UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
CENTRO INTERDISCIPLINAR DE NOVAS TECNOLOGIAS NA EDUCAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA NA EDUCAÇÃO

PAULO ROBERTO CÓRDOVA

ETHOSCOOL: AVANÇANDO A ÉTICA EM IA NA EDUCAÇÃO POR MEIO DE AGENTES MORAIS ARTIFICIAIS

TESE

PORTO ALEGRE
2023

PAULO ROBERTO CÓRDOVA

ETHOSCOOL: AVANÇANDO A ÉTICA EM IA NA EDUCAÇÃO POR MEIO DE AGENTES MORAIS ARTIFICIAIS

Tese apresentado(a) como requisito para obtenção do título(grau) de Doutor em Informática na Educação, do Programa de Pós-Graduação em Informática na Educação, da Universidade Federal do Rio Grande do Sul (UFRGS).

Orientador(a): Prof(a). Dr(a). Rosa Maria Vicari

PORTO ALEGRE

2023

CIP - Catalogação na Publicação

Cordova, Paulo Roberto
ETHOSCOOL: AVANÇANDO A ÉTICA EM IA NA EDUCAÇÃO POR MEIO DE AGENTES MORAIS ARTIFICIAIS / Paulo Roberto Cordova. -- 2024.
137 f.
Orientadora: Rosa Maria Vicari.

Tese (Doutorado) -- Universidade Federal do Rio Grande do Sul, Centro de Estudos Interdisciplinares em Novas Tecnologias na Educação, Programa de Pós-Graduação em Informática na Educação, Porto Alegre, BR-RS, 2024.

1. Inteligência Artificial. 2. Ética. 3. Educação. 4. Aprendizagem. 5. Alinhamento de Valores. I. Vicari, Rosa Maria, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

Paulo Roberto Córdova

ETHOSCOOL: AVANÇANDO A ÉTICA EM IA NA EDUCAÇÃO POR MEIO DE AGENTES MORAIS ARTIFICIAIS

Tese apresentada ao Programa de Pós-Graduação em Informática na Educação do Centro Interdisciplinar de Novas Tecnologias na Educação da Universidade Federal do Rio Grande do Sul, como requisito para obtenção do título de Doutor em Informática na Educação.

Aprovada em ____/____/_____.

_____

Prof(ª). Dr(ª). Rosa Maria Vicari – UFRGS – Orientadora

_____

Prof. Dr. Dante Augusto Couto Barone – UFRGS – Membro da Banca

_____

Prof. Dr. Edson Prestes e Silva Junior – UFRGS – Membro da Banca

_____

Prof. Dr. Joel Haroldo Baade – UNIARP – Membro da Banca

RESUMO

O avanço das tecnologias de Inteligência Artificial (IA) tem levantado preocupações acerca dos impactos éticos deste avanço em diversos campos. Para mitigar tais impactos, estudos recentes sugerem que a IA deve ser alinhada a valores humanos. Uma abordagem importante, nesse contexto, é a ética por design, que busca desenvolver sistemas autônomos com capacidade de raciocínio ético. No campo da educação, entretanto, as pesquisas existentes têm se concentrado principalmente nas implicações sociais da expansão da IA, negligenciando questões éticas fundamentais, como aquelas abordadas pelo framework *Fairness, Accountability, Transparency* e *Ethics* (FATE). Nesse sentido, a presente pesquisa propõe e descreve um modelo de Agente Moral Artificial (AMA) capaz de promover o engajamento comportamental em grupos de aprendizagem colaborativa. Os resultados evidenciaram, apesar das fragilidades identificadas, a capacidade do agente de tomar decisões éticas alinhadas a princípios deontológicos pré-estabelecidos e reconhecidos pela comunidade de pesquisa e a sua proficiência em lidar com dilemas éticos usando raciocínio utilitarista. Tais resultados marcam um progresso fundamental em direção a uma IA confiável no ambiente de ensino e aprendizagem. A integração ética da IA na educação e o estabelecimento de diretivas que permitem futuros avanços neste campo estão entre as principais contribuições desta pesquisa para o campo da Informática na Educação.

Palavras-chave: Inteligência Artificial. Ética. Educação. Aprendizagem. Alinhamento de Valores.

ABSTRACT

The advancement of Artificial Intelligence (AI) technologies has raised concerns about the ethical impacts of this progress in various fields. To mitigate these impacts, recent studies suggest that AI should be aligned with human values. An important approach in this context is ethics by design, which aims to develop autonomous systems with ethical reasoning capabilities. In the field of education, however, existing research has primarily focused on the social implications of AI expansion, neglecting fundamental ethical issues addressed by the Fairness, Accountability, Transparency, and Ethics (FATE) framework. In this regard, the current research proposes and describes a model of Artificial Moral Agent (AMA) capable of promoting behavioral engagement in collaborative learning teams. The results, despite identified weaknesses, demonstrated the agent's ability to make ethical decisions aligned with pre-established deontological principles recognized by the research community and its proficiency in handling ethical dilemmas using utilitarian reasoning. These outcomes mark a significant step toward reliable AI in the educational environment. The ethical integration of AI in education and the establishment of guidelines enabling future advancements in this field are among the main contributions of this research to the field of Informatics in Education.

Keywords: Artificial Intelligence. Ethics. Education. Learning. Value Alinments.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A utilização de tecnologias de Inteligência Artificial (IA) para dar suporte a tarefas cotidianas vem ganhando novos espaços e mostrando-se capaz de promover transformações reais na forma como as pessoas interagem, resolvem problemas e tomam decisões (SARKER *et al.*, 2021; KIM *et al.*, 2021). Com isso, torna-se cada vez mais necessário pensar profundamente sobre os potenciais impactos, se positivos ou negativos, que isto pode ter (MABASO, 2020). Esta preocupação se justifica, em razão das evidências acumuladas acerca dos riscos que o mau uso da IA pode representar, seja este mau uso planejado ou não (HAN *et al.*, 2021).

Por esta razão, Bostrom (2003) argumenta que os esforços no desenvolvimento destas tecnologias deveriam estar mais focados em evitar consequências negativas do que em atingir resultados positivos, mesmo que isso implique na perda de bons resultados. Dessa forma, segundo Dignum (2018), para que seja possível aproveitar plenamente os benefícios potenciais da IA, é necessário que ela esteja alinhada com os valores morais e princípios éticos humanos. Assim, pesquisadores de diferentes áreas têm buscado soluções relacionadas à ética de máquina, enfrentando uma série de desafios no intuito de encontrar soluções capazes de tornar a IA mais previsível, confiável e, consequentemente, segura (CóRDOVA *et al.*, 2021).

Nesse contexto, existem três dimensões principais que abordam este problema: a ética por design, interessada em desenvolver soluções algorítmicas para dotar sistemas artificiais autônomos com capacidade de raciocínio ético; a ética em design, que aborda métodos regulatórios e de engenharia para apoiar a análise das implicações éticas que sistemas de IA podem gerar para a sociedade e; a ética para o design, que foca em códigos de condutas que tentam garantir a integridade de desenvolvedores e usuários de IA (DIGNUM *et al.*, 2018). O presente trabalho busca investigar soluções em ética por design. Para esta dimensão de abordagem, o conjunto de esforços empreendido pela comunidade de pesquisa, para a produção de sistemas inteligentes alinhados a valores humanos, está organizado em um campo denominado Alinhamento de Valores (VA[1]) em IA (KIM *et al.*, 2021).

