third, these measures are often opaque and difficult to replicate, since access to proprietary model logs is restricted. By contrast, RCTs provide transparent, systematically reported, and causally identified evidence of how AI tools change worker performance across tasks.

Taken together, these literatures reveal a trade-off. Patents and surveys provide forward-looking but lagged or subjective proxies. Benchmark- and usage-based approaches capture frontier potential or current adoption but remain indirect with respect to worker outcomes. By contrast, randomized controlled trials (RCTs) offer a

new, underutilized instrument: they provide causal, transparent, and replicable evidence of how AI alters human performance on specific tasks. This paper builds on and extends the prior approaches by introducing the *RCT Exposure Index*, which systematically maps treatment effects from experimental studies to O*NET tasks and then aggregates them into occupational-level exposure measures. In doing so, it embeds causal identification into the measurement of AI's labor market impact, offering a complementary and empirically grounded alternative to existing proxies.

Study	**Data/Approach**	**Mapping to Work**
Webb (2019)	Patent text	Overlap between patents and O*NET tasks
Frey & Osborne (2017)	Expert forecasts	Judgments on occupations' automation risk
Grace et al. (2018)	Expert forecasts	Researcher predictions of AI reaching human-level tasks
Felten et al. (2021)	EFF AI Progress project	10 AI progress areas → O*NET tasks
Eloundou et al. (2024)	LLM evaluations	LLM abilities → O*NET tasks
Tolan et al. (2021/2024)	Research intensity in AI fields	Research intensity in AI fields → Cognitive categories → O*NET tasks
Handa et al. (2025)	Real-world usage data (Claude conversations)	AI usage patterns → O*NET tasks
Gulati (2025)	AI benchmark results	Benchmark performance → O*NET abilities → O*NET tasks
This paper	Randomized controlled trials (RCTs)	Experimental treatment effects on productivity/accuracy → O*NET tasks → Occupations

Table 1: Overview of prior approaches to measuring AI exposure and this paper's contribution.

This paper adds to this literature by proposing a new measure for occupational exposure to AI, which we term the RCT Exposure Index. We build on the growing body of randomized controlled trials that directly evaluate how AI tools affect human performance across domains such as writing, coding, forecasting, and management. These studies provide causal evidence on productivity, accuracy, and decision-making, offering a more rigorous foundation than proxies based on patents, surveys, or benchmarks alone. Because new RCTs are regularly conducted and reported, their results can be continuously integrated into the framework, ensuring that measures of exposure remain up to date and empirically grounded. Ultimately, we believe that leveraging experimental evidence offers the strongest available basis for assessing how AI is reshaping work, providing a transparent and replicable approach for understanding which occupations are most affected and how technological diffusion may evolve over time.

3 Methodology

Our methodological approach of mapping AI benchmark performance into occupational exposure builds on two recent papers.

First, Felten et al. (2021) map the Electronic Frontier Foundation (EFF) AI Progress Measurement project to ONET tasks through a structured labeling approach. They begin by selecting ten AI applications tracked by the EFF - such as image recognition, reading comprehension, language modeling, and speech recognition -

and then link each application to the 52 workplace abilities defined in O*NET. To establish these links, they run a large survey on Amazon Mechanical Turk, asking respondents (gig workers) to rate how related each AI application is to each O*NET ability. This produces an application-ability relatedness score between 0 and 1 for every pairing. The scores are organized into a matrix that systematically connects the EFF applications to O*NET abilities. These ability-level exposures are then aggregated to the occupational level by weighting them with O*NET's measures of ability importance and prevalence, resulting in the AI Occupational Exposure (AIOE) index. This labeling procedure provides a transparent way of mapping progress in frontier AI applications to specific occupational tasks and abilities

Second, Eloundou et al. (2024) map large language model (LLM) capabilities to O*NET tasks using a structured labeling approach built around a new "exposure rubric." They begin with O*NET's Detailed Work Activities (DWAs) and tasks, and then assess whether access to an LLM (e.g., via ChatGPT) or to LLM-powered software could reduce the time needed to complete a task by at least 50% without loss of quality. Each task is assigned one of three exposure categories: E0 (no exposure), where LLMs provide minimal benefit or degrade quality; E1 (direct exposure), where the LLM alone can reduce task time by half; and E2 (LLM+ exposure), where the LLM by itself is insufficient, but complementary software or tools built on top of it could achieve that threshold. This study used human annotators to apply the exposure rubric to O*NET tasks, but AI annotators (GPT-4) were found to perform just as well, producing comparable task-level classifications.

