Systematic Review of Reliability and Trustworthiness on Multi-Agent Architectures

Michael da Silva[1], **Anderson Aparecido Alves da Silva**[1]

[1] Instituto de Pesquisas Tecnológicas do Estado de São Paulo (IPT)
São Paulo, São Paulo – Brasil

michael.silva@ensino.ipt.br,andersonalves@ipt.br

Abstract. *Multi-Agent Systems (MAS) have emerged as a promising approach for building adaptive, distributed defense strategies. This paper presents a Systematic Literature Review (SLR) aimed at identifying essential architectural aspects of reliability in MAS. The results are organized around: attacks and vulnerabilities in MAS, proposed mitigation solutions, methods for detecting behavioral anomalies and architectural characteristics of reliable and trustworthy systems. Finally, we summarize a set of technical approaches that can be applied into MAS to address the requirement, and we single out one of them as a candidate for future implementations in this context.*

Abstract. *Os Sistemas Multiagente (SMAs) emergem como uma abordagem promissora para a construção de estratégias de defesa adaptativas e distribuídas. Este artigo apresenta uma Revisão Sistemática da Literatura (RSL) com o objetivo de identificar aspectos arquiteturais essenciais de confiabilidade em SMAs. Os resultados foram organizados em torno das seguintes dimensões: ataques e vulnerabilidades em SMAs, soluções propostas para mitigação, métodos para identificar anomalias comportamentais e características arquiteturais de sistemas confiáveis e dignos de confiança. Por fim, resumindo um conjunto de abordagens que podem ser aplicadas em MAS para atender a este requisito, e destacando-se uma delas para futuras implementações nesse contexto.*

1. Introdução

A adoção de arquiteturas baseadas em inteligência distribuída tem impulsionado o interesse pelos Sistemas Multiagentes (SMAs) em domínios que exigem respostas adaptativas em tempo real, como plataformas de agentes baseadas em modelos de linguagem de grande escala [Ahmed et al. 2021]. Nesses contextos, múltiplos agentes autônomos colaboram por meio de trocas de mensagens para alcançar objetivos comuns [Kott 2023].

Em SMAs, cada agente possui capacidade autônoma de percepção, raciocínio e ação, de modo que o comportamento do sistema emerge das interações locais entre seus componentes. Isso implica que a confiabilidade não é uma propriedade estática, verificável pontualmente, mas um atributo dinâmico que depende do alinhamento comportamental dos agentes ao longo do tempo e sob condições adversas [Masters et al. 2025].

Motivado por desenvolver SMAs com confiabilidade na comunicação desde a concepção, sendo assim, o presente trabalho apresenta os resultados de uma Revisão

Sistemática da Literatura (RSL), com o objetivo de responder à seguinte questão de pesquisa: Quais são os aspectos arquiteturais essenciais de um sistema multiagente com comunicação confiável?

O presente artigo está estruturado da seguinte forma: a Seção 2 detalha o protocolo e a condução da RSL, bem como a análise dos resultados segundo as quatro questões de pesquisa derivadas; a Seção 3 apresenta as considerações finais.

2. Revisão Sistemática da Literatura

A revisão sistemática está dividida em três fases principais de acordo com [Felizardo et al. 2017] divide-se nas seguintes fases: Planejamento, Condução e Análise dos resultados.

2.1. Planejamento da Revisão Sistemática

A revisão será decomposta e organizada utilizando a estratégia PICO. A População: sistemas multiagentes; Intervenção: ataques na comunicação prejudiciais à confiança; Comparação: artigos utilizados na etapa exploratória; Resultados: características principais de confiabilidade.

A lista das **bases de busca:** Web of Science (WOS), IEEE, Association for Computing Machinery (ACM). Strings de Busca: *"agent AND reliable AND communication AND system AND attack"*. Compreendendo o período entre 2015 e 2026. Os seguintes **critérios de inclusão:** Descreve características de multiagentes; Propõe solução para mitigar ataques e vulnerabilidades em multiagentes; Versa sobre vulnerabilidades e ataques em multiagentes. Para os **critérios de exclusão:** Independência de fatores ambientais e da natureza; Não encontrado; Não versa sobre agentes na área de computação; Versa sobre agentes estritamente físicos ou segurança física;

A **estratégia para seleção dos trabalhos:** por meio da leitura dos metadados, os critérios são verificados. Os documentos que não forem excluídos nesta etapa seguem para a próxima fase, na qual são lidas a Introdução e a Conclusão.

A **estratégia de qualidade dos trabalhos:** Os documentos que não forem eliminados até esta etapa avançam para a fase de avaliação de qualidade. Cada resposta será classificada da seguinte forma: Em linha, com peso de 16,667 pontos; Parcialmente em linha, com peso de 8,445 pontos; Fora de linha, com peso de 0,0 pontos. Os documentos foram analisados com base nas seguintes perguntas:

- As vulnerabilidades e ameaças são descritas?
- As soluções e mitigações são bem descritas?
- Existem métodos quantitativos para detecção de mudança de comportamento dos agentes?
- Descreve as principais características de arquiteturas de multiagentes confiáveis?
- Existe o mapeamento de brechas relacionadas à comunicação entre agentes?
- Estabelece-se uma definição sobre confiabilidade em arquitetura multiagente?