Pesquisadores deste campo têm buscado responder a importantes questões, entre elas: como traduzir princípios éticos em modelos computacionais, como evitar viés de dados que implique na replicação de preconceitos humanos, como tornar os sistemas inteligentes responsáveis

[1] Neste trabalho será usada a sigla para o termo em inglês, *Value Alignment*, por ser este o mais difundido na comunidade científica.

por suas decisões, entre outros (CóRDOVA *et al.*, 2021). Este conjunto de questões deu origem ao desenvolvimento de um tipo especial de agente artificial, capaz de demonstrar comportamentos considerados éticos: o Agente Moral Artificial (AMA).

A noção de agentes **agentes morais** resgata uma abordagem antropomórfica da IA, ou seja, a visão na qual suas característica e funcionalidades são concebidas, caracterizadas e descritas por traços humanos (WATSON, 2019). Esta abordagem têm sido objeto de discussão no campo da IA, levantando questões éticas e epistemológicas (WATSON, 2019; SALLES *et al.*, 2020). A este respeito não há consenso. É notório que pesquisadores da área adotam, com frequência, terminologias e conceitos antropomórficos no intuito de compreender e controlar melhor o desenvolvimento da IA (WATSON, 2019). Entretanto, é importante que tais termos não sejam tomados literalmente, mas como referências que facilitem a organização e compreensão das características de cada tecnologia de IA (SALLES *et al.*, 2020). No presente trabalho serão consideradas diferentes abordagens e terminologias antropomórficas clássicas da IA, sendo o AMA uma delas.

Feita essa ressalva, pode se afirmar que contexto educacional, mais especificamente quando se trata de soluções para ensino e aprendizagem, questões éticas em IA podem se tornar especialmente graves (CASAS-ROMA *et al.*, 2021). Isto porque, a exemplo do que ocorre em outras áreas, a tendência é que haja um aumento do uso de soluções de IA para apoiar o ensino e a aprendizagem (VICARI, 2021). Nesse particular, podem ser destacados alguns recursos de IA que vem sendo aplicados de forma dispersa, como: Processamento de Linguagem Natural (PLN), para tradução análise e interpretação de texto; computação afetiva, integrada a PLN e a Sistemas Tutores Inteligentes (STI); sistemas de recomendação de conteúdo pedagógico, integrados a *Learning Management Systems* (LMS); *Learning Analytics* e *Big Data*, para, entre outras coisas, realizar previsões sobre o comportamento futuro de alunos com base em seus comportamentos passados (VICARI, 2021); e, mais recentemente, linguagens generativas, com destaque para o ChatGPT, que têm impactado notavelmente os processos de ensino e aprendizagem (GARCíA-PEñALVO *et al.*, 2023).

Além disso, discussões contemporâneas sobre a Educação 4.0 enfatizam a Inteligência Artificial como um componente central desse novo paradigma, resultando em novos requisitos para a inclusão da IA nos currículos modernos (OLIVEIRA *et al.*, 2023). A crescente complexidade do cenário digital também está exigindo a incorporação de elementos mais avançados da ciência da computação, incluindo aprendizagem de máquina e IA, nas fases finais da educação

básica (BOCCONI *et al.*, 2022).

Entretanto, apesar desta versatilidade, os resultados do uso da IA no contexto educacional têm se mostrado mais profícuos para o ensino personalizado, sendo ainda um desafio encontrar soluções satisfatórias para a aprendizagem colaborativa (VICARI, 2021). Nota-se, portanto, que, não só há espaço para novas soluções neste segmento, como é desejável que haja, uma vez que a aprendizagem colaborativa é capaz de aperfeiçoar as práticas de aprendizagem por meio da promoção da autonomia, interdisciplinaridade e participação ativa dos alunos, consistindo em uma alternativa que tem se mostrado bastante efetiva (CARNEIRO *et al.*, 2020). Com isso, alguns trabalhos nesse sentido têm sido desenvolvidos, mas os principais resultados ainda se restringem a algoritmos de recomendação de conteúdo educacional, suporte à formação de grupos e analisadores de diálogos, capazes de identificar alunos que não colaboraram e emitir lembretes sobre questões em aberto (VICARI, 2021).

De todo modo, é notório que as tecnologias de IA estão cada vez mais presentes nas interações sociais contemporâneas (SARKER *et al.*, 2021). Por esta razão, é importante considerar que, no caso da sua aplicação para fins de ensino e aprendizagem, o que está em jogo são processos formativos de seres humanos (CóRDOVA *et al.*, 2021). É preciso, portanto, refletir sobre os efeitos que a aplicação de tecnologias de IA para a educação poderá causar nas gerações atuais e futuras (CASAS-ROMA *et al.*, 2021). Assim, implicações éticas do uso de sistemas capazes de apoiar, conduzir ou influenciar a aprendizagem precisam ser consideradas com a mesma seriedade que têm sido em outras áreas. Ainda assim, apesar da crescente utilização de soluções de IA para fortalecer o ensino e a aprendizagem, as pesquisas atuais têm se concentrado predominantemente nas implicações sociais de sua implementação e avanço, deixando questões relacionadas à ética por design, como o framework *Fairness, Accountability, Transparency* e *Ethics* (FATE), sem a atenção apropriada (WOOLF, 2022).

Nesse sentido, visando atender às oportunidades de investigação apresentadas, a presente pesquisa tem por objetivo propor e descrever um modelo de agente moral artificial capaz de promover o engajamento comportamental em grupos de aprendizagem colaborativa.

Para atender a este objetivo, foi descrita uma solução baseada em Sistemas Multiagentes (SMA) contendo um AMA com o objetivo de gerar engajamento comportamental em alunos em um contexto de aprendizagem colaborativa. Esse AMA, por ser implementado utilizando uma abordagem *top-down*, faz uso da estrutura ética deontológica para guiar seus objetivos e da estrutura ética utilitarista para lidar com dilemas éticos. A prova de conceito mostrou que o

modelo proposto, apesar das reconhecidas fragilidades, foi capaz de demonstrar comportamento considerado ético, que segundo Dignum (2018), são aqueles que envolvem decisões relacionadas com, ou que impactam diretamente a dignidade ou bem-estar humano. Além disso, os testes também evidenciaram a capacidade do modelo para lidar com dilemas éticos, constituindo, com isso, um importante passo na direção de uma IA mais confiável para o ambiente de ensino.

1.1 PROBLEMA DE PESQUISA

Até alguns anos, acreditava-se que o uso da IA para tomada de decisões era a forma mais confiável de eliminar preconceitos e decisões enviesadas, uma vez que tais decisões seriam baseadas em dados objetivos e neutras com relação a interesses e preconceitos. Contudo, logo percebeu-se que a forma como os dados eram coletados, representados, selecionados e usados, bem como o modo como os algoritmos e suas regras eram codificados, poderiam facilmente, de diferentes maneiras, encapsular preconceitos pessoais, sociais e históricos (CASAS-ROMA *et al.*, 2021). Um exemplo disso ocorreu nos Estados Unidos, quando um sistema usado para avaliar o risco de reincidência entre réus foi considerado discriminatório contra negros (FAVARETTO *et al.*, 2019).

Este é apenas um exemplo de risco real oferecido por tecnologias de IA. Outros cenários descritos na literatura podem ser igualmente preocupantes. Estes riscos explicam tantos esforços no sentido de construir sistemas capazes de demonstrar comportamento ético para suportar diferentes áreas, como é possível observar em um estudo sobre o estado da arte dos AMAs publicado por Cervantes *et al.* (2020). Apesar disso, não é possível, até o momento, encontrar estudos sobre o desenvolvimento deste tipo de solução aplicada à área da educação.

