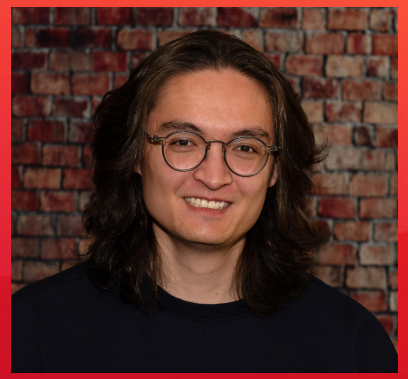
Fairness in Multi-Agent Systems for Software Engineering: An SDLC-Oriented Rapid Review



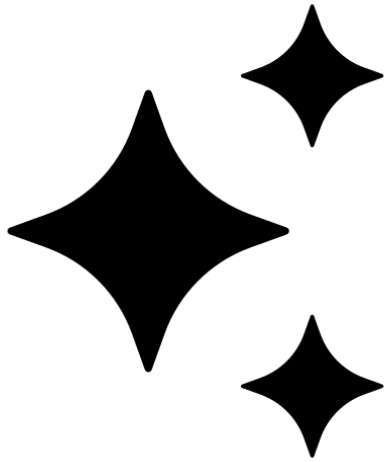
Corey Yang-Smith
corey.yangsmith@ucalgary.ca

Ronnie de Souza Santos
ronnie.desouzasantos@ucalgary.ca

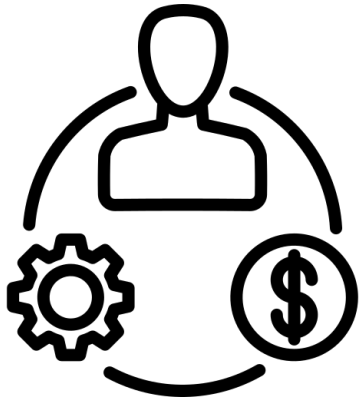
Ahmad Abdellatif
ahmad.abdellatif@ucalgary.ca

UNIVERSITY OF CALGARY

The Problem in the SDLC



84% developers in 2025 using AI

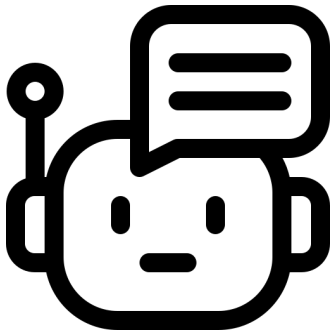


89% of organizations integrating AI

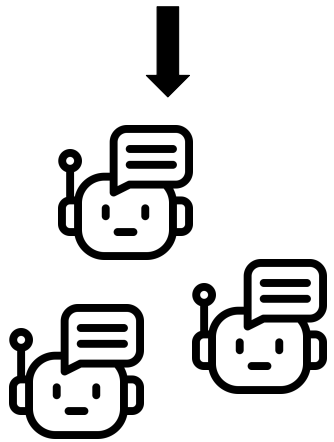
Source: Stack Overflow Developer Survey (2025)

Source: DORA - Strategies for Organizational AI Adoption (2025)

The Growth of Multi-Agent Systems



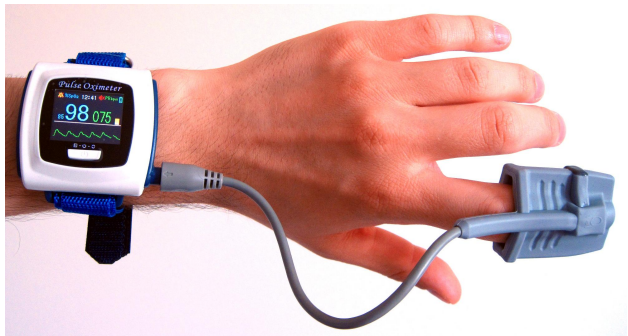
Chatbot QA

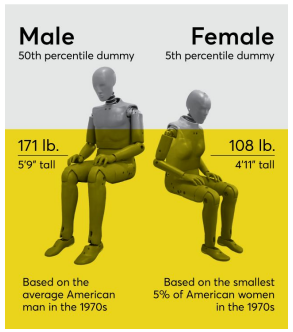


Multi-Agent Systems

94 papers on LLM-based multi-agent systems for SE

Fairness for SE + LLMs is underexplored





Source: [Designing LLM-based Multi-Agent Systems for Software Engineering Tasks (Cai, Arxiv 2025)]

When multiple AI agents plan, code, review, and test together, how do we understand fairness in their interactions?

RQ1: How do existing studies **define and measure fairness** in multi-agent systems?

RQ1: How do existing studies **define and measure fairness** in multi-agent systems?