A **estratégia de extração e síntese dos dados:** foram preenchidos formulários de extração de dados para cada documento válido, após sua leitura integral e resumo em função dos dados mais importantes extraídos:

- Quais são os principais ataques e vulnerabilidades em sistemas multiagentes?
- Quais são as soluções propostas para as vulnerabilidades?
- Quais são as formas de identificar anomalias e mudança de comportamento em multiagentes devido a ataques?
- Quais são os principais aspectos arquiteturais de sistemas multiagentes confiáveis?

A **estratégia para análise dos dados para publicação dos resultados:** após a leitura minuciosa e completa dos trabalhos qualificados, foi elaborado um resumo destacando os métodos ou técnicas como se inter-relacionam e também como respondem ou ajudam a responder à pergunta de pesquisa.

2.2. Condução

A condução da revisão sistemática foi realizada pelo autor. Todos os trabalhos retornados pelas buscas foram devidamente documentados na ferramenta de apoio Parsifal (https://parsif.al). **Identificação**: é aplicada a estratégia descrita previamente no planejamento. *Data de busca:* 7 de fevereiro de 2026. *Amplitude da busca:* Tópico: TS. *Tipos de documentos:* artigos de revista, congresso, de revisão e material editorial. *Período considerado:* 2015 até 2026.

- ACM (133 documentos). String executada: [Title: agent*] and [Abstract: *reliab*] and [Abstract: communic*] and [All: system*] and [All: attack]
- IEEE (37 documentos). String executada: (”Document Title”:agent*) and (”Abstract”:*reliab*) and (”Abstract”:communic*) AND (”All Metadata”:system*) and (”All Metadata”:attack)
- WOS (32 documentos). String executada: TI=(agent*) and AB=(reliab*) and ALL=(communic*) and ALL=(system) and ALL=(attack)

Seleção dos estudos: foi aplicada a estratégia de seleção, conforme detalhada na subseção de planejamento. Sendo 96 documentos selecionados ao total, outros 91 não selecionado e 15 marcados como duplicados. **Qualificação dos estudos:** os estudos responderam às perguntas conforme a subseção de planejamento. A pontuação mínima para prosseguir é de 67 pontos, haja vista que a qualidade dos artigos frente aos interesses da pesquisa decai quando a pontuação se apresenta inferior ao patamar, culminando em 33 qualificados e 63 desqualificados. **Extração e síntese dos dados:** para realizar a extração e a síntese dos dados, foi aplicada a estratégia detalhada no planejamento da revisão. Nessa etapa, 33 documentos tiveram seus dados extraídos integralmente.

2.3. Análise dos resultados

Esta subseção analisa as perguntas de pesquisa da revisão sistemática sob o prisma das principais características de confiabilidade em arquiteturas SMAs, respondendo, em particular, a cada pergunta de pesquisa. A Figura 1 sumariza as principais abordagens encontradas na literatura e destaca aquelas mais estritamente correlacionadas à estabilidade, na perspectiva do presente trabalho, quanto à confiabilidade.

2.3.1. Quais são os principais ataques e vulnerabilidades em sistemas multiagentes?

Sob a ótica do princípio da disponibilidade, devido às características distribuídas dessa abordagem arquitetural, os ataques de DDoS (diretos, indiretos ou

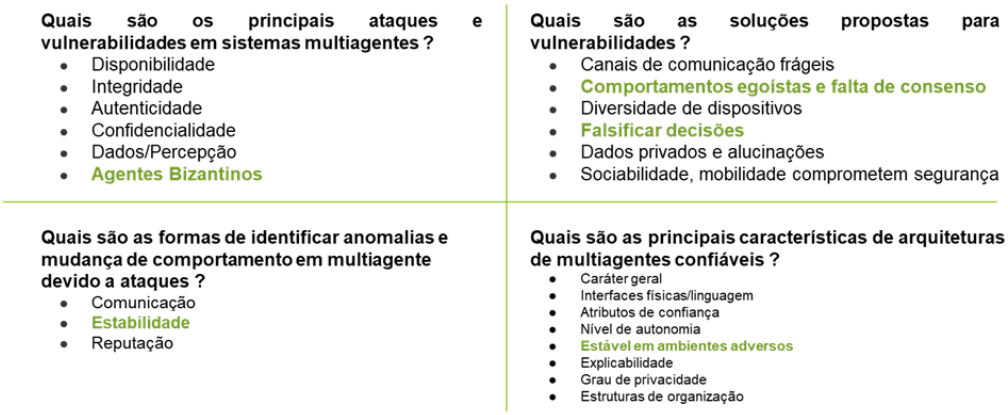

Figura 1. Perguntas e principais respostas

híbridos) [Wen et al. 2026] são amplamente referenciados [Baradarannia et al. 2025] [Lygizou and Kalles 2025] [Jain and Solomon Jebaraj 2023] [Deng et al. 2024] [Lin et al. 2025], seja pela volumetria de dados enviados [Yuan et al. 2024] ou sobrecarga dos agentes envolvidos [Hajar et al. 2023], provocando a negação de serviço propriamente dita [Li et al. 2020]. A segmentação da rede podem ser alvos na medida em que as áreas de abrangência de cada agente podem ser afetadas [Pretila et al. 2024], ocasionando indisponibilidade [Hao et al. 2025].