Este pode ser considerado um fato grave, uma vez que tecnologias de IA aplicadas à educação não podem ser equiparadas a outras Tecnologias de Informação e Comunicação (TIC), que podem ser reguladas por meio de codificação de regras. Sistemas baseados em IA podem agir de forma autônoma, reagir e afetar o seu ambiente (CASAS-ROMA *et al.*, 2021). Além disso, ao interagir com os alunos, estes sistemas estão constantemente sujeitos a ter de lidar com situações inéditas que podem exigir a resolução de dilemas éticos (CORDOVA; VICARI, 2021).

Por fim, é importante destacar novamente a carência de soluções baseadas em IA para suportar processos de aprendizagem colaborativa. Tais tecnologias têm sido aplicadas com êxito no âmbito da aprendizagem personalizada (VICARI, 2021), o que é muito significativo. Entretanto, considerando os benefícios da aprendizagem colaborativa, que incluem a promoção

da autonomia, da participação ativa e do foco no aluno (CARNEIRO *et al.*, 2020), entre outras, pode-se afirmar que a investigação de soluções para suprir esta demanda é igualmente desejável.

Dentro do contexto apresentado, pergunta-se: Quais requisitos são relevantes em um modelo de agente moral artificial capaz de incentivar o engajamento comportamental em grupos de aprendizagem colaborativa?

1.2 OBJETIVOS DA PESQUISA

1.2.1 Objetivo Geral

A presente pesquisa visa propor e descrever um modelo de agente moral artificial capaz de promover o engajamento comportamental em grupos de aprendizagem colaborativa.

1.2.2 Objetivos Específicos

a Propor uma arquitetura com características funcionais e éticas para agentes morais artificiais, capaz de suportar processos de ensino e aprendizagem;

b Modelar a arquitetura proposta usando técnicas de Engenharia de Software baseada em Agentes (ESA);

c Implementar o modelo proposto e realizar testes de laboratório para avaliar sua capacidade de demonstrar comportamento considerado ético.

1.3 JUSTIFICATIVA

Na medida em que as tecnologias de IA avançam em direção a diferentes áreas e ocupam cada vez mais espaço na sociedade contemporânea, mais importantes se tornam ações no sentido de aumentar sua confiabilidade e previsibilidade para que seus usuários se sintam mais seguros em compartilhar ou delegar a elas tarefas do dia a dia (DIGNUM, 2018; MABASO, 2020; CóRDOVA *et al.*, 2021). Nesse sentido, embora cada domínio suscite preocupações específicas, a maioria das discussões se concentrou em carebots e bots militares, pois eles parecem levantar mais questões éticas (FORMOSA; RYAN, 2020).

No entanto, visto que trata da formação humana, abarcando, inclusive, a capacitação para a proficiência nas mais diferentes técnicas, nos seus âmbitos político, social, ético e estético

(RIOS, 2009), e ainda enfrentando os influxos advindos das transformações tecnológicas causadas pela expansão da IA (VICARI, 2021), torna-se imprescindível conferir à área da Educação uma atenção análoga àquela dispensada a outros campos do conhecimento, no tocante ao uso ético da IA. Com isso, a importância do presente trabalho se evidencia por fortalecer a discussão acerca do alinhamento de valores em IA na área da educação, propor um modelo usando ética por design para ser usado em processos de ensino, e, com isso, iniciar a pesquisa aplicada neste campo.

2 REFERENCIAL TEÓRICO

O presente capítulo apresenta as publicações, conceitos e definições que fundamentam o presente trabalho. Primeiramente são tratados temas relacionados à IA, Sistemas Multiagentes e autonomia nesse contexto. Em seguida, são apresentados alguns problemas e soluções vigentes para a Ética por Design. Para isso, são introduzidos alguns conceitos centrais deste campo, bem como do campo da Ética enquanto área da Filosofia, para o desenvolvimento de Agentes Morais Artificiais. Por fim, são abordadas questões referentes à aprendizagem colaborativa, seus fundamentos teóricos e as principais contribuições da Aprendizagem Colaborativa com Suporte Computacional – do inglês, *Computer Supported Collaborative Learning* (CSCL).

2.1 INTELIGÊNCIA ARTIFICIAL

O campo da Inteligência Artificial pode ser considerado um dos mais recentes em ciências e engenharia. Seu desenvolvimento teve início logo após a segunda guerra mundial, tendo seu nome cunhado em 1956, em um seminário liderado por John McCarthy em Dartmouth (RUSSEL; NORVIG, 2013). Atualmente, a IA abrange uma grande variedade de subcampos, como: Processamento de Linguagem Natural (PLN), Representação de Conhecimento, Raciocínio Automatizado, Aprendizagem de Máquina (ML[1]), Visão Computacional e Robótica. Trata-se, portanto, de uma área bastante complexa, difícil de caracterizar, delimitar e relevante para qualquer tarefa intelectual (RUSSEL; NORVIG, 2013).

Do ponto de vista filosófico, a asserção de que as máquinas talvez possam agir de forma inteligente é chamada hipótese de IA fraca, enquanto a asserção de que as máquinas que o fazem estão realmente pensando, ao invés de simulando o pensamento, é chamada de hipótese de IA forte (RUSSEL; NORVIG, 2013). Searle (1990) argumenta que o objetivo da IA forte é projetar e construir máquinas que, ao invés de representar um modelo da mente, seriam literalmente uma mente, no mesmo sentido que uma mente humana. Por outro lado, o objetivo da IA fraca é seguir uma abordagem mais cautelosa, assumindo que os modelos de inteligência simulados por computador podem ser úteis para estudar problemas específicos, como o clima, a economia, a biologia molecular, entre outras.

Com relação à definição de IA, de acordo com Kim *et al.* (2019), a IA é uma tentativa

[1] Neste trabalho foi adotada a sigla em inglês para *Machine Learning* (ML), por ser mais amplamente utilizada.

de imitar a inteligência humana. Nath e Sahu (2020), por sua vez, afirmam que o principal objetivo da IA é entender, recriar e possivelmente superar a inteligência humana em entidades artificiais. Para McCarthy (2007), entretanto, ela é a ciência e a engenharia de construir máquinas, em especial, programas de computador, dotados de inteligência. Nesse sentido, o mesmo autor (2007) define inteligência como a parte computacional da habilidade de atingir objetivos no mundo, sendo variados os graus e entidades – como pessoas, animais ou máquinas – em que esta inteligência pode ocorrer.

Entretanto, a inteligência não é algo que possa ser abstraída e explicada facilmente, tão pouco simplificada. O que se conseguiu até agora é a compreensão de alguns dos seus mecanismos, sendo que a pesquisa em IA descobriu, também, como fazer com que computadores executem alguns destes mecanismos (McCarthy, 2007). Com isso, a IA precisa considerar métodos que vão além de simplesmente imitar a inteligência humana (RUSSEL; NORVIG, 2013). É possível desenvolver técnicas para fazer as máquinas resolverem problemas observando pessoas, mas a maior parte dos esforços em IA envolvem estudar problemas que o mundo apresenta à inteligência, ao invés de estudar o comportamento de pessoas ou animais a fim de replicá-los. Assim, pesquisadores em IA devem ser livres para usar os métodos que não são observados em seres humanos ou que superem a sua capacidade computacional (McCarthy, 2007).

