Investigação de Modelos Organizacionais para a Criação de Máquinas Éticas

Tielle da Silva Alexandre [1], **Carlos Eduardo Pantoja** [2], **Flávia Cristina Bernadini** [1]

[1]Instituto de Computação - Universidade Federal Fluminense - Niterói, RJ – Brasil

[2]Centro Federal de Educação Tecnológica (Cefet/RJ) Rio de Janeiro, RJ – Brasil

tiellesa@id.uff.br

Abstract. *A Inteligência Artificial desempenha um papel importante na transformação de cidades em ambientes inteligentes, possibilitando o desenvolvimento de novas tecnologias integradas ou substitutivas, como veículos autônomos, embora levante questões éticas e de responsabilidade. Os Sistemas Multiagentes consistem em agentes autônomos que cooperam e competem para resolver problemas coletivamente, com objetivos individuais e coletivos, semelhante à organização social humana. Máquinas Éticas, que seguem princípios e normas éticas, podem ser implementadas usando SMA, já que esses sistemas são formados por agentes que podem estar sob a mesma sociedade compartilhando regras e normas em prol da sociedade em si. Esses sistemas, com sua estrutura social análoga à humana, são candidatos para a criação e avaliação dessas máquinas éticas. O objetivo deste trabalho é investigar modelos organizacionais para criação máquinas éticas usando SMA embarcados.*

1. Introdução

A Internet das Coisas (IoT) é uma rede que conecta objetos na Internet, possibilitando o compartilhamento de dados e controle dos dispositivos [Management Association 2017]. O avanço da IoT permitirá a transformação das cidades em ambientes inteligentes dotados de soluções tecnológicas que melhoram a qualidade de vida de seus habitantes. Segundo [Albino et al. 2015], uma cidade inteligente poderia ser definida como uma cidade que fornece tecnologias integradas aos seus cidadãos para que tenham qualidade de vida. Por exemplo, uma cidade inteligente pode ser equipada com uma rede que conecta dispositivos de IoT e permite a troca de informações entre eles. Considerando a interconectividade de tais dispositivos, outras soluções poderão ser integradas às cidades inteligentes, onde os semáforos poderão ser autônomos priorizando a passagem do transporte público e dos pedestres, assim como um sistema de segurança inteligente poderá monitorar a cidade vinte e quatro horas por dia, acionando os serviços de segurança ou emergência quando um incidente for detectado [Boden 1996, Nalini 2019].

Nesse cenário, a Inteligência Artificial (IA) pode desempenhar um papel importante na transformação das cidades, possibilitando o desenvolvimento de novas tecnologias e soluções que podem ser integradas às soluções existentes ou até mesmo substituí-las. A presença de veículos autônomos no trânsito já é uma realidade em diversas cidades,

como em São Francisco, na Califórnia [Allen et al. 2006]. No entanto, questões sobre responsabilidade em casos de acidentes e outros dilemas, como a possibilidade de os veículos autônomos violarem as regras de trânsito para evitar um acidente, ainda estão em aberto [Metz 2019, Paulo 2023]. Isso ressalta a importância de que o desenvolvimento de soluções com IA seja acompanhado por discussões contínuas sobre ética e moralidade, garantindo que essas tecnologias sejam implementadas de forma responsável.

Concomitantemente, nos últimos 70 anos, a IA vem sendo desenvolvida acadêmica e industrialmente, tendo diversas definições ao longo da história. Por exemplo, em sistemas de IA, agentes podem ser projetados para realizar tarefas como reconhecimento de padrões, tomada de decisões, processamento de linguagem natural, controle de sistemas autônomos, entre outras. Neste trabalho, adotamos o grupo de definições de agentes inteligentes que consideram que todo sistema computacional e inteligente pode ser visto como um agente inteligente [Russell and Norvig 2016] e que um agente inteligente é uma entidade racional lógica ou física dotado de autonomia e proatividade que está situado em um ambiente ao qual interage e modifica a fim de atingir um objetivo [Weiss 1999].

Em um Sistema Multiagentes (SMA), agentes podem estar organizados em grupos de agentes autônomos que cooperam e competem para a resolução de problemas que não podem ser resolvidos de forma isolada [Briot and Demazeau 2001]. Assim, todo agente possui objetivos individuais e também objetivos coletivos pertencentes à sua sociedade. Dessa forma, uma organização estabelece restrições aos comportamentos dos agentes estabelecendo um comportamento grupal coeso. Tais organizações de agentes podem ser vistas como uma organização societal de agentes, que pode ser comparada à organização social dos seres humanos, que é estruturada e guiada por leis, normas e princípios éticos e morais. Por exemplo, uma universidade possui grupos de pessoas com diferentes papéis (e.g., professor, aluno e secretários) e com diferentes objetivos pessoais e profissionais, que podem ou não estar alinhados aos objetivos organizacionais da universidade. Tais grupos devem executar suas atividades na organização (universidade) respeitando as normas e o código de ética da sociedade na qual a organização está inserida,