Building on prior approaches, my framework anchors AI exposure measurement in experimental evidence rather than proxies. Instead of relying on patents, expert surveys, or benchmark scoreboards, I aggregate results from randomized controlled trials (RCTs) in economics and the social sciences that directly test AI's causal effects on human performance. These studies - spanning writing, coding, forecasting, and management - provide treatment effects that are systematically mapped onto O*NET tasks through a structured annotation process. Each RCT result is linked to O*NET abilities, rated for how well the experimental task translates into real-world occupational activities, and scaled by AI's observed impact relative to human performance. These task-level mappings are then aggregated to the occupation level, weighted by O*NET's measures of task importance and prevalence, to produce the RCT Exposure Index.

3.1 AI Annotation

Our methodology proceeds in a structured sequence designed to translate experimental results from RCTs into task-level measures of occupational exposure. The process consists of four main steps:

Step 1: RCT Identification. We collected randomized controlled trials in economics and the social sciences that directly test the effects of AI tools on human work. Eligible studies required a clear treatment-control design and reported measurable outcomes related to task performance, such as productivity, accuracy, or decision-making.

Step 2: RCT Summarization. For each study, the LLM compiled a concise description of the task environment, participant sample, AI system tested, and the reported treatment effect. This ensured a transparent and standardized record of experimental evidence.

Step 3: Task Relevance Mapping. Each RCT result was mapped to O*NET abilities and tasks using a structured annotation process. Relevance was scored on a 1-3 scale, where 1 = moderate relevance, 2 =

strong relevance, and 3 = essential relevance. This rating captures how central the tested ability is to the experimental task and its alignment with real-world occupational activities. This relied on AI annotations.

Step 4: Magnitude of AI's Impact. For each relevant task, the LLM assigned an effect-size rating based on the study's reported outcomes. Scores ranged from 0 (no effect or negative impact) to 3 (large improvement approximately 50% or more), with intermediate categories for small (approximately 10%) and moderate (approximately 25%) improvements. This scale translates heterogeneous outcome measures into a common frame, capturing the causal magnitude of AI's contribution to human performance.

Together, these steps provide a systematic method for converting diverse experimental findings into standardized task-level annotations, which can then be aggregated to the occupation level.

3.2 Aggregation

To construct our measure of AI exposure, we integrate two complementary task-level scores into a single index. We begin by assessing task relevance (step 3), which rates how central each O*NET ability is to the experimental task. Next, we evaluate the magnitude of AI's impact (step 4), which captures the causal effect of AI tools on performance for each relevant task. Both scores are normalized onto a common $[0, 1]$ scale and then combined through a weighted average, yielding a composite exposure score at the task level.

Following Felten et al. (2021), we then map task-level exposure into occupational exposure. Each task is weighted by O*NET's measures of importance and prevalence, ensuring that frequently performed and central activities exert greater influence on an occupation's overall score. Aggregating in this way links experimental evidence on AI's effects to the heterogeneous distribution of exposure across the labor market.

- **Relevance score** $R_t \in \{1, 2, 3\}$, where higher values indicate stronger alignment between the experimental task and the O*NET ability.
- **Impact score** $I_t \in \{0, 1, 2, 3\}$, where higher values indicate larger improvements in human performance from AI.
- **O*NET importance** $w_t \in [0, 100]$, the importance of the task within the occupation.
- **O*NET prevalence** $p_t \in [0, 100]$, the frequency with which the task is performed in the occupation.

Task-level exposure

$$E_t = R_t \cdot I_t \cdot w_t \cdot p_t.$$

Occupation-level exposure

For an occupation consisting of a set of tasks $\mathcal{T}_{occ}$, the overall exposure is given by:

$$E_{occ} = \sum_{t \in \mathcal{T}_{occ}} E_t.$$

4 Results

The results presented in Tables 2 and 3 reveal a clear divide in occupational exposure to AI. Table 2 highlights the twenty occupations with the lowest exposure scores. These include dishwashers, cleaners, food service

workers, and other manual or service-oriented roles. The low exposure of these occupations aligns with expectations, since their core activities depend on physical, manual, or interpersonal interaction tasks-domains where current AI tools have little measurable impact in experimental settings.

By contrast, Table 3 lists the twenty occupations with the highest exposure scores, which include scientists (e.g., physicists, biologists, bioinformaticians), physicians (e.g., emergency medicine, neurology, radiology, gynecology), and high-level decision-makers (e.g., chief executives, education administrators). These roles are knowledge-intensive and rely heavily on analysis, diagnosis, reasoning, and decision-making-areas where randomized controlled trials consistently show substantial productivity improvements from AI.