Existem ataques nos quais pacotes de comunicação são manipulados: frequências, duração, multiplicação ou substituição dos pacotes de informação [He et al. 2022], como *man-in-the-middle* com repetição de pacotes [Marques et al. 2021]. Esses ataques podem ocorrer diretamente contra o agente ou contra os seus sensores [Li et al. 2020]. A adulteração de descarte indevido de informações [Hajar et al. 2023] ou interrupções no canal [Pretila et al. 2024], o que reflete em dados incorretamente transmitidos no canal de comunicação [Gao and Yang 2024] até mesmo com adição de ruídos [Hao et al. 2025].

Outros ataques violam a autenticidade no controle de acesso inicial [Yuan et al. 2024] e na autenticação [Çımen et al. 2025], utilizando técnicas de obtenção de credenciais [Choi et al. 2020]. O controle de acesso pós-autenticação também é afetado por mudanças abruptas de requisições de recursos [Hajar et al. 2023], refletindo-se em casos de roubo de uma identidade de maneira indevida [Marques et al. 2021]. A ausência ou a baixa aderência a práticas de autorização dos agentes [Alruqi et al. 2021] pode permitir a adulteração de configurações indevidamente, sem a devida rastreabilidade, além de possibilitar o abuso do uso de recursos [Zhang et al. 2025c].

Em topologias de agentes baseadas em consenso mútuo, pode ocorrer pela clonagem indevida de agentes, subvertendo o balanço do consenso do sistema [Lin et al. 2023]. Uma vez que outros princípios de segurança sejam violados, a confidencialidade do sistema também fica comprometida [Jain and Solomon Jebaraj 2023]. Ataques de exfiltração de informações e configurações são relatados [Yuan et al. 2024] devido a fragilidades na criptografia, bem como o vazamento de informações por brechas na segurança [He et al. 2022], infringindo e abusando da privacidade dos usuários e da plataforma [Toumi et al. 2017] ou pela injeção de *prompts* [Zhang et al. 2025a].

Alguns invasores buscam afetar diretamente as fontes de informação, manipu-

lando dados que interfiram no comportamento final [Jain and Solomon Jebaraj 2023], mas também modificando parâmetros para envenenar os modelos [Esmaeili et al. 2024]. Existem relatos frequentes de ingestão de dados erráticos ou falsos [Zhang et al. 2025b] [Huang and Dong 2020] diretamente nos agentes, ou indiretamente através dos sensores e perceptores [Li et al. 2020], bem como nos atuadores e ações [Hao et al. 2025]. Estes são comumente chamados de ataques de *deception* [He et al. 2022], pois realizam a falsificação ou adição indevida de informações. Nesse sentido, em relação à percepção existem ataques por supressão e visão parcial [Xue et al. 2022], que geram uma compreensão incompleta ou distorcida da memória de contexto [Zhang et al. 2025a].

Em topologias SMAs de forte consenso, ataques de transações conflitantes provocando sobrecarga [Lin et al. 2023]. Alguns autores apontam ataques violando vários princípios de segurança [Mao et al. 2025]. Devido à interdependência, a falha ou o comprometimento dos agentes poder ocasionar o comprometimento do todo, seja por: disponibilização de informações falsas [Lu et al. 2024], recomendações e compartilhamento de informações, mudança abrupta de comportamento [Lygizou and Kalles 2025] ou por adulteração da comunicação entre agentes [Wen et al. 2026].

Dessa forma, esses ataques interferem na confiabilidade entre os agentes, quando um agente sinaliza positivamente o alinhamento [Zhang et al. 2025a], mas realiza ações opostas ao alinhamento de objetivos projetados [Phiri 2025]. Além disso, ao ocultar determinadas ações [Xue et al. 2022], a rede pode ficar tomada por agentes maliciosos [Li et al. 2020], que podem conspirar construindo planos conflituosos para degradar a operação [Masters et al. 2025]. Para além disso, a adição ou exclusão de agentes [Barambones et al. 2021], sejam modificados ou clonados indevidamente, pode descoordenar os mecanismos de tomada de decisão do sistema [Jain and Solomon Jebaraj 2023].