Por esta razão, para melhor compreender a IA, é preciso considerar também as diferentes abordagens por meio das quais este campo vem se desenvolvendo. Tais abordagens incluem a estratégia da Modelagem Cognitiva, a abordagem das Leis do Pensamento e a abordagem do Agente Racional (RUSSEL; NORVIG, 2013).

- A modelagem cognitiva abrange a ideia de que, se "[...] um dado programa pensa como um ser humano, temos de ter alguma forma de determinar como os seres humanos pensam" (RUSSEL; NORVIG, 2013, p.5). Portanto, busca compreender como funcionam os mecanismos da cognição humana para propor modelos computacionais capazes de representar estes mecanismos. Recebe muitas contribuições e também contribui com a ciência cognitiva, um campo interdisciplinar que reúne modelos computacionais e técnicas experimentais da psicologia para construir teorias mais acuradas e verificáveis dos processos de funcionamento da mente (RUSSEL; NORVIG, 2013). Em suma, esta abordagem inclui tecnologias concebidas para pensar, em algum grau, como um ser humano.

- A abordagem das leis do pensamento, por sua vez, representa a chamada tradição logicista

dentro da IA. Com isso, tendo como base a existência de certas leis que conduzem ao pensamento correto e a raciocínios irrefutáveis, esta abordagem faz uso da linguagem lógica, que dispõe de notações precisas para representar coisas do mundo e as relações entre elas. Além disso, também provê meios para a realização de inferências lógicas (RUSSEL; NORVIG, 2013). “Por volta de 1965, existiam programas que, em princípio, podiam resolver qualquer problema solucionável descrito em notação lógica” (RUSSEL; NORVIG, 2013, p.6). Entretanto, é bastante difícil enunciar conhecimento informal nos termos da notação lógica formal, principalmente quando este conhecimento é menos que 100% correto. Além disso, a resolução de problemas complexos, com algumas centenas de fatos podem esgotar os recursos computacionais disponíveis na maioria das máquinas. Essas dificuldades, embora sejam comuns a qualquer tentativa de construir sistemas de raciocínio computacional, surgiram primeiro na tradição logicista (RUSSEL; NORVIG, 2013).

- Por fim, a abordagem do agente racional é explicada por Russel e Norvig (2013) como sendo aquela que se ocupa com o desenvolvimento de agentes racionais. Nesse contexto, espera-se que agentes racionais sejam capazes de operar sob controle autônomo, perceber seu ambiente, persistir por um período de tempo prolongado, se adaptar a mudanças, além de criar e perseguir metas. Em suma, “um agente racional é aquele que age para alcançar o melhor resultado ou, quando há incerteza, o melhor resultado esperado” (RUSSEL; NORVIG, 2013, p.6).

Entre as três abordagens, a do agente racional é a mais abrangente e traz ao menos duas vantagens em relação às demais: ela é mais geral que a abordagem de leis do pensamento, pois a inferência correta usando raciocínio lógico é apenas um entre os vários mecanismos possíveis para que um agente alcance a racionalidade; ela é mais acessível ao desenvolvimento científico do que a estratégia de modelagem cognitiva, que é baseada no comportamento e no pensamento humano (RUSSEL; NORVIG, 2013). Além disso, por ser a abordagem de maior interesse para os fins do presente trabalho, maior foco será dado a ela no decorrer deste trabalho.

2.1.1 Agentes e Sistemas Multiagentes

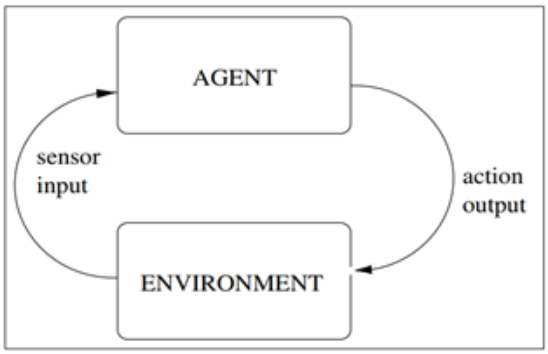
Além da definição estabelecida no capítulo 2.1, agentes racionais, também chamados de agentes artificiais inteligentes, mas, a partir de agora referidos apenas de Agentes Artificiais

(AA) para simplificar, se caracterizam por serem programas de computador capazes de decidir por si mesmos o que precisam fazer para satisfazer seus objetivos de design (WOOLDRIDGE, 2001). Para Cardoso e Ferrando (2021), AAs deste tipo são capazes de decidir, de forma reativa ou proativa, sobre um curso de ação, raciocinando sobre as informações que estão disponíveis sobre o mundo, incluindo o meio ambiente, o próprio agente e outros agentes.

Outra característica importante dos AAs é que eles devem ser capazes de operar de forma robusta em ambientes abertos, imprevisíveis ou que mudam rapidamente (WOOLDRIDGE, 2001). Estas características centrais, entretanto, parecem ser o único ponto de consenso com relação a definição de AA no contexto da IA. Para Wooldridge (2001) a discordância em relação a uma definição universal está relacionada ao tipo de problemas aos quais são aplicados tais agentes. Em alguns casos, por exemplo, a capacidade de aprendizagem é de extrema importância, em outros, é indesejável.

Os AAs devem também, segundo Wooldridge (2001), ter a capacidade de interagir com o seu ambiente e, em determinadas condições, alterá-lo. Nesse sentido, estes agentes devem dispor de um repertório de ações possíveis, que representam a sua capacidade potencial para modificar o ambiente. A figura 1 apresenta uma arquitetura básica que representa um agente em seu ambiente.

Figura 1 – Um agente em seu ambiente.



Fonte: Wooldridge (2001).

No modelo apresentado, “o agente obtém a entrada sensorial do ambiente e produz como saída ações que o afetam. A interação é geralmente contínua, sem término” (WOOLDRIDGE, 2001, p.6). Com isso, o principal problema que um agente deve resolver internamente, é decidir qual das suas ações deve ser executada para satisfazer seus objetivos de design. Esta decisão, para ser considerada racional, deve maximizar a medida de desempenho do agente, considerando o conhecimento prévio e atual que ele tem sobre o ambiente e o conjunto de ações que dispõe para executar (RUSSEL; NORVIG, 2013).

Além disso, a complexidade do processo de tomada de decisão, por parte do AA, pode ser influenciada por um conjunto de diferentes propriedades ambientais (WOOLDRIDGE, 2001). Dada a importância que o ambiente representa para o projeto e construção de agentes, Russel e Norvig (2013) sugeriram uma classificação considerando as suas propriedades. Segundo esta classificação, os ambientes podem ser:

- Acessíveis ou inacessíveis: ambientes acessíveis são aqueles sobre quais o AA pode obter informações completas, precisas e atualizadas sobre o estado. Ambientes inacessíveis, ao contrário, não possibilitam ao agente esta percepção.

- Agente único ou multiagentes: ambientes de agente único são compostos por apenas um agente tentando maximizar sua medida de desempenho. Ambientes multiagentes podem ser competitivos, quando, ao tentar maximizar sua medida de desempenho, determinado agente acaba reduzindo a de outro agente, ou cooperativo, quando ao maximizar sua medida de desempenho, o agente maximiza também a de outro(s) agente(s).

- Determinístico ou não determinístico: um ambiente determinístico é aquele no qual qualquer ação tem um único efeito garantido e previsível, não havendo incertezas sobre o estado que resultará da execução de uma ação. Ambientes não determinísticos, por outro lado, são imprevisíveis.

- Episódico ou sequencial: em ambientes episódicos as percepções e ações do agente são divididas em episódios atômicos, de modo que uma ação não influencia ou é influenciada por outra. Ambientes sequenciais, ao contrário, demandam ações que podem influenciar todas as decisões futuras.

- Estático ou dinâmico: ambientes estáticos são aqueles que podem permanecer inalterados até que o agente execute alguma ação. Ambientes dinâmicos são aqueles que possuem processos que podem causar mudanças no seu estado e que estão além do controle do agente.

- Discreto ou contínuo: ambientes discretos dispõe de um conjunto finito e bem limitado de ações e estados. O jogo de xadrez é um exemplo de ambiente discreto. Ambientes contínuos podem apresentar infinitos estados. Dirigir um carro pode ser considerado um ambiente contínuo. Esta distinção está ligada as formas de percepção e ação do agente.

Existem diferentes arquiteturas e modelos de AA que podem ser aplicados a ambientes e problemas distintos. As categorias nas quais AAs podem ser organizadas são agentes reativos, agentes baseados em lógica e agentes cognitivos. Os agentes reativos usam uma implementação mais simples, onde sensores e ações estão conectados de tal modo, que um dado estado sensorial leva sempre à mesma ação (NOLFI, 2002). Entretanto, Brooks (1991) apresenta uma abordagem de planejamento reativo na qual a soma dos comportamentos produz inteligência emergente. Esta abordagem busca seus princípios na biologia, etologia, cibernética e neurociência. Os agentes reativos possuem pouca autonomia e interagem com o ambiente por meio de sensores com comportamento reflexo e episódico. Em geral, a percepção-ação pode ser implementada por meio de uma tabela de regras condição-ação pré-definida. Esta abordagem inibe a aprendizagem, sendo necessário reconstruir a tabela toda vez que uma adaptação do agente for necessária (BERCHT, 2001).

Os agentes baseados em lógica, por sua vez, usam dedução lógica em seus processos de tomada de decisão (WOOLDRIDGE, 2001). Segundo Toni e Bentahar (2008), a lógica computacional tem sido usada para prover ferramentas e técnicas poderosas para abordar diferentes questões em contextos único e multiagentes. Desde a comunicação entre AA até linguagens de programação de agentes, argumentação e tomada de decisão.

Por fim, os agentes cognitivos – também chamados de deliberativos – são baseados em teorias de diferentes áreas como Psicologia, Filosofia e Biologia, para descrever o comportamento de indivíduos (FISHER *et al.*, 2002). Agentes desta categoria possuem conhecimento sobre seu ambiente e conhecimento pragmático sobre como usá-lo. Além disso, também possuem conhecimento sobre outros agentes e, assim, são capazes de manter um histórico de interações e ações passadas. Seu processo de deliberação pode ser implementado dentro de um enfoque mentalístico, que atribui ao agente estados mentais, dentre os quais, destacam-se: crenças, intenções, capacidades, objetivos, compromissos e expectativas (BERCHT, 2001; SHOHAM, 1993).

Entre os modelos supracitados, os agentes cognitivos são o de maior relevância para os fins deste trabalho. Por esta razão, maior foco será dado a esta categoria de AA. Sendo assim, segundo Braubach *et al.* (2005), entre as teorias mais influentes que implementam arquiteturas cognitivas podem-se destacar o modelo *Belief-Desire-Intention* (BDI) e a teoria da Programação Orientada a Agentes (AOP). Além disso, a arquitetura *Agent_Zero*, proposta por Epstein (2014), também pode ser incluída nesta categoria.

Uma outra forma de classificar os AA é quanto à sua funcionalidade de maior destaque ou ao seu domínio de maior atividade. Considerando estes critérios, pode-se dar destaque às classes de: agentes migrantes, agentes sociais, assistentes pessoais e agentes pedagógicos Bercht (2001). Agentes migrantes, segundo Bordini (1999) são AA capazes de migrar entre diferentes ambientes ou sociedades, visando cumprir com seus objetivos ou com os objetivos sociais da sociedade para a qual o agente emigrou. No contexto dos Sistemas Multiagentes (MAS[2]), os agentes migrantes podem desempenhar diferentes tarefas, como coletar e distribuir informações sobre outros agentes e ambientes, realizar ações sobre diferentes ambientes e compartilhar recursos Dorri *et al.* (2018).

Agentes sociais, por sua vez, são caracterizados por possuir habilidades de interação social com outros agentes Bercht (2001). Este tipo de AA tem sido aplicado em contextos de MAS (RUZZI *et al.*, 2017), na interação humano computador Holtgraves *et al.* (2007), Subagdja e Tan (2019) e no desenvolvimento de jogos sérios para aprendizagem mediante interação (AUGELLO *et al.*, 2016).

Os assistentes pessoais abrangem agentes capazes de desempenhar uma grande variedade de funções de suporte a outro agente biológico. Tais agentes são capazes de gerenciar tarefas, agendas, transações de vendas e processos de pagamento, sincronização de calendário e email, previsão do tempo, entre outras. Além disso, estes aplicativos são capazes, por exemplo, de processar consultas de um usuário contendo sons, imagens, textos e outras informações contextuais para responder à questões, ações e recomendações. Em geral assistentes pessoais também podem demonstrar capacidade de aprendizagem (SAPUTRA; MANONGGA, 2021).

Por fim, os agentes pedagógicos têm como objetivo "auxiliar os alunos ou aprendizes no processo de aprendizagem" (BERCHT, 2001, p.40). De acordo com Giraffa (1999), estes agentes podem ser de dois tipos: *goal-driven*, ou guiados por objetivo; e *utility-driven*, ou guiados pela utilidade. Os agentes *goal-driven* são capazes de realizar tarefas em colaboração ou competindo com os alunos. Os agentes *utility-driven* são aqueles que suportam os alunos em atividades diversas, como buscas de arquivos, agendamento de encontros de grupos, lembretes de compromissos e atividades a entregar, entre outras. Martha e Santoso (2019), entretanto, restringem agentes pedagógicos a personagens antropomórficos virtuais, usados em ambientes de aprendizagem online para servir a propósitos instrucionais. Este tipo de agente antropomórfico capaz de se comunicar com os estudantes e demonstrar emoções é mais amplamente conhecido

[2] Adotamos a sigla para o termo em inglês *Multi Agent Systems*, que é mais amplamente usada.

como agente pedagógico animado (LIN *et al.*, 2020).

Ao adotar-se uma abordagem que utiliza mais de um AA atuando em determinado ambiente, tem-se um Sistema Multiagente (SMA). Tais sistemas consistem em grupos de agentes que podem adotar papeis específicos em uma estrutura organizacional (BERCHT, 2001). Um padrão chave de interação em SMAs é, tanto em situações de cooperação como de competição, a coordenação orientada a objetivos e tarefas (WEISS, 1999). Além disso, questões relacionadas à robustez são de crucial importância em SMA (BALDONI *et al.*, 2020).

Segundo Alvares e Sichman (1997), os SMAs podem ser divididos em três categorias, sendo: Sistemas Multiagentes Reativos (SMAR), Sistemas Multiagentes Cognitivos (SMAC) e Sistemas Multiagentes Híbridos (SMH). Os SMARs, segundo o mesmo autor (1997), têm como características a não representação explícita de conhecimento, a não representação do ambiente, a não memorização das ações executadas, a organização etológica – mais semelhante à dos animais do que a das sociedades humanas –, e o grande número de membros. Com isso, o modelo de funcionamento dos agentes neste tipo de SMA é dado pelo padrão estímulo-resposta ou ação-reação.

Em uma SMAC, por outro lado, os agentes são capazes de manter uma representação explícita do ambiente, bem como de outros agentes da sociedade, são capazes de manter o histórico de suas ações e interações passadas, a comunicação entre os agentes é realizada de modo direto, por meio da troca de mensagens e o mecanismo de controle dos agentes é deliberativo. Além disso, a organização das sociedades nos SMAC é baseada em modelos sociológicos humanos. Tais sistemas, diferentemente dos SMAR são, geralmente, compostos por poucos agentes (ALVARES; SICHMAN, 1997).

No contexto de SMA, Alvares e Sichman (1997) identifica ainda cinco classes de agentes cognitivos, organizados de acordo com o seu nível de complexidade e conceitos explicitamente implementados em sua arquitetura. Tais classes, em ordem de complexidade decrescente, são:

- agentes organizados: mantém perspectivas múltiplas acerca de determinado problema e obedecem a regras e leis sociais;
- agentes negociantes: são capazes de resolver conflitos por meio de negociação;
- agentes intencionais: possuem representações internas de noções como intenções, comprometimento, objetivos, e planos parciais;

- agentes cooperativos: são dotados de representações mútuas uns dos outros e são organizados em um esquema de alocação de tarefas;
- módulos comunicantes: são estruturados sobre protocolos de comunicação;
- atores, processos: implementam primitivas de comunicação.

Por fim, SMAHs, como o nome sugere possuem em sua organização tanto agentes com estrutura e comportamento cognitivo como agentes com papéis que exigem um comportamento reativo (BERCHT, 2001). Com relação à comunicação entre os agentes em um SMA, o padrão mais amplamente usado é o especificado pela *Foundation for Intelligent Physical Agents* (FIPA) e se chama *Agents Communication Language* (ACL), podendo ser abreviado como FIPA-ACL. Esta linguagem é baseada na teoria dos atos de fala de Searle e Searle (1969) e é composta de diferentes sentenças que podem representar ações ou atos comunicativos. Apesar de ter sido criada na década de 1990, esta linguagem ainda é bastante utilizada atualmente (RăILEANU *et al.*, 2018).

As diferentes abordagens, arquiteturas e formas de classificação de AA apresentadas até agora, dão uma ideia de quão amplo e complexo é este campo de pesquisa. Existem ainda, uma infinidade de protocolos de negociação, plataformas para simulação de agente, argumentação multiagente, abordagens de planejamento multiagente, ferramentas e linguagens, entre outros (CARDOSO; FERRANDO, 2021), que, embora não tenham sido apresentados por não terem relação com este trabalho, tem ajudado a impulsionar o campo de pesquisa sobre AA. No próximo capítulo será apresentada com mais detalhe a arquitetura BDI, que consiste em um aparato teórico importante e amplamente utilizado para representar e raciocinar sobre os agentes inteligentes (CRUZ *et al.*, 2021).

2.1.2 Arquitetura BDI

O modelo BDI, segundo Cardoso e Ferrando (2021) aborda o comportamento autônomo de agentes por meio de duas teorias relacionadas do conceito filosófico de intencionalidade: a noção de um Sistema Intencional como uma entidade com crenças, desejos e outras atitudes proposicionais, proposta por Dennett (1989); e a teoria do Raciocínio Prático proposta por Bratman (1987), que se fundamenta em crenças, desejos e intenções na forma de planos parciais. Essas duas teorias de intencionalidade relacionadas, fornecem os fundamentos para descrição

de agentes em um nível apropriado de abstração em termos de crença, desejos e intenções (BDI). Com isso, é possível adotar a postura intencional e projetar agentes compatíveis com tais descrições, ou seja, como sistemas de raciocínio prático (CARDOSO; FERRANDO, 2021).

Dessa forma, segundo Georgeff *et al.* (1998), a arquitetura BDI, combina este modelo filosófico de raciocínio humano prático, uma série de implementações bem testadas, várias aplicações de sucesso e uma semântica lógica abstrata elegante, que foi adotada e elaborada amplamente na comunidade de pesquisa de agentes. O modelo BDI, resumidamente, consiste em uma abordagem baseada em três estados mentais, sendo: crenças, desejos e intenções, do inglês, respectivamente, *Beliefs, Desires* e *Intentions*.

De acordo com Georgeff *et al.* (1998), estes estados mentais podem ser descritos da seguinte forma:

- Crenças: Em termos de IA, crenças representam conhecimento sobre o mundo em que o agente se encontra, ou seja, seu ambiente. Em termos computacionais, elas são apenas uma representação do estado deste ambiente, seja na forma de valores em variáveis, bancos de dados relacionais ou expressões simbólicas em lógica de predicados.

- Desejos: em termos de IA representam o conjunto de objetivos de um agente, também podendo ser representados computacionalmente por meio de expressões simbólicas em alguma lógica, valores em variáveis ou estruturas de registro.

- Intenções: representam o compromisso que o AA assume com um plano de ações para atingir um dado resultado futuro. Constituem, na prática, um subconjunto de desejos selecionados de acordo com as crenças – situações do mundo – em um momento específico. Além disso, podem também ser caracterizadas como um estado de eventos a alcançar em seu ambiente. Computacionalmente, intenções podem ser um conjunto de threads em execução em um processo que pode ser interrompido à medida que feedbacks apropriados são recebidos do mundo em mudança.

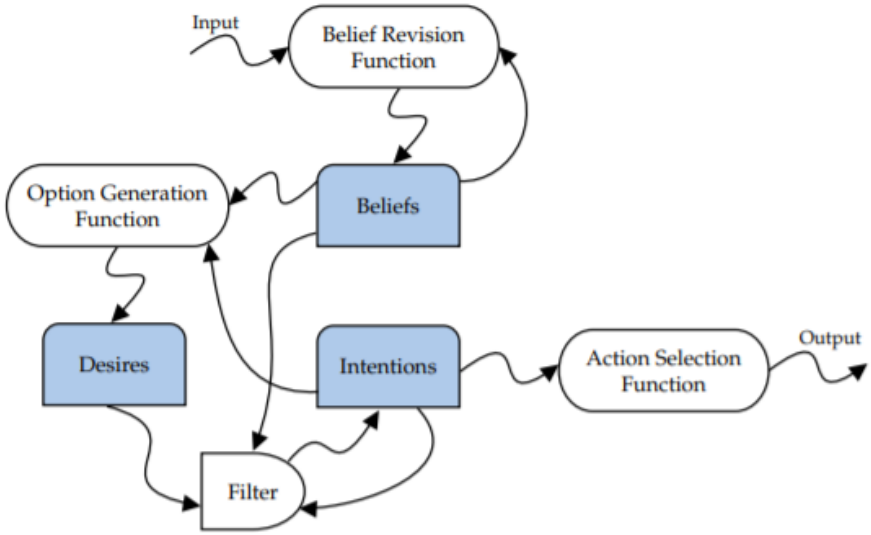
Nesse sentido Weiss (1999), vai um pouco mais além e descreve alguns componentes para o modelo BDI, sendo eles:

- Um conjunto de crenças que representam as informações que o agente possui sobre seu ambiente;

- Uma função de revisão de crenças capaz de analisar e atualizar as crenças do AA com base nos dados de entrada coletados do ambiente e em suas crenças atuais;
- Uma função geradora de opções, que determina as opções disponíveis para o agente executar, ou seja, seus desejos, com base em suas crenças e intenções atuais;
- Um conjunto de opções – desejos – que representam possíveis cursos de ação à disposição do agente;
- Uma função de filtro de admissibilidade que determina as intenções dos agentes, com base em suas crenças, desejos e intenções atuais;
- Um conjunto de intenções, representando o foco atual do agente – o estado de coisas que se comprometeu a alcançar;
- Uma função de seleção de ações, responsável por determinar a ação a ser executada pelo agente com base no atual conjunto de intenções.

A figura 2 apresenta a estrutura básica de componentes de uma arquitetura BDI genérica.

Figura 2 – Arquitetura BDI genérica.



Fonte: Cardoso e Ferrando (2021).

No modelo BDI, de forma geral, o raciocínio prático é dado por um processo de deliberação no qual o agente se compromete com uma intenção, a partir de suas próprias crenças, desejos e intenções anteriores. Além disso, o modelo também prevê o raciocínio meio-fim, por meio do qual o agente constrói um plano, ou seja, um conjunto de estados a serem alcançados sequencialmente, para atender a intenção escolhida (CRUZ *et al.*, 2021).

Esta arquitetura, bem como sua formalização, tal qual aqui apresentada, foi proposta por Rao e Georgeff (1995). Neste modelo, é importante destacar o papel da intenção, que segundo Weiss (1999), desempenha um papel fundamental. As intenções são selecionadas pelo agente e não pelo projetista, que deverá definir apenas os desejos e as crenças iniciais. A fonte para a seleção de intenções são os desejos e outras intenções, considerando sempre, as crenças do agente e suas possibilidades de ação. Intenções são compromissos que o agente assume com um estado futuro possível. O agente deve crer na possibilidade de satisfazer esta intenção e planejará ações para realizá-la. O agente se comprometerá com as intenções que assumiu até que elas se mostrem já alcançadas ou impossíveis de alcançar. Por fim, o agente pode ter mais de uma intenção, mas elas não podem ser contraditórias.

Ainda hoje a arquitetura BDI é o modelo mais popular para a construção de agentes deliberativos, sendo usado em sete de quinze novas linguagens de programação para agentes, como mostra uma revisão sistemática publicada por Cardoso e Ferrando (2021). É possível que a principal contribuição da abordagem baseada em BDI seja a inteligibilidade e a previsibilidade do comportamento do agente (CARDOSO; FERRANDO, 2021). Esta característica, conhecida como Inteligência Artificial Explicável (XAI[3]), possibilita o desenvolvimento de sistemas autônomos capazes de explicar suas decisões e ações aos usuários humanos (GUNNING; AHA, 2019).

Segundo Cardoso e Ferrando (2021), isto acontece porque a abordagem baseada em intenções torna mais intuitivo, para os usuários finais, o porquê de um agente estar fazendo o que está fazendo, bem como prever as suas próximas ações. Além disso, a forma com a qual agentes BDI são estruturados torna mais fácil a extração de respostas que justifiquem o porque das ações escolhidas, uma vez que se baseiam em intenções devidamente fundamentadas pelo processo de raciocínio do agente. Esta relativa transparência dos programas do agente BDI facilita também a implementação de programas capazes de raciocinar sobre os aspectos éticos de suas decisões.

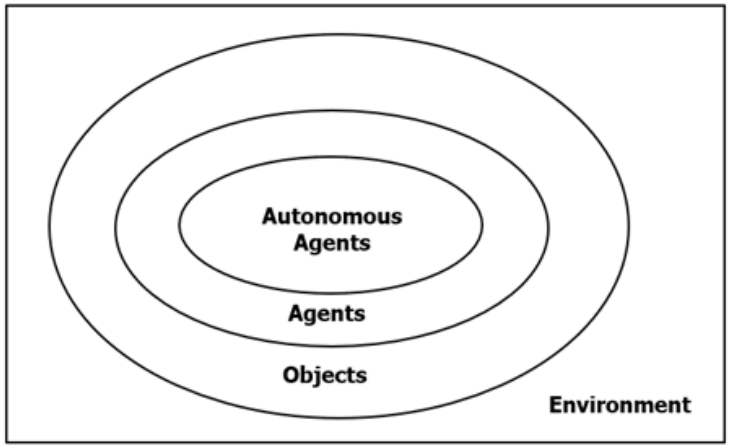
2.1.3 Autonomia em Agentes

Conforme demonstrado no capítulo 2.1.2, a autonomia é uma característica central para muitos autores, no que diz respeito à definição de agentes. Para Luck e d'Inverno (1995), entretanto, não faz sentido considerar que todo o agente é autônomo. Segundo estes autores, existe uma hierarquia de entidades, conforme apresentado na figura 3. Nesta hierarquia, pode-se

[3] Adotamos a sigla XAI para o termo em inglês *Explainable Artificial Intelligence*, que é mais amplamente usada

identificar: o ambiente com o qual os agentes interagem; os objetos, que abrangem o conjunto de todas as entidades conhecidas no ambiente; os agentes, que estão contidos no conjunto de objetos; e, por fim, os agentes autônomos, que estão contidos no conjunto de todos os agentes.

Figura 3 – Hierarquia de entidades.



Fonte: Luck e d'Inverno (1995).

Nesta abordagem, o que torna um agente autônomo e, portanto, diferente de outros agentes encontrados em um dado ambiente, é a sua capacidade de gerar os próprios objetivos por meio da auto motivação. A motivação, por sua vez, pode ser caracterizada como qualquer desejo ou preferência que leve a geração ou adoção de objetivos, e que afete os resultados das tarefas de raciocínio ou comportamento do agente. Sendo assim, agentes motivados são aqueles que seguem sua própria agenda de raciocínio e comportamento de acordo seus desejos ou preferências internas, sendo agentes autônomos, portanto, aqueles que possuem o seu próprio conjunto de motivações (LUCK; D'INVERNO, 1995).

Para Castelfranchi (2000), a autonomia é derivada da arquitetura e das teorias de ação do agente. Nesse sentido, ela pode ocorrer em dimensões e graus diferentes, a depender do contexto no qual o agente está inserido. Esta abordagem traz a ideia de que o comportamento autônomo não deve ser completamente determinado e previsível com base nas entradas que o agente recebe "como uma bola de bilhar sob forças mecânicas" (CASTELFRANCHI, 2000, p.1).

Ser autônomo, significa, de modo geral, não ser hétero-dirigido, ou seja, não ser determinado ou dirigido de fora. No caso dos agentes, isto significa não ser uma entidade causal, mas, no mínimo, orientado por objetivos (ROSENBLUETH; WIENER, 1950; CASTELFRANCHI, 2000). Nesse sentido, dizer que autonomia está relacionada à arquitetura interna e às teorias de ação do agente, significa que ele deve dispor de meios, estruturas e recursos capazes de possibilitar a execução de ações para atingir seus objetivos, sem depender da ação de outros

agentes (CASTELFRANCHI *et al.*, 1992). E dizer que a autonomia pode ocorrer em diferentes graus e dimensões, significa que determinado agente pode ser autônomo com relação a uma ação, objetivo ou outro agente, mas não com relação a outros (CASTELFRANCHI, 2000) (CASTELFRANCHI, 2000).

Quanto às dimensões da autonomia, Castelfranchi (2000) define treze pontos relacionados à dependência do agente para agir, sendo:

1. Autonomia/dependência de informação: caracteriza se um agente depende de outro agente para obter informações sobre o ambiente a fim de conseguir executar suas ações;
2. Autonomia/dependência de interpretação: especifica se um agente é capaz e tem permissão para interpretar, por si só, os dados ou informações sensoriais;
3. Autonomia/dependência de know-how: especifica se o agente é capaz de executar um plano sem a necessidade de delegar atividades ou consultar outro agente sobre como executá-lo;
4. Autonomia/dependência de plano-arbítrio: especifica se um agente tem capacidade e prerrogativa de escolher um plano para outro agente;
5. Autonomia/dependência de planejamento: especifica se um agente é autorizado e capaz de elaborar um plano e executá-lo;
6. Autonomia/dependência motivacional: especifica se um agente tem objetivos próprios, não dependendo de outro para saber o que perseguir;
7. Autonomia/dependência da dinâmica de metas: especifica se um agente é capaz e autorizado a suspender um determinado plano ou abandonar uma determinada intenção ou de trocar suas preferências para responder a possíveis mudanças;
8. Autonomia/dependência para critérios de objetivo: especifica se um agente é capaz e tem o direito de escolher entre diferentes objetivos ou tarefas próprias, ou seja, se ele pode exercer suas preferências;
9. Autonomia/dependência de raciocínio: especifica se o agente é capaz e permitido a fazer suas próprias inferências, raciocinar e confiar em suas conclusões.
10. Autonomia/dependência para monitorar as próprias ações: especifica se ‘um agente depende de outro para monitorar e avaliar suas próprias ações – se foram ou não bem-sucedidas ou corretas.

11. Autonomia/dependência de habilidades: especifica se um agente depende de outro para dispor de alguma opção de ação ou habilidade necessária para executar seu planejamento.
12. Autonomia/dependência de recursos: especifica se o agente depende de outro agente para ter acesso a recursos materiais necessários à execução dos seus planos.
13. Autonomia/dependência para habilitar ou condicionar suas ações: especifica se um agente depende de outro agente para habilitar ou permitir as condições necessárias à execução de suas ações.

Ainda com relação às dimensões da autonomia, Rieder *et al.* (2020), argumentam que é preciso considerar o escopo, o caráter e o grau no qual um agente pode agir. Nesse contexto, o escopo diz respeito ao espaço ou domínio dentro do qual um agente é capaz ou autorizado a exercer sua autonomia, bem como o conjunto de escolhas ou ações disponíveis. O caráter, por sua vez, diz respeito às capacidades, poderes e habilidades que compõe suas possibilidades de ação. Entre tais potencialidades destacam-se a capacidade de compreender e tomar decisões justificadas, a capacidade de autogoverno, de formar e executar planos, de selecionar suas próprias metas, de se comprometer com outros agentes ou projetos e a capacidade de tomar decisões. Finalmente, o grau de autonomia se refere à habilidade de um agente para utilizar suas capacidades dentro de um determinado escopo (RIEDER *et al.*, 2020)

A autonomia é um conceito complexo, sendo composto por uma coleção de atributos, capacidades, habilidades e poderes. Nesse sentido, fatores como questões legais, políticas e éticas podem influenciar na autonomia dos agentes (RIEDER *et al.*, 2020). As questões éticas, em especial, têm sido bastante consideradas nos últimos anos, a fim de viabilizar agentes autônomos alinhados a valores humanos (CóRDOVA *et al.*, 2021).

2.2 ÉTICA EM INTELIGÊNCIA ARTIFICIAL

No campo da inteligência artificial, pesquisas têm sido direcionadas, desde a sua origem na década de 1950, para questões relacionadas ao fato de poder ou não desenvolvê-la. Este foco implicou em diferentes dimensões de investigação, como: o que se pode desenvolver neste campo, como se pode desenvolver e quando se pode desenvolver, possibilitando, com isso, grandes avanços tecnológicos nesta área. Entretanto, há outra questão que também precisa ser considerada com igual interesse: a que diz respeito ao fato de dever ou não desenvolver IA (RUSSEL; NORVIG, 2013).

É notório que, na medida em que sistemas de IA – por exemplo: *chatbots*, robôs, sistemas de recomendação, entre outros agentes inteligentes –, estão deixando de ser vistos como simples ferramentas para serem percebidos como agentes dotados de autonomia, companheiros de equipe ou auxiliares nas mais diferentes tarefas, um novo e importante foco de pesquisa é entender o impacto ético destes sistemas na sociedade (DIGNUM, 2018). Para Aliman e Kester (2019), esta têm sido uma questão urgente e de relevância internacional.

Não se trata, contudo, de especulações distópicas presentes em obras de ficção científica, nas quais as máquinas dominariam o mundo e acabariam com a vida humana. Trata-se, sim, de situações práticas nas quais sistemas de IA já estão mudando a rotina das pessoas. Este é o caso, por exemplo da área de transportes, onde o desenvolvimento de veículos autônomos têm ganhado grande destaque (THORNTON *et al.*, 2017), da área da saúde (REDDY *et al.*, 2020), educação (VICARI, 2021; BERENDT *et al.*, 2020), aplicativos móveis inteligentes (SARKER *et al.*, 2021), militar (SVENMARCK *et al.*, 2018), setor público e segurança pública (KANKANHALLI *et al.*, 2019), entre outras.

Nesse sentido, a partir do momento em que se podem observar sistemas de IA tomando decisões e guiando ou influenciando decisões humanas, algumas questões devem ser levantadas, por exemplo: quais são as consequências morais, legais e sociais destas decisões?; um sistema de IA pode ser responsabilizado por tais decisões?; como estes sistemas podem ser controlados se a sua capacidade de aprendizagem for capaz de mudar seu estado em relação ao projeto inicial?; este tipo de inovação deveria sequer ser permitida?. Os meios de que a sociedade dispuser para lidar com estas questões irá determinar, em grande parte, o quanto será possível confiar e coexistir de forma segura com estes sistemas (DIGNUM, 2018).

Assim, a emergente necessidade de considerar questões éticas acerca do desenvolvimento e aplicação de sistemas inteligentes e interativos, têm levado, devido a sua relevância, a diferentes iniciativas, entre as quais se destacam: IEEE *Global Initiative on Ethics of Autonomous System* [4], a *Foundation for Responsible Robotics*[5] e a *Partnership for AI* [6] (DIGNUM, 2018). Além disso, diferentes esforços governamentais e intergovernamentais têm sido direcionados para a promoção de debates, visando a regulamentação da área. Nesse contexto, merecem destaque os princípios de Asilomar[7], que trazem um conjunto de diretrizes que os pesquisadores deste campo devem respeitar, os princípios da IA criados pela Organização para a Cooperação

[4] http://ethicsinaction.ieee.org
[5] http://responsiblerobotics.org
[6] http://www.partnershiponai.org
[7] https://futureoflife.org/ai-principles

e Desenvolvimento Econômico (OCDE)[8] e, por fim, ações da *United Nations Educational, Scientific and Cultural* (UNESCO)[9] que visam produzir um código de ética global para pesquisa em IA.

Essas iniciativas evidenciam a complexidade e a abrangência deste tema. Por esta razão, para melhor organizar a discussão em torno do assunto, segundo (DIGNUM, 2018), a ética em IA pode ser compreendida a partir das seguintes dimensões:

- Ética por Design: termo oriundo do inglês, *ethics by design*, que trata da integração técnica e/ou algorítmica de capacidades de raciocínio ético como parte do comportamento de sistemas autônomos artificiais;
- Ética em Design: do inglês, *ethics in design*, abrange os métodos regulatórios e de engenharia que apoiam a análise e avaliação das