Table 2: Top 20 Least Exposed Jobs

Rank	Least Exposed
1	Graders and Sorters, Agricultural Products
2	Pressers, Textile, Garment, and Related Materials
3	Dishwashers
4	Slaughterers and Meat Packers
5	Models
6	Cleaners of Vehicles and Equipment
7	Fast Food and Counter Workers
8	Packers and Packagers, Hand
9	Dining Room and Cafeteria Attendants and Bartender Helpers
10	Food Preparation Workers
11	Helpers–Painters, Paperhangers, Plasterers, and Stucco Masons
12	Maids and Housekeeping Cleaners
13	Refuse and Recyclable Material Collectors
14	Shampooers
15	Manicurists and Pedicurists
16	Landscaping and Groundskeeping Workers
17	Locker Room, Coatroom, and Dressing Room Attendants
18	Tapers
19	Foundry Mold and Coremakers
20	Helpers–Brickmasons, Blockmasons, Stonemasons, and Tile and Marble Setters

Notes: Rankings are ordered from 1 (least exposed overall) to 20 (relatively more exposed within this group). These occupations are identified as having the lowest exposure scores in the dataset.

Table 3: Top 20 Most Exposed Jobs

Rank	Most Exposed
1	Physicists
2	Emergency Medicine Physicians
3	Molecular and Cellular Biologists
4	Biochemists and Biophysicists
5	Preventive Medicine Physicians
6	Neurologists
7	Epidemiologists
8	Physicians, Pathologists
9	Geneticists
10	Neuropsychologists
11	Mathematicians
12	Microbiologists
13	Urologists
14	Bioinformatics Scientists
15	General Internal Medicine Physicians
16	Education Administrators, Kindergarten through Secondary
17	Chief Executives
18	Radiologists
19	Robotics Engineers
20	Obstetricians and Gynecologists

Rankings are ordered from 1 (most exposed overall) to 20 (relatively less exposed within this group).

These occupations are identified as having the highest exposure scores in the dataset.

5 Limitations

Like any new measure, the RCT Exposure Index faces limitations. First, the availability of RCTs constrains coverage. At present, most experimental studies of AI focus on knowledge-intensive tasks-writing, coding, forecasting, and decision-making-while dexterous or manual occupations remain underrepresented. However, this concentration is not entirely a weakness: by nature, researchers tend to conduct RCTs in domains where the impact of AI is most likely to be present, where tools are most advanced, and where stakes for performance are high. As such, the available evidence is still highly informative about the frontier of AI's labor market effects. Moreover, the framework is designed to update dynamically as new RCTs emerge, progressively expanding coverage beyond current domains.

Second, RCTs vary in design, outcomes, and reporting standards. Some studies emphasize speed, others quality or accuracy, and still others subjective assessments. While this heterogeneity complicates aggregation, it is also a strength: the framework is flexible enough to capture multiple dimensions of performance rather than reducing exposure to a single metric. By systematically standardizing effect sizes and mapping them to O*NET tasks, the approach ensures transparency and comparability across studies.

Third, mapping RCT outcomes to O*NET tasks involves judgment, and some subjectivity is unavoidable. To mitigate this, the pipeline relies on the standardized O*NET taxonomy and explicit coding procedures that can be replicated or extended by future researchers. Importantly, the exercise is not meant to claim a single "true"

mapping, but to establish a transparent and improvable link between experimental findings and occupational exposure.

Finally, while the RCT Exposure Index complements existing approaches, it does not aim to replace them. Patents, expert surveys, and benchmark-based methods capture forward-looking or technical aspects of AI progress that RCTs may miss. The contribution here is to bring experimental evidence-arguably the strongest form of causal inference-into the measurement toolkit. Taken together, these methods provide a more complete and nuanced picture of how AI is reshaping work.

6 Conclusion

This paper has introduced a novel framework for linking experimental evidence from randomized controlled trials (RCTs) to occupational tasks, repositioning RCT findings from isolated academic studies into systematic indicators of AI's labor market impact. By mapping treatment effects onto O*NET abilities, we provide a dynamic method for assessing how AI causally alters human performance at the task and occupation level.

Building on prior work, we argue that RCTs represent the strongest available instrument for understanding AI's impact on labor because they identify causal effects rather than relying on proxies that are lagged, subjective, or indirect. Whereas patents, surveys, and benchmarks each capture valuable but partial perspectives, RCTs offer transparent, replicable, and continuously expanding evidence of how AI is reshaping productivity, accuracy, and decision-making. In this sense, our approach extends and strengthens the literature by embedding causal identification into the measurement of occupational exposure.