Outros ataques envolvem interação humana para a elevação de privilégios [Zhang et al. 2025b] e a manipulação indireta da janela de contexto [Zhang et al. 2025c]. Quanto a SMAs que possuem atuadores físicos, o comprometimento desses componentes pode impactar a tomada de decisão [Lin et al. 2025]; em alguns casos a falta de visão geral do ambiente de forma consistente é prejudicial [Phiri 2025]. Na análise exploratória de vulnerabilidades relatos sobre o uso de ataques adversariais [Esmaeili et al. 2024] contra a comunicação para mapeamento de brechas [Çımen et al. 2025], bem como a utilização de canais laterais para a inferência de comportamento [Damle et al. 2022]. A exploração de configurações indevidas quando outras camadas de segurança são violadas [de la Hoz et al. 2017] configura também uma grave ameaça para essa arquitetura, podem ser empregadas em ameaças persistentes [Marques et al. 2021].

2.3.2. Quais são as soluções propostas para vulnerabilidades?

Os canais não íntegros de comunicação possuem vulnerabilidades relativas a gargalos e atrasos que podem ser exploradas [Gao and Yang 2024]. Estas, por sua vez, podem estar relacionadas a topologias centralizadas [Yuan et al. 2024] ou à falta de controle centralizado para verificação de resultados e operações [Phiri 2025], o que pode abrir espaço para pontos únicos de falha na topologia [Alruqi et al. 2021], possibilitando a saturação dos canais [Deng et al. 2024] ou a criação de malhas complexas de serem operacionalmente roteadas, ainda mais em situações de redundância [Pretila et al. 2024]. Nessa linha, os

agentes tornam-se fragilizados quanto a variações de latência da rede [Zhang et al. 2025b] e à falta de mecanismos de controle de tráfego [Hao et al. 2025], mais notoriamente em protocolos industriais [Choi et al. 2020].

Dado o caráter distribuído do SMAs, são exploradas vulnerabilidades que impactam a integridade e a autenticidade quanto aos objetivos e ao alinhamento dos agentes [Li et al. 2020], como a presença de agentes egoístas [Hajar et al. 2023], refletindo diretamente na falta de consenso [He et al. 2022] e no entrave na tomada de decisão ou obtenção de informação [de la Hoz et al. 2017]. Em outras palavras, topologias muito descentralizadas [Baradarannia et al. 2025] também apresentam problemáticas para a segurança. Por outro lado, o isolamento na tomada de decisão demonstra-se como uma fraqueza [Barambones et al. 2021]. Em outro extremo, a forte dependência entre os agentes pode representar uma fragilidade [Alruqi et al. 2021], sendo essas dependências externas ou globais interagentes [Huang and Dong 2020]. Por fim, a falta de formas de lidar com medições parciais ou a forte dependência de dados históricos debilita a arquitetura [Masters et al. 2025].

A diversidade dos agentes e dos dispositivos abre espaço para incompatibilidades na interoperabilidade [Deng et al. 2024] que podem afetar a disponibilidade. A própria limitação de recursos propicia riscos aos agentes [Lygizou and Kalles 2025]. A baixa robustez da proteção dos dispositivos [Esmaeili et al. 2024] expostos em ambientes hostis e intempéries pode afetar diretamente a tomada de decisão [Lin et al. 2025]. Isso significa que conexões e dispositivos tornam-se passíveis de exploração [Lu et al. 2024] - vide a alta mobilidade [Alruqi et al. 2021] dos agentes em rodovias de alta velocidade onde decisões relevantes precisam ser tomadas [Wen et al. 2026].

A necessidade de consenso e estabilidade correta na tomada de decisão pode ser explorada, justamente alterando a postura dos agentes levando: conluio, omissão ou falsificação pelos agentes [Phiri 2025]. Estes também são conhecidos como agentes bizantinos [Mao et al. 2025] que podem colaborar entre si ou individualmente para fraudar votações [Lin et al. 2023], o que pode acarretar a prática deliberada de injustiça na tomada de decisão [Esmaeili et al. 2024].

As brechas referentes à autenticidade e na irretratabilidade estão associadas à falta de gestão de acessos [Toumi et al. 2017], podendo contar com *rootkits* para exploração. Citam-se os ataques de *prompt* amplamente divulgados que se aproveitam das alucinações [Zhang et al. 2025c] das LLM para expor informações ou dados pessoais por meio do ultra detalhamento [Choi et al. 2020]. Nessa linha, como a IA tem sido cada vez mais empregada nos agentes, herda-se uma série de ataques relacionados a essa tecnologia [Xue et al. 2022]. Estes, por sua vez, demandam explicabilidade, sob o risco de ocorrência de vazamentos de preferência ou de utilidade [Damle et al. 2022], os quais, por meio de inferência indireta, podem violar a privacidade [Yuan et al. 2024].

2.3.3. Quais são as formas de identificar anomalias e mudança de comportamento em multiagente devido a ataques?

Na comunicação pode-se avaliar a quantidade de *hops* e redundâncias ativadas na rede [Gao and Yang 2024] e o rastreamento em profundidade da rede [de la Hoz et al. 2017].

Além disso, analisa-se a frequência e a duração da comunicação, identificando comportamentos de Zeno [Deng et al. 2024], observando-se também comportamentos que se interrelacionam com a comunicação [Lin et al. 2025]. Assim a observação de sinais de interferência e da variação média da comunicação é relevante [He et al. 2022].

De maneira mais indireta, realiza-se a observação lateral dos agentes vizinhos cuja comunicação apresente perda de pacotes [Hajar et al. 2023] e sobre as mensagens transmitidas [Zhang et al. 2025b]. Adicionalmente, o acompanhamento do tráfego de saída permite a detecção de ataques de exfiltração, monitorando volume, tempo, origem e destino [Marques et al. 2021]. Nessa linha, a observação do atraso ao longo da comunicação [Pretila et al. 2024] ou o envelhecimento das mensagens [Choi et al. 2020] fornecem indícios de anomalias. Por outro lado, analisando a comunicação segundo a teoria do controle *Uniformly Ultimately Bounded* para remontagem ou análise assintótica das comunicações [Huang and Dong 2020].

Quanto ao estabelecimento de limites a distribuição de Bernoulli, que facilita o entendimento dos intervalos de operação dos ataques e intermitências [Baradarannia et al. 2025]. Na perspectiva da estabilidade dos sistemas, diferentes trabalhos apontam para o método de Lyapunov [Huang and Dong 2020] [Baradarannia et al. 2025], mensurando o distanciamento do comportamento normal esperado ao longo do tempo. De modo similar a medição desvio da distribuição de referência Kolmogorov-Smirnov [Lygizou and Kalles 2025]. Avalia-se também a correlação $k{-}out{-}of{-}n$ entre o estímulo inicial e a ação realizada [Lu et al. 2024], como alucinações e coordenações incorretas [Mao et al. 2025], e o monitoramento estatístico das tarefas em execução [Zhang et al. 2025b]. Podem ser avaliadas, ainda, a entropia da informação e a função de utilidade em alguns casos [Damle et al. 2022], além de limites fixos estabelecidos para o monitoramento de gerações residuais [Li et al. 2020] e limiares de erro [Lin et al. 2025].

Quanto à reputação, o monitoramento das votações e a análise da convergência do consenso [Gao and Yang 2024] se demonstra fundamental, primordialmente em topologias de enxame [Lu et al. 2024]. Além disso, destaca-se a indicação de abusos frequentes de limites [Jain and Solomon Jebaraj 2023], acompanhada pela análise de ruído gaussiano. Também há análise ativa de: falsos positivos, assinaturas, violações de postura e de regras [Toumi et al. 2017], aplicando-a diretamente aos líderes de equipe dos agentes e aos seus subordinados [Esmaeili et al. 2024]. Consequentemente, efetua-se o acompanhamento da estabilidade de entrada [Deng et al. 2024], bem como dos passos intermediários tomados pelos subagentes [Zhang et al. 2025c]. De maneira complementar, a análise de arquivos de logs e auditorias [Phiri 2025], a inspeção de controles de conformidade, gargalos e metas definidas [Masters et al. 2025] auxiliam na verificação da integridade dos componentes e dos contratos estabelecidos [Alruqi et al. 2021].

Em soluções pautadas em votação ou consenso, a pontuação maliciosa ou cálculos de reputação incorretos podem ser prejudicados [Wen et al. 2026]. Isso possui relação direta com propriedades de *agreement and consistency* [Mao et al. 2025] as quais são apontadas como de necessária observação. Ademais, quando os agentes são baseados em aprendizado por reforço, as cadeias de Markov mostram-se como abordagem relevante [Lygizou and Kalles 2025], considerando fatores como o tempo de confirmação de consenso, a duplicidade de gastos e a fila de atividades pendentes [Lin et al. 2023].

2.3.4. Quais são as principais características de arquiteturas de multiagentes confiáveis?

Os sistemas devem ser, por concepção, descentralizados [Huang and Dong 2020], com tomada de decisão distribuída e modularizada[Esmaeili et al. 2024], apresentando sociabilidade, autonomia e proatividade [Jain and Solomon Jebaraj 2023], bem como uma separação claramente definida entre as camadas de rede, consenso, contrato e aplicação [Lu et al. 2024]. A comunicação pode ser assíncrona, com heterogeneidade entre os agentes constituintes [Lin et al. 2023]. Os sistemas, além de manterem um certo grau de autonomia, devem incluir mecanismos de planejamento, ação, raciocínio e uso de memória [Zhang et al. 2025c]. Além disso, é fundamental que as aplicações sejam escaláveis [He et al. 2022], recorrendo a *containers* [de la Hoz et al. 2017], podendo ainda operar em um modelo de *multi-tenant* [Toumi et al. 2017].

Os agentes podem utilizar LLM para realizar tarefas de inferência textual em combinação com RAG e *chain-of-thought* para resultados mais detalhados [Zhang et al. 2025c]. Adicionalmente, deve haver colaboração e integração entre o meio físico e o digital dos agentes [Choi et al. 2020], incluindo suporte a ambientes 3D e *digital twins* aptos a operar em modo *off-load* [Wen et al. 2026]. Dessa forma, provê-se tolerância a falhas ciberfísicas [Hao et al. 2025], além de se demandar robustez dos hardwares de transmissão [Lin et al. 2025]. Deve-se, suportar a mobilidade dos agentes e a avaliação de confiança entre eles [Lygizou and Kalles 2025], com análise de competência e confiança por meio de *Tit-3-for-tat*.

Nesse cenário, a confiança também precisa considerar uma janela deslizante de eventos e evidências [Lygizou and Kalles 2025], ou seja, um mecanismo de gerenciamento de confiança que envolva a aplicação de punições, definição de rotas seguras e proteção de dados e recursos [Hajar et al. 2023]. Esse contexto viabiliza o uso de políticas ajustadas aos comportamentos dos agentes, o que confere certo grau de previsibilidade [Wen et al. 2026], ou ainda de protocolos de consenso com múltiplas rodadas [Mao et al. 2025], bem como de compensadores de consenso [Deng et al. 2024]. Adicionalmente, é possível empregar estratégias de propagação de crenças, com coordenação heurística orientada ao contexto [de la Hoz et al. 2017], em alguns casos impondo consenso por dominância de grupo [Li et al. 2020].

Os agentes devem apresentar certo grau de subjetividade e consciência de contexto [Lygizou and Kalles 2025], além de garantir interoperabilidade com protocolos como ACP e A2A, incorporando mecanismos de aprendizado adaptativo, por exemplo com aprendizado federado [Çımen et al. 2025]. Devem ser autoadaptáveis a partir de treinamentos com dados benignos e maliciosos com centralização [Xue et al. 2022], levando em conta dados tanto atuais quanto históricos [Esmaeili et al. 2024], utilizar protocolos padronizados como MCP [Zhang et al. 2025b] e oferecer descoberta dinâmica de recursos e autodetecção [Toumi et al. 2017].

Em síntese, isso exige o atendimento a certos critérios de conformidade *soft* ou *hard*, com um sistema estrategicamente controlável e sujeito a intervenção humana [Masters et al. 2025], além de atributos consagrados de *compliance*, como auditabilidade, responsabilidade legal, protocolos de testemunha e governança da plataforma [Phiri 2025], vários trabalhos ainda apontam para a utilização de blockchain

[Alruqi et al. 2021], bem como de privacidade diferencial [Zhang et al. 2025a]. A habilidade de manter o funcionamento em condições adversas e ainda assim recuperar o desempenho [Barambones et al. 2021], bem como aplicar filtragem de mensagens [Xue et al. 2022] para evitar problemas decorrentes de erros de quantização [Huang and Dong 2020], torna plausível definir *triggers* de acionamento da comunicação [Deng et al. 2024].

A comunicação em ao menos duas etapas e a capacidade de identificar a variabilidade entre períodos de demanda [Hao et al. 2025], associadas à detectabilidade do comportamento dos vizinhos [Gao and Yang 2024], devem ser consideradas em comunicação *broadcast* [Mao et al. 2025] quanto em redes *ad-hoc*, incorporando segmentação e redundância [Pretila et al. 2024]. Estratégias de filtragem de mensagens inspiradas em mecanismos biológicos, permitindo o reconhecimento de padrões e a autorregulação [Marques et al. 2021]. Adicionalmente, é relevante empregar detectores cooperativos e *hard-limits* para a detecção de anomalias [Li et al. 2020], sem negligenciar a necessidade contínua de criptografia [Yuan et al. 2024].

A coordenação característica entre os agentes [Çımen et al. 2025], aliada à possibilidade de estabelecer formações temporárias para ampliar a dinamicidade da comunicação, dificultando a previsibilidade para agressores [Baradarannia et al. 2025], bem como de realizar segmentação hierárquica quando necessário [Choi et al. 2020], fazendo uso de normalizadores para distribuir de forma mais equilibrada a carga [Hao et al. 2025]. Torna-se possível uma organização holônica e autossimilar dos agentes, com interações em níveis local e global [Esmaeili et al. 2024], visando preservar a auto-organização entre os agentes, a ocultação ou camuflagem para evitar vazamento de dados ou perdas, além de formarem coalizões e atuarem sobre domínios específicos do problema [Zhang et al. 2025b].

O uso em regime de operação contínua com velocidade constante e ênfase na eficiência operacional [Baradarannia et al. 2025], combinado com a adoção de políticas de gestão de credenciais e a realização de varreduras sistemáticas por vulnerabilidades [Choi et al. 2020], com a utilização de listas restritivas [Alruqi et al. 2021].Da mesma forma, o estabelecimento de conexões entre agentes [Damle et al. 2022] possibilita a adoção de uma postura de segurança proativa em tempo real [Wen et al. 2026], fortalecida por extensas baterias de testes adversariais [Çımen et al. 2025]. Adicionalmente, aplicam-se a introdução de atraso artificial e a coleta aperiódica de logs e métricas [Lin et al. 2025], levando em conta os custos de comunicação [Yuan et al. 2024].

3. Considerações Finais

A partir da análise, foi possível consolidar um panorama abrangente sobre confiabilidade respondendo às quatro questões de pesquisa propostas. No que diz respeito aos ataques e vulnerabilidades, ficou evidenciado que os SMAs herdam ameaças reconhecidas em sistemas distribuídos, como ataques de negação de serviço, injeção de dados falsos e afins, porém apresentam uma superfície de ataque adicional e característica: as arestas de confiança entre os agentes. O Problema dos Generais Bizantinos sintetiza bem essa condição, na qual agentes comprometidos podem atuar de forma coordenada para subverter os mecanismos de decisão coletiva sem que o sistema detecte a anomalia de forma imediata.

A literatura converge para abordagens que combinam descentralização controlada, mecanismos de consenso com múltiplas rodadas, privacidade diferencial, criptografia e rastreabilidade. Sendo assim, a revisão denotou que as abordagens mais promissoras combinam análise da comunicação, avaliação de reputação e estabilidade comportamental. Nesse conjunto, destaca-se o método de Lyapunov como um dos instrumentos analíticos recorrentes e tecnicamente consistentes, sendo empregado para verificar formalmente se o comportamento do sistema permanece estável.

A análise de Lyapunov permite derivar condições suficientes de estabilidade para sistemas multiagentes não lineares e heterogêneos, oferecendo uma medida contínua do distanciamento em relação ao comportamento esperado. Assim a confiança na comunicação, portanto, não deve ser tratada como um estado binário, mas como um atributo dinâmico, continuamente recalibrado com base no histórico de interações e no contexto operacional de cada agente.

3.1. Limitações e ameaças a validade

O trabalho emprega estritamente três bases de buscas, ou seja, poderiam ser adicionados mais bases na revisão. Além disso não foram considerados linguagens ou frameworks desenvolvimento dos agentes, algo relevante mapeamento das principais tendências de uso.

3.2. Trabalhos Futuros

Em trabalhos futuros, sobretudo em cenários nos quais a confiabilidade das mensagens trocadas é determinante para a qualidade das decisões tomadas de forma distribuída, pode-se aplicar o método de Lyapunov de modo experimental para mensurar a estabilidade sistêmica, portanto implementando um mecanismo para a detecção precoce de desvios comportamentais provocados por agentes maliciosos ou comprometidos.

Referências

Ahmed, M., Ansar, K., Muckley, C. B., Khan, A., Anjum, A., and Talha, M. (2021). A semantic rule based digital fraud detection. *PeerJ Computer Science*, 7:e649. Modelo ontológico para detecção e alerta de fraudes digitais financeiras.

Alruqi, M., Hsairi, L., and Eshmawi, A. (2021). Secure mobile agents for patient status telemonitoring using blockchain. page 224–228.

Baradarannia, M., Hashemzadeh, F., and Kumbasar, T. (2025). Fixed-time surrounding control of multi-agent systems underdenial-of-service attacks. *JOURNAL OF THE FRANKLIN INSTITUTE*, 362.

Barambones, J., Richoux, F., Imbert, R., and Inoue, K. (2021). Resilient Team Formation with Stabilisability of Agent Networks for Task Allocation. *ACM Trans. Auton. Adapt. Syst.*, 15.

Choi, I.-S., Hong, J., and Kim, T.-W. (2020). Multi-Agent Based Cyber Attack Detection and Mitigation for DistributionAutomation System. *IEEE ACCESS*, 8:183495–183504.

Damle, S., Triastcyn, A., Faltings, B., and Gujar, S. (2022). Differentially Private Multi-Agent Constraint Optimization. page 422–429.

de la Hoz, E., Gimenez-Guzman, J. M., Marsa-Maestre, I., Cruz-Piris, L., and Orden, D. (2017). A Distributed, Multi-Agent Approach to Reactive Network Resilience. page 1044–1053.

Deng, F., Xu, C., Guan, Z.-H., and Chen, G. (2024). Dynamic event-triggered quantized secure output consensus ofheterogeneous multi-agent systems under DoS attacks. *CHAOS*, 34.

Esmaeili, A., Ghorrati, Z., and Matson, E. T. (2024). Holonic Learning: A Flexible Agent-based Distributed Machine Learning Framework. page 525–533.

Felizardo, K. et al. (2017). *Revisão sistemática da literatura em engenharia de software: teoria e prática*. Elsevier Brasil.

Gao, R. and Yang, G.-H. (2024). Trust-Based Distributed Secure State Estimation Against Malicious Agentsvia Two-Hop Communication. *IEEE TRANSACTIONS ON AUTOMATIC CONTROL*, 69:1006–1013.

Hajar, M. S., Kalutarage, H. K., and Al-Kadri, M. (2023). 3R: A reliable multi agent reinforcement learning based routing protocolfor wireless medical sensor networks. *COMPUTER NETWORKS*, 237.

Hao, Q., Luo, M., Cheng, J., and Shi, K. (2025). Distributed fault-tolerant consensus for two-time-scale multiagentsystems against multiple faults and random attacks via a generalizedtwo-step transmission mechanism. *JOURNAL OF THE FRANKLIN INSTITUTE*, 362.

He, W., Xu, W., Ge, X., Han, Q.-L., Du, W., and Qian, F. (2022). Secure Control of Multiagent Systems Against Malicious Attacks: A BriefSurvey. *IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS*, 18:3595–3608.