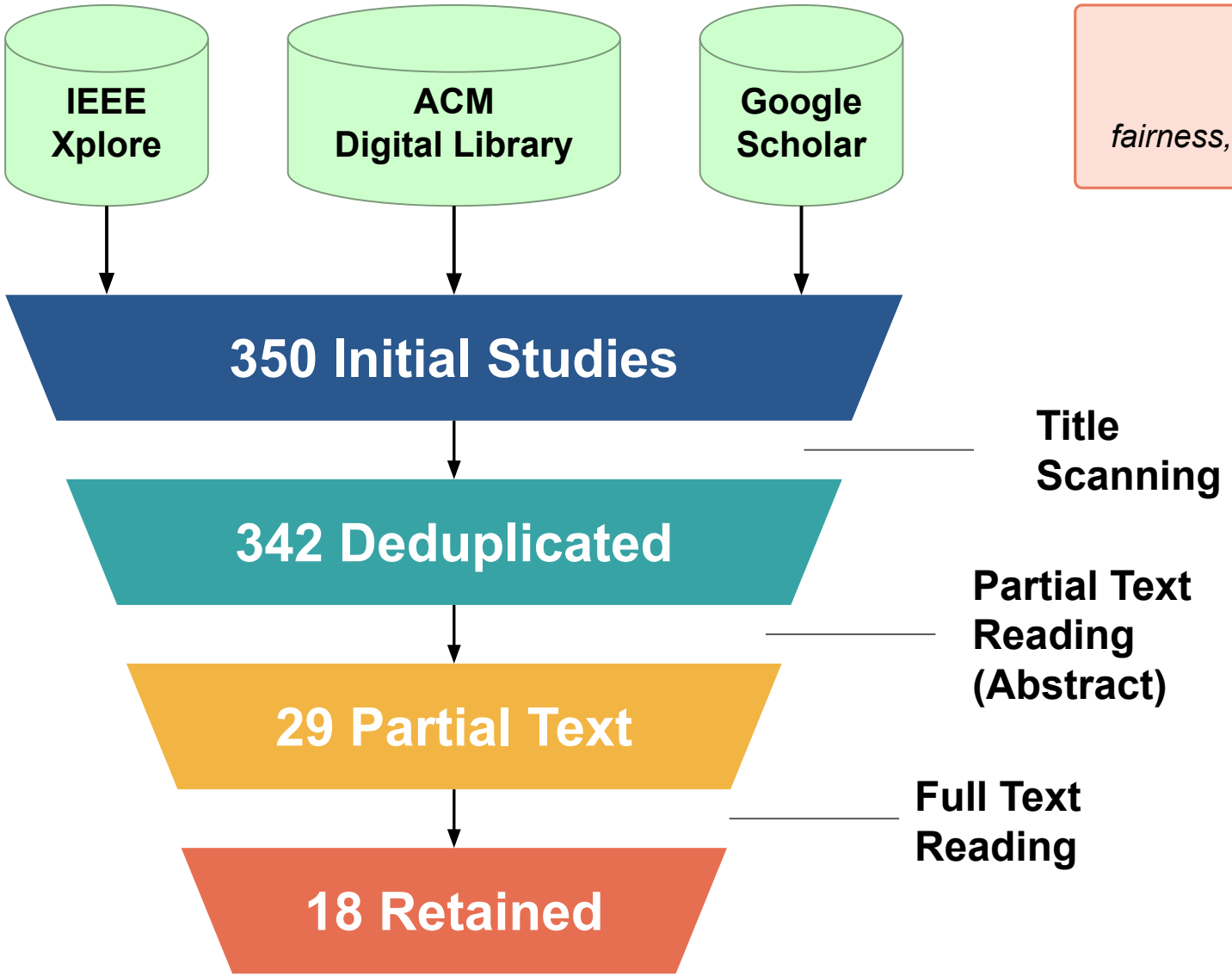
RQ2: What **biases or harms** appear across the SDLC when using these systems?

RQ1: How do existing studies **define and measure fairness** in multi-agent systems?

RQ2: What **biases or harms** appear across the SDLC when using these systems?

RQ3: What **gaps remain** in making multi-agent systems for software engineering **fair, accountable, and transparent?**

Methodology



Keyword Search

fairness, multi-agent systems, SDLC/software engineering

Extraction Schema

- Fairness Definitions (RQ1)
- Fairness Metrics (RQ1)
- Harms and Inequities (RQ2)
- SDLC Relevance (RQ2)
- Gaps and Future Work (RQ3)

RQ1: How do existing studies **define and measure fairness** in multi-agent systems?

Fairness Definitions

Bias Reduction

BIAS

Reducing stereotypes and demographic disparities across groups.

9 Studies

Inter-Agent Dynamics

MAS

Fairness as a **property of MAS interaction**, conformity, amplification, responsibility allocation, and unequal agent influence.

4 Studies

Trustworthy AI

TRUST

Fairness as **justice, transparency, accountability, non-discrimination**, and oversight.

5 Studies

Measurement Categories

- **Benchmark Performance**
- **Group Disparity**
- **Interactional / MAS Behavioural Metrics**
- **Conceptual / Normative**

RQ2: What **biases or harms** appear across the SDLC when using these systems?

Representation → Outputs can **reinforce stereotypes**, marginalize protected groups, and **encode biased assumptions** into prompts, roles, or generated artifacts.

Quality of Service → MAS can **perform unevenly across groups or contexts**, producing less accurate, less useful, hallucinated, or **unsafe recommendations for some users**.

Security & Privacy → MAS can **introduce risks** such as prompt injection, cascading attacks, or unsafe interactions that **expose sensitive information or compromise behavior**.

Governance & Trust → MAS can create **opaque decision pathways**, unclear responsibility across agents and humans, weak oversight, and **uncertainty about accountability for unfair outcomes**.

RQ2: What **biases or harms** appear across the SDLC when using these systems?

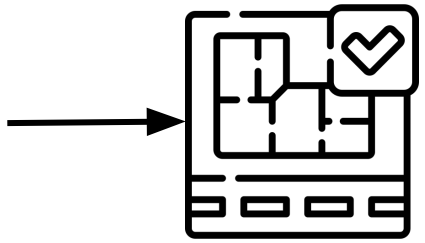
LOW (0-1) | MED (2-3) | HIGH (4-6)

Harm Type	Requirements	Design	Development	Testing	Maintenance
Representation	1	1	1	6	0
Quality of Service	3	3	2	6	2
Security & Privacy	1	2	1	3	2
Governance & Trust	4	2	2	6	2

RQ3: What **gaps remain** in making multi-agent systems for software engineering **fair, accountable, and transparent?**

Fragmented Evaluation

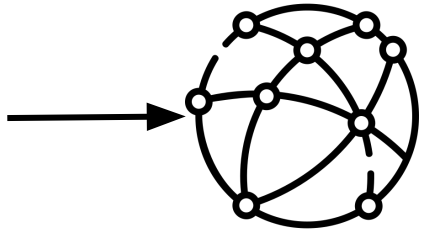
Ad-hoc tasks and metrics make studies difficult to compare.



MAS-Aware Evaluation Protocols

Limited Generalization

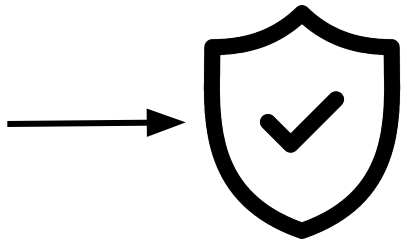
Results often depend on one model, protocol, task, language, or attribute.



Expanded Evaluations

Weak Mitigation

Safeguards are proposed more often than they are implemented and tested.



Validated Interventions

RQ1: How do existing studies **define and measure fairness** in multi-agent systems?

Fairness Definitions

Bias Reduction (BIAS)

Reducing stereotypes and demographic disparities across groups.

9 Studies

Inter-Agent Dynamics (MAS)

Fairness as a **property of MAS interaction**, conformity, amplification, responsibility allocation, and unequal agent influence.

4 Studies

Trustworthy AI (TRUST)

Fairness as **justice, transparency, accountability, non-discrimination**, and oversight.

5 Studies

Measurement Categories

- Benchmark Performance
- Group Disparity
- Interactional / MAS Behavioural Metrics
- Conceptual / Normative

RQ2: What **biases or harms** appear across the SDLC when using these systems?

LOW (0-1) | MED (2-3) | HIGH (4-6)

Harm Type	Requirements	Design	Development	Testing	Maintenance
Representation	1	1	1	6	0
Quality of Service	3	3	2	6	2
Security & Privacy	1	2	1	3	2
Governance & Trust	4	2	2	6	2

RQ1: Fairness is fragmented across bias, trust, and agent interactions.

RQ2: MAS harms span the SDLC, but evidence is uneven and weakly tied to real SE workflows.

RQ3: The field identifies risks, but lacks shared evaluation, generalizable evidence, and deployable mitigations.

RQ3: What **gaps remain** in making multi-agent systems for software engineering **fair, accountable, and transparent?**

Fragmented Evaluation

Ad-hoc tasks and metrics make studies difficult to compare.

→ **MAS-Aware Evaluation Protocols**

Limited Generalization

Results often depend on one model, protocol, task, language, or attribute.

→ **Expanded Evaluations**

Weak Mitigation

Safeguards are proposed more often than they are implemented and tested.

→ **Validated Interventions**