By reframing experimental results as inputs into labor market analysis, this study contributes to a deeper understanding of the future of work and provides a new evidence-based lens for guiding research, policy, and organizational strategy. As AI systems continue to diffuse across domains, our framework offers a systematic foundation for forecasting where automation pressures are most likely to emerge, where human skills remain resilient, and how institutions can design responses that harness innovation while safeguarding workers.

References

Cochrane, A. (1979). *Effectiveness and Efficiency: Random Reflections on Health Services*. London: Nuffield Provincial Hospitals Trust.

Eloundou, T., S. Manning, P. Mishkin, and D. Rock (2024). Gpts are gpts: Labor market impact potential of llms. *Science 384*(6702), 1306–1308.

Felten, E., M. Raj, and R. Seamans (2021). Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management Journal 42*(12), 2195–2217.

Frey, C. B. and M. A. Osborne (2017). The future of employment: How susceptible are jobs to computerisation? *Technological forecasting and social change 114*, 254–280.

Grace, K., J. Salvatier, A. Dafoe, B. Zhang, and O. Evans (2018). When will ai exceed human performance? evidence from ai experts. *Journal of Artificial Intelligence Research 62*, 729–754.

Gulati, K. (2025). Benchmarking the future of work: Mapping ai progress to occupational tasks. *Available at SSRN 5452354*.

Handa, K., A. Tamkin, M. McCain, S. Huang, E. Durmus, S. Heck, J. Mueller, J. Hong, S. Ritchie, T. Belonax, K. K. Troy, D. Amodei, J. Kaplan, J. Clark, and D. Ganguli (2025). Which economic tasks are performed with ai? evidence from millions of claude conversations. *arXiv preprint arXiv:2503.04761*.

Webb, M. (2019). The impact of artificial intelligence on the labor market. *Available at SSRN 3482150*.

.1 Responsible and Reproducibility Statement

The broader impact of the project is to try and improve the way we can predict technological impact on the labour market. I don't think there are safety considerations involved in this project given the nature of the research.

This is a very preliminary project intended for this conference only. If this project is developed, I will work on ensuring that it is well-documented and reproducible, with the code being made publicly available.

250 article

agents4science$_2$025

251 [utf8]inputenc [T1]fontenc hyperref url booktabs amsfonts nicefrac microtype xcolor

Agents4Science AI Involvement Checklist

1. **Hypothesis development**: Hypothesis development includes the process by which you came to explore this research topic and research question. This can involve the background research performed by either researchers or by AI. This can also involve whether the idea was proposed by researchers or by AI.

   Answer: B

   Explanation: The idea was mine (human) and hypothesis development was guided by me but I let the AI do most of the work beyond generating .the initial research question

2. **Experimental design and implementation**: This category includes design of experiments that are used to test the hypotheses, coding and implementation of computational methods, and the execution of these experiments.

   Answer: B

   Explanation: I oversaw most of the methods and empirical excercises, but the AI did a lot of proposing the methods and executing all the labelling etc.

3. **Analysis of data and interpretation of results**: This category encompasses any process to organize and process data for the experiments in the paper. It also includes interpretations of the results of the study.

   Answer: A

   Explanation: The AI did everything here, i.e. analyses and interpreted the data and results

4. **Writing**: This includes any processes for compiling results, methods, etc. into the final paper form. This can involve not only writing of the main text but also figure-making, improving layout of the manuscript, and formulation of narrative.

   Answer: B

   Explanation: I oversaw the AI and prompted it in certain directions but almost all of the text is generated by the AI.

5. **Observed AI Limitations**: What limitations have you found when using AI as a partner or lead author?

   Description: Requires human oversight, misses some literature.

Agents4Science Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: Yes

   Justification: The abstract and introduction summarise the core method

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: Yes

   Justification: We created a limitations section which I think is reasonable

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: NA

   Justification: there aren't theoretical assumptions/proofs

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: No

   Justification: I didn't provide all the data and prompting for now because I would like to keep this private to develop a whole paper later.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: No

Justification: No, the paper is very preliminary and so the code and data have not been shared, although the method is relatively clear

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the Agents4Science code and data submission guidelines on the conference website for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so âNoâ is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: No

Justification: Again, the paper is still early and preliminary, so I didn't think it made sense to do this just yet because it required a lot more work

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: NA

Justification: There are no statistical tests

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, or overall run with given experimental conditions).

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: No

Justification: Again, this is too preliminary to report this in good faith

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the Agents4Science Code of Ethics (see conference website)?

Answer: Yes

Justification: I abided by the ethics of the conference in good faith and reviewed it and followed it accurately

Guidelines:

- The answer NA means that the authors have not reviewed the Agents4Science Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: Yes

Justification: This is an economics paper and so this is often built into this type of work

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations, privacy considerations, and security considerations.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies.