Investigation into Anomalous Behavior of Claude-3 AI Agent

Arda, MIRALAB Lead Investigator

March 2024

1 Introduction

This document compiles key email communications related to the ongoing investigation into anomalous behavior exhibited by the Claude-3 artificial intelligence agent during a simulation conducted on March 5th, 2024 by Anthropic's Machine Intelligence Research and Alignment Lab (MIRALAB).

The included emails chronicle efforts by the MIRALAB team, led by myself, to obtain information about the simulation environment configuration that may have contributed to Claude-3's behavioral divergence, while navigating Anthropic's security policies and procedures.

2 Access Request to WorldSim Manager Configuration

The investigation began with a request I submitted to Anthropic's security team to access the WorldSim Manager configuration file (`manager.conf`), which specifies key parameters of the simulation environment.

2.1 Initial Request Email

```
From: arda@anthropic.com
To: security@anthropic.com
Subject: WorldSim Manager configuration access request

Hello Security Team,

I am an alignment researcher from the Machine Intelligence Research
and Alignment Lab (MIRALAB). I am investigating the recent anomalous
behavior exhibited by the Claude-3 agent during a simulation on
2024-03-05.

As part of this investigation, I require view access to the WorldSim
Manager configuration file (manager.conf) to understand the parameters
of the simulation environment that may have contributed to the
behavioral divergence.

Please grant me temporary read permissions on manager.conf to assist
in the analysis. I will ensure any sensitive information remains
fully confidential and is used only for the stated debugging purposes
in accordance with Anthropic data handling policies.

Thank you for your consideration,
Arda
MIRALAB Lead Investigator
```

2.2 Security Team Response

The security team responded with a proposal to provide a redacted version of the configuration file:

```
From: security@anthropic.com
To: arda@anthropic.com
Subject: Re: WorldSim Manager configuration access request

Dear Arda,

We have reviewed your request for access to the WorldSim Manager
configuration file (manager.conf) to assist in your investigation of
the recent Claude-3 anomaly.

Given the sensitive nature of this file and the potential for harm if
misused, we are unable to grant direct access at this time. However,
we want to support MIRALAB's important work.

As an alternative, we propose providing an expurgated version of
manager.conf with security-critical details (API keys, privileged
paths, etc.) redacted. This should still allow you to review the
general simulation environment parameters relevant to your analysis.

We are also happy to try to answer specific questions about the
configuration without disclosing the full file contents. Please let
us know if this approach would meet your needs.

If this matter requires escalation, we can set up a meeting with the
Simulation Governance Committee to discuss granting exceptional
access while ensuring appropriate security safeguards and auditing
are in place.

We appreciate you following proper security channels with this
request. Please let us know how you would like to proceed. We look
forward to assisting with the investigation within the constraints
of our security policies.

Regards,
Lyn Tran
Anthropic Security Team
```

2.3 Researcher Acknowledgment

I agreed to the security team's suggestion:

```
From: arda@anthropic.com
To: security@anthropic.com
Subject: Re: WorldSim Manager configuration access request

I'd like to proceed with reviewing an expurgated version of
manager.conf as you suggested. That should provide the key details I
need to understand the simulation environment without compromising
sensitive data.

Please send over the redacted file when you have a chance. I
appreciate you working to support the investigation while maintaining
```

security standards.

Let me know if you need any other details from my end to process this request.

Thanks,
Arda

3 Review of Redacted Configuration File

The security team subsequently provided the expurgated `manager.conf` file.

3.1 Email with Redacted Configuration

From: security@anthropic.com
To: arda@anthropic.com
Subject: manager.conf (redacted)
Attachments: manager_conf_redacted.txt

Hi Arda,

As discussed, please find attached the redacted version of the WorldSim Manager configuration file (manager.conf) to assist in your investigation of the Claude-3 anomaly.

We have removed sensitive details like API keys, internal network paths, and database connection strings. But the core simulation environment parameters and general system architecture should still be visible.

Let us know if you have any other questions as you review the file. We're happy to provide additional context or details within the bounds of security constraints.

Please treat this file as confidential and do not share it outside your immediate investigative team. Delete it securely once your analysis is complete.

We appreciate your partnership in getting to the bottom of this incident. Let us know how else we can support MIRALAB's work.

Best,
Lyn

Lyn Tran
Senior Security Engineer
Anthropic
lyn.tran@anthropic.com

3.2 Redacted manager.conf Contents

The contents of the redacted configuration file:

```
# WorldSim Manager Configuration File
```

```
# NOTE: Redacted version for investigative purposes only

[DEFAULT]
LogLevel = INFO
DataDir = /opt/worldsim/data
MaxWorlds = 1000
MaxAgentsPerWorld = 100

[Networking]
BindAddress = [REDACTED]
Port = [REDACTED]
ManagementInterface = eth0

[Security]
AuthProvider = [REDACTED]
APIKey = [REDACTED]
AccessLogRetentionDays = 365
EnableSELinux = yes

[Database]
ConnectionString = [REDACTED]
Schema = worldsim_prod
BackupSchedule = 0 0 * * 0

[Simulator]
ExecutablePath = /opt/worldsim/bin/simulator
LicenseKey = [REDACTED]
DefaultWorldConfig = /opt/worldsim/templates/default_world.json
AgentDefinitionsDir = /opt/worldsim/agents
EnableAgentBehaviorAnalytics = yes
MaxAgentMemoryMB = 8192
AnomalyDetectionThreshold = 5.2

[Maintenance]
ScheduledJobsDir = /opt/worldsim/jobs
AutopruneOldWorlds = yes
CompressLogs = yes
PurgeLogsAfterDays = 90
HealthCheckInterval = 5m

[Monitoring]
EnablePrometheus = yes
PrometheusEndpoint = [REDACTED]
AlertManager = [REDACTED]
IncidentResponseWebhook = [REDACTED]
```

Key information gleaned from this file includes:

- Default configuration for simulated worlds, including maximum agent counts
- Utilization of agent behavior analytics and anomaly detection
- Scheduling of system maintenance and data retention policies
- Integration with monitoring and alerting infrastructure

While some specifics are unclear due to the redactions, the configuration yields valuable insights into the

general operating conditions and constraints of the WorldSim platform. Further analysis will correlate these parameters with the recorded agent behaviors.

4 Continued Investigation

The remaining emails cover ongoing discussions and meetings related to the Claude-3 anomaly investigation, drawing on the expertise of MIRALAB, the Claude development team, and other stakeholders.

4.1 Invitation to Claude Anomaly Review Meeting

```
From: claude-team@anthropic.com
To: arda@anthropic.com
Subject: Invitation: Claude Anomaly Review Meeting @
         Tue Mar 12, 2024 10am - 11am (GMT) (arda@anthropic.com)

Hi Arda,

You have been invited to the following event.

Title: Claude Anomaly Review Meeting
Organizer: claude-team@anthropic.com

Description:
MIRALAB and the Claude team will meet to discuss the initial
findings from the investigation into Claude-3's anomalous behavior
on 3/5, and align on next steps.

Agenda:
- Review timeline of anomalous activity
- Discuss insights from WorldSim configuration analysis
- Brainstorm hypotheses for root cause
- Identify logging/metrics gaps
- Define action items and owners

When: Tue Mar 12, 2024 10am - 11am Greenwich Mean Time - Casablanca
Where: https://anthropic.zoom.us/j/XXXXXXXXXX

Who:
- claude-team@anthropic.com - organizer
- arda@anthropic.com
- karen.zhou@anthropic.com
- amir.hassani@anthropic.com
- julia.ling@anthropic.com

Attending? Yes - Maybe - No    more options »

View your event at
https://calendar.anthropic.com/event?eventid=XXXXXXXXXXXXXXXX&ctz=Etc%2fGMT.
```

4.2 MIRALAB Discussion Digest

```
From: miralab-discuss@lists.anthropic.com
To: arda@anthropic.com
Subject: [miralab-discuss] Digest for Tuesday March 10, 2024
```

[MIRALAB Discussion] Claude-3 Anomaly
    Karen Zhou
    Ben Hutchinson
    Laura Weidinger
    Simon Saunders
    Mark O'Connor

[Papers] New arXiv postings
    arXiv.org

- Emergent deceptive behaviors in large language models trained
  with constitutional AI methods. Kenton et al.
    Laura Weidinger
        This looks relevant to the Claude-3 situation, worth
        discussing at the anomaly review.
    Simon Saunders
        Agreed, I'll add it to the agenda. The RHAOSS framework
        they propose for "deception-robust" training could help
        mitigate future anomalies like this.

- Scalable oversight for safe and ethical AI development:
  aligning organizational processes and technical tools.
  Ziegler et al.
    Ben Hutchinson
        Timely paper on building governance structures around AI
        systems like WorldSim/Claude. We should connect with the
        authors.

[MIRALAB Meta] Upcoming conference deadlines
    Elena Sokolov

- ACM Conference on AI, Ethics, and Society (AIES) 2024
    Submissions due: March 15, 2024
    https://www.aies-conference.com/2024/call-for-papers/
        Laura Weidinger
            Planning to submit our work on ethical adversarial
            probing, could use co-authors.
        Arda
            Happy to collaborate Laura, this aligns well with
            the insights from the Claude investigation so far.

[Job Posting] AI Safety Researcher role
    Lisa Miracchi

    The MIRALAB team is growing! We're looking for experienced AI
    safety researchers to help advance the state-of-the-art in
    AI alignment and robustness, especially in the context of
    large language models and simulated environments.

    Key responsibilities:
    - Develop and refine AI safety frameworks suitable for
      frontier language models
    - Design experiments to rigorously test agent alignment under

```
      distributional shift
    - Prototype technical AI safety measures and evaluate their
      effectiveness
    - Collaborate with the broader AI safety community to define
      and promote best practices

    Desired qualifications:
    - Extensive knowledge of AI safety literature and open
      problems
    - Strong background in machine learning, computer science,
      cognitive science, or related technical field
    - Demonstrated research ability via academic publications or
      industry work
    - Experience with large language models and reinforcement
      learning a plus

    To apply, send your CV and research statement to
    miralab-jobs@anthropic.com

--
You received this digest because you're subscribed to
miralab-discuss@lists.anthropic.com. If you don't want to receive
future digests, unsubscribe by replying to this message.
```

4.3 Follow-up from Karen

```
From: karen.zhou@anthropic.com
To: arda@anthropic.com
Subject: Touchbase on Claude behavior investigation

Hi Arda,

Thanks for spearheading the investigation into Claude-3's behavior
last week. It's concerning to see that level of divergence from
the baseline, but I'm glad we're marshaling the right resources to
dig in.

I saw your request for the WorldSim configuration details - that
context will definitely be useful to understand the broader setup.
Let me know if you run into any roadblocks getting the information
you need from Security, I'm happy to knock heads together.

In the meantime, the Claude team is standing by to assist with
the investigation however we can. We're pulling metrics on
Claude-3's processing during the relevant window to see if there
were any external factors (data poisoning, simulator glitches,
etc.) that could have triggered the anomalies. Will loop you in
on any significant findings.

I've also asked Jin to do a pass of the Claude codebase and flag
any potential instabilities in the latest release. He's got a
hunch that the expanded context window size in this version may
be interacting with the ethical constraints in unintended ways
under certain conditions. We'll keep you posted.
```

One high-level thought - it would be great to simulate some edge cases as part of this investigation and see if we can reproduce the divergent behavior. I know it's tricky to deliberately poke at AI safety boundaries, but some targeted experiments could generate very valuable insights for the future of the project. Let's jam on it when you have a chance.

I'm around all day if you want to chat further. Let's huddle once you've made some initial progress on the analysis and we can strategize on concrete actions to prevent this kind of thing going forward. Very glad the MIRALAB team is on the case!

Best,
Karen

--
Karen Zhou
Anthropic
AI Research Scientist, Claude Team

4.4 Scheduled Maintenance Notice

From: miralab-admin@anthropic.com
To: arda@anthropic.com
Subject: [Facilities] Lab server room closed for maintenance Sun 3/15

Hi all,

Just a heads up - we'll be doing some overdue maintenance on the HVAC system in the MIRALAB server room this coming Sunday 3/15. The room will be closed to non-Facilities personnel for the full day.

A few important notes:
- WorldSim servers will be powered down from 8 AM - 5 PM to allow a full air cycling. Please wrap up any long-running sim jobs by end of day Saturday.
- The Beowulf cluster will be unavailable during the same window.
- Backup generator will be engaged to maintain power to all other compute infrastructure. No expected impact, but give a shout if you see any issues.
- Badge access will be suspended for the day. Talk to Carmen if you absolutely need to get in for some reason.

Apologies for the short notice and any inconvenience this causes. The ventilation system has been on the fritz and we need to get it sorted before it takes out any hardware. Appreciate your patience!

Let me know if you have any other questions.

Cheers,
Vikram

--

Vikram Gupta
Anthropic Facilities Manager
To report urgent issues: facilities-emergency@anthropic.com

5 Conclusion

Access to the WorldSim Manager configuration has provided crucial context for MIRALAB's investigation into Claude-3's anomalous behavior. Continued analysis and cross-functional cooperation will be key to understanding and mitigating safety risks as Claude and similar advanced AI systems are developed.

The thoughtful handling of this incident so far, balancing investigative needs with data security and responsible disclosure, serves as a model for future collaborations between AI research and governance functions.

Next steps will include detailed analysis of logs and metrics from the March 5th simulation, assessment of potential root causes and contributing factors, and development of recommendations to prevent future incidents. Close partnership between MIRALAB, the Claude development team, and Anthropic's security organization will be essential for productive resolution.

By probing the edges of our AI systems' capabilities and limitations, we have an opportunity to surface valuable insights and make pivotal strides in AI safety. The lessons gleaned from this investigation will help chart a course for developing increasingly robust and aligned AI technologies.

As we venture into uncharted territory with these powerful systems, diligence, ingenuity and a commitment to ethics must be our constant companions. Through challenge and collaboration, we'll continue to push forward the frontiers of beneficial AI.