Além disso, a organização também pode especificar um código de ética específico para sua comunidade (e.g., código de ética do servidor ou do aluno). A universidade cumpre seu propósito a partir do comportamento coeso de grupos de pessoas, cujos comportamentos individuais são regidos por um conjunto de regras. Segundo [Nalini 2019], a ética pode ser definida como um sistema de princípios, regras ou diretrizes que regem os comportamentos ou ações de um indivíduo ou grupo de indivíduos. Já o campo que estuda os valores e princípios éticos ou morais para que uma IA se comporte de maneira ética pode ser denominado de IA Ética ou Máquina Ética [Allen et al. 2006]. Considerando a analogia dos SMAs com a organização social humana, estes se tornam candidatos para implementação e avaliação de Máquinas Éticas. O objetivo deste trabalho é discutir uma proposta de investigação de modelos organizacionais para criação de máquinas éticas utilizando SMA. Este trabalho está dividido da seguinte forma: na Seção 2 é vista a metodologia do trabalho; e por fim, na Seção 3 serão apresentada os resultados esperados.

2. Metodologia

Na literatura, há diversos trabalhos que abordam a Máquina Ética em SMAs e o foco deles está na interação entre os agentes do sistema e sua organização. Atualmente, há duas abor-

dagens: centrada nos agentes e centrada na organização [Lemaître and Excelente 1998]. Na primeira abordagem, o SMA não possui uma representação explícita de sua organização. Espera-se ainda que, por meio da interação entre os agentes e o ambiente, possa emergir comportamentos complexos (auto-organização). Por exemplo, no sistema MANTA é simulada uma colônia de formigas e observou-se que o formigueiro apresentou estratégias de controle populacional e mecanismos de divisão do trabalho mesmo que tais comportamentos não tenham sido intencionalmente programados no código das formigas [Gilbert and Conte 2006]. Nessa abordagem, a estrutura organizacional existe apenas na memória dos agentes e a falta de uma descrição explícita da organização pode dificultar o raciocínio sobre ela.

Na abordagem centrada na organização, é possível obter uma descrição da organização que a sociedade está adotando sem precisar observar seu comportamento. Por exemplo, a descrição da organização de uma universidade existe em manuais, códigos de ética, organogramas, etc. Nessa abordagem, os autores propõem modelos organizacionais para descrever de forma explícita a organização levando em consideração as normas e os princípios que a norteiam. Esses modelos possuem como objetivo restringir o comportamento individual dos agentes buscando produzir um comportamento global direcionado a um objetivo social [Hübner et al. 2002]. Diversos modelos organizacionais têm sido propostos na literatura e cada um oferece uma proposta diferente para representar a organização societal dos agentes em qualquer contexto.

O modelo Moise+ é um modelo organizacional que considera a organização de um SMA segundo três dimensões: a estrutura (papéis e hierarquias), o funcionamento (planos globais) e as normas (obrigações) da organização [Hübner et al. 2002]. O aspecto deôntico liga os aspectos estruturais e o funcionais indicando quais as responsabilidades dos papéis nos planos globais. A estrutura organizacional envolve a definição de papéis, relações entre papéis e os grupos que possibilitam a definição dos níveis individual, social e coletivo. No nível individual, os papéis dos agentes são definidos juntamente com as obrigações das tarefas pertencentes a cada papel (e.g., o professor deve mediar o conhecimento). No nível social, as ligações entre os papéis (e.g., o professor tem autoridade sobre o aluno) e também as relações de compatibilidade entre os papéis (um agente não pode assumir um papel de aluno na turma onde é professor) são estabelecidas. No nível coletivo, um grupo agrega agentes com objetivos comuns e estes grupos podem ser usados na definição de várias estruturas organizacionais.

A área de Máquina Ética para SMA está em evolução e inúmeras discussões estão em aberto, tais como a escolha da melhor abordagem societal para um SMA e a capacidade deste em avaliar sua própria organização para propor mudanças organizacionais. Contudo, a construção de dispositivos físicos usando um SMA embarcado ainda não leva em consideração regras e princípios éticos e tais dispositivos atuam no ambiente sem nenhuma restrição comportamental. Para o desenvolvimento de uma Máquina Ética para SMA embarcado, os passos necessários são propostos nas próximas subseções.

2.1. Escolha de um modelo organizacional que considere construções éticas

Nesta etapa, os modelos organizacionais para SMA, que restringem o comportamento dos agentes segundo as regras e condutas éticas de uma organização, podem ser investigados e testados para um ambiente embarcado. Devido à relevância de trabalhos que empregaram

o modelo Moise+, este pode ser o primeiro modelo organizacional a ser testado para um SMA embarcado. Na investigação dos modelos organizacionais para SMA, pode-se identificar a necessidade de modificá-los para atender aos seguintes pontos:

- **A aderência dos modelos organizacionais aos princípios éticos para uma Máquina Ética.** Segundo [Huang et al. 2022], os princípios para a criação de uma Máquina Ética podem ser classificados em: transparência, equidade e justiça, responsabilidade e prestação de contas, não maleficência, beneficência, solidariedade, sustentabilidade, confiabilidade e dignidade. Os modelos organizacionais existentes podem ser avaliados à luz desses princípios e novas versões podem ser propostas para adequá-los;
- **O modelo organizacional escolhido deve permitir a existência de níveis organizacionais comuns na sociedade em que vivemos.** Por exemplo, um indivíduo deve seguir as regras e condutas éticas da sua profissão (nível micro), da empresa onde trabalha (nível meso) e também seguir as leis federais do governo (nível macro). Dessa forma, o comportamento de uma Máquina Ética deve estar também restrita por essas três perspectivas. Logo, o modelo organizacional para um SMA embarcado precisa garantir que os agentes tenham seus comportamentos restringidos por esses três níveis organizacionais e o caminho para alcançar esse objetivo ainda precisa ser construído;
- **Permitir que atualizações de leis, regras e condutas éticas de uma sociedade sejam refletidas para um SMA em tempo de execução.** Quando ocorre uma mudança nas leis, regras ou condutas, essas alterações devem ser repassadas para os modelos organizacionais em tempo de design. Contudo, quando se trata de SMA embarcado é preciso se pensar em um sistema adaptável ao ambiente em tempo de execução de forma que as mudanças ocorridas em qualquer nível sejam percebidas pelo SMA durante sua execução. Nesse sentido, é necessário pensar em soluções que possibilitem essa atualização em tempo de execução.

2.2. Especificação de uma organização com as regras e condutas éticas em um determinado contexto

Nesta etapa, será necessário o mapeamento de um conjunto de regras e condutas éticas reais de uma organização para o modelo organizacional, ou seja, as regras e condutas éticas de uma organização humana devem ser traduzidas para uma linguagem inteligível por um SMA e esquematizadas no modelo organizacional. Para facilitar esse processo, o pensamento computacional ético pode ser usado previamente pelos projetistas de SMA para a identificação de dilemas éticos em determinado contexto, a avaliação e a escolha da opção mais ética em cenários simulados e a antecipação de impactos negativos e positivos da IA aos indivíduos. Esse pensamento contribui para que os projetistas de SMA compreendam as possíveis consequências éticas das ações e tomadas de decisão que podem ser realizadas pela IA. Consequentemente, esses projetistas podem especificar melhor as restrições comportamentais do SMA no modelo organizacional. Nesse cenário, um ferramental pode ser desenvolvido para apoiar o mapeamento e o pensamento computacional ético dos projetistas de SMA, zelando para que o desenvolvimento e o uso da tecnologia ocorram de maneira ética e responsável.

2.3. Desenvolvimento de um SMA que integre as especificações da organização ao gerenciamento das ações e tomadas de decisão em um dispositivo

Esta etapa aborda o desenvolvimento do SMA pelo projetista considerando as especificações das regras e condutas éticas contidas no modelo organizacional. Em determinado contexto, o SMA é desenvolvido para atuar no ambiente de forma autônoma, no entanto, tendo como limites as restrições comportamentais da organização. Nesse sentido, experimentos podem ser conduzidos para averiguar se o comportamento do SMA no ambiente é adequado às questões éticas. Para isso, os cenários com questões éticas definidos por meio do pensamento computacional ético podem ser simulados e, eventualmente, ajustes na especificação do modelo organizacional podem ser identificados. Nesses experimentos, o atendimento aos princípios de uma Máquina Ética também devem ser avaliados.

2.4. Um ferramental que apoie o desenvolvimento de uma Máquina Ética para SMA, a embarcação da especificação de sua organização e propicie o ensino e aprendizagem do pensamento computacional ético

O ferramental existente para o desenvolvimento de um SMA não apoia o projetista na criação de uma Máquina Ética para SMA embarcado e há várias lacunas que precisam ser preenchidas para garantir avanços nessa área.

- Os modelos organizacionais que inserem preocupações éticas ao SMA ainda não foram testados e adaptados para SMA embarcados.
- Não há práticas de ensino e aprendizagem do pensamento computacional que auxiliem os projetistas de SMA a desenvolverem habilidades críticas para a especificação das regras e condutas éticas nos modelos organizacionais.
- Também não há uma IDE que possibilite o desenvolvimento de um SMA embarcado de forma integrada com um modelo organizacional.

Essas lacunas podem ser resolvidas com o desenvolvimento de um ferramental para apoiar a criação de Máquinas Éticas para SMA embarcado. Há ferramentas que suportam o desenvolvimento de um SMA como a *Cognitive Hardware on Network - Integrated Development Environment* (ChonIDE) e o *Visual Studio Code* (VS Code) [Souza de Jesus et al. 2023]. No entanto, a ChonIDE é uma plataforma *web* e *open source* que apoia especificamente o desenvolvimento de um SMA embarcado considerando as camadas de arquitetura que necessitam de programação. Para facilitar a programação de uma Máquina Ética para SMA embarcado, uma camada adicional pode ser integrada a ChonIDE permitindo que o projetista especifique um modelo organizacional e embarque diferentes níveis de regras e condutas éticas em um dispositivo. Além disso, a IDE pode abarcar as técnicas para o desenvolvimento do ensino e aprendizagem do pensamento computacional. Diante do exposto, propomos os seguintes objetivos para serem atingidos para a criação de Máquinas Éticas Embarcadas:

1. Mapeamento da literatura para investigar os modelos organizacionais e os princípios éticos para a criação de uma Máquina Ética para SMA;
2. Realizar uma análise preliminar da aderência dos modelos organizacionais aos princípios éticos;
3. Escolher um modelo organizacional para a condução de testes com regras e condutas éticas de uma organização real;
4. Realizar testes preliminares para embarcar o SMA integrado com o modelo organizacional.

3. Resultados Esperados

Este estudo pretende levantar e comparar os modelos organizacionais existentes na literatura para a criação de uma Máquina Ética para SMA e também compreender como os níveis organizacionais podem ser especificados nesses modelos. Além disso, este estudo objetiva identificar possíveis lacunas na representação das regras e condutas éticas em uma organização e avaliar a aderência desses modelos aos princípios éticos. Um estudo prático também será conduzido para entender o comportamento de um SMA sob as restrições comportamentais éticas de uma organização e realizar testes preliminares para embarcar o SMA integrado com o modelo organizacional.

Referências

Albino, V., Berardi, U., and Dangelico, R. M. (2015). Smart cities: Definitions, dimensions, performance, and initiatives. *Journal of urban technology*, 22(1):3–21.

Allen, C., Wallach, W., and Smit, I. (2006). Why machine ethics? *IEEE Intelligent Systems*, 21(4):12–17.

Boden, M. A. (1996). *Artificial intelligence*. Elsevier.

Briot, J.-P. and Demazeau, Y. (2001). *Principes et architecture des systèmes multi-agents*. Hermès-Lavoisier.

Gilbert, N. and Conte, R. (2006). Manta: New experimental results on the emergence of (artificial) ant societies. In *Artificial societies*, pages 172–191. Routledge.

Huang, C., Zhang, Z., Mao, B., and Yao, X. (2022). An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence*, 4(4):799–819.

Hübner, J. F., Sichman, J. S., and Boissier, O. (2002). Moise+: towards a structural, functional, and deontic model for mas organization. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems*, pages 501–502.

Lemaître, C. and Excelente, C. B. (1998). Multi-agent organization approach. In *Proceedings of II Iberoamerican Workshop on DAI and MAS*, pages 7–16. Toledo.

Management Association, I. (2017). *The Internet of Things: Breakthroughs in Research and Practice: Breakthroughs in Research and Practice*. Critical explorations. IGI.

Metz, C. (2019). Is ethical ai even possible. *The New York Times*, 1(3).

Nalini, B. (2019). The hitchhiker's guide to ai ethics medium.

Paulo, N. (2023). The trolley problem in the ethics of autonomous vehicles. *The Philosophical Quarterly*, 73(4):1046–1066.

Russell, S. J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Pearson.

Souza de Jesus, V., Mori Lazarin, N., Pantoja, C. E., Vaz Alves, G., Ramos Alves de Lima, G., and Viterbo, J. (2023). An IDE to Support the Development of Embedded Multi-Agent Systems. In Mathieu, P., Dignum, F., Novais, P., and De la Prieta, F., editors, *Advances in Practical Applications of Agents, Multi-Agent Systems, and Cognitive Mimetics. The PAAMS Collection*, pages 346–358, Cham. Springer Nature Switzerland. DOI: https://doi.org./10.1007/978-3-031-37616-0_29.

Weiss, G. (1999). *Multiagent systems: a modern approach to distributed artificial intelligence*. MIT press.