Huang, X. and Dong, J. (2020). Reliable Leader-to-Follower Formation Control of Multiagent SystemsUnder Communication Quantization and Attacks. *IEEE TRANSACTIONS ON SYSTEMS MAN CYBERNETICS-SYSTEMS*, 50:89–99.

Jain, G. and Solomon Jebaraj, N. R. (2023). Taxonomy of Multi-Agent Systems Attacks and their Defense Tactics in Certifying Security of Cyber Physical Systems. pages 648–652.

Kott, A. (2023). *Autonomous Intelligent Cyber Defense Agent (AICA): A Comprehensive Guide*, volume 87 of *Advances in Information Security*. Springer.

Li, Y., Fang, H., and Chen, J. (2020). Anomaly Detection and Identification for Multiagent Systems Subjected toPhysical Faults and Cyberattacks. *IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS*, 67:9724–9733.

Lin, B.-Y., Dziubałtowska, D., Macek, P., Penzkofer, A., and Müller, S. (2023). TangleSim: An Agent-based, Modular Simulator for DAG-based Distributed Ledger Technologies. pages 1–5.

Lin, H., Dong, J., and Park, J. H. (2025). An Artificial-Delay-Based Looped Functional for Dynamic Event-Triggered Fault-Tolerant Control of T-S Fuzzy Multi-Agent Systems. *IEEE Transactions on Automation Science and Engineering*, 22:6050–6060.

Lu, C., Zhu, J., and Ouyang, L. (2024). Mission Reliability and Optimization in Multi-Agent Systems Using Hyperledger Fabric Platform. pages 1–6.

Lygizou, Z. and Kalles, D. (2025). A Biologically Inspired Trust Model for Open Multi-Agent Systems That IsResilient to Rapid Performance Fluctuations. *APPLIED SCIENCES-BASEL*, 15.

Mao, Y., Kang, Y., Li, P., Zhang, N., Xu, W., and Zhang, C. (2025). IBGP: Imperfect Byzantine Generals Problem for Zero-Shot Robustness in Communicative Multi-Agent Systems. page 2654–2656.

Marques, R. S., Epiphaniou, G., Al-Khateeb, H., Maple, C., Hammoudeh, M., De Castro, P. A. L., Dehghantanha, A., and Choo, K. K. R. (2021). A Flow-based Multi-agent Data Exfiltration Detection Architecture for Ultra-low Latency Networks. *ACM Trans. Internet Technol.*, 21.

Masters, C., Vellanki, A., Shangguan, J., Kultys, B., Gilmore, J., Moore, A., and Albrecht, S. (2025). Orchestrating Human-AI Teams: The Manager Agent as aUnifying Research Challenge. page 91–107.

Phiri, C. C. (2025). Creating Characteristically Auditable Agentic AI Systems. page 1–14.

Pretila, H., Campbell, B., and Szabo, C. (2024). Contested Communication in C2 Multi-agent Simulations. page 25–29.

Toumi, H., Marzak, B., Khazri, Y., Talea, A., Eddaoui, A., and Talea, M. (2017). Mobiles Agents and Virtual Firewall to Secure the Shared Network for Virtual Machines in IaaS cloud.

Wen, X., Wen, J., Xiao, M., Kang, J., Zhang, T., Li, X., Chen, C., and Niyato, D. (2026). Defending Against Network Attacks for Secure AI Agent Migration in Vehicular Metaverses. *IEEE Internet of Things Journal*, 13:4153–4166.

Xue, W., Qiu, W., An, B., Rabinovich, Z., Obraztsova, S., and Yeo, C. K. (2022). Mis-spoke or mis-lead: Achieving Robustness in Multi-Agent Communicative Reinforcement Learning. page 1418–1426.

Yuan, T., Chung, H.-M., and Fu, X. (2024). PP-MARL: Efficient Privacy-Preserving Multi-Agent Reinforcement Learning for Cooperative Intelligence in Communications. *IEEE Network*, 38:196–203.

Zhang, S., Ma, Y., Chen, J., Li, S., Yi, X., and Li, H. (2025a). Towards Aligning Personalized AI Agents with Users' Privacy Preference. page 33–42.

Zhang, X., Dong, X., Wang, Y., Zhang, D., and Cao, F. (2025b). A Survey of Multi-AI Agent Collaboration: Theories, Technologies and Applications. page 1875–1881.

Zhang, Z., Yao, Y., Zhang, A., Tang, X., Ma, X., He, Z., Wang, Y., Gerstein, M., Wang, R., Liu, G., and Zhao, H. (2025c). Igniting Language Intelligence: The Hitchhiker's Guide from Chain-of-Thought Reasoning to Language Agents. *ACM Comput. Surv.*, 57.

Çımen, S., Karahan, S. N., Karhan, D., Güllü, M., and Osmanca, M. S. (2025). The Integration of Agentic AI in 6G Wireless Networks: State-of-the-Art, Challenges, and Future Perspectives. pages 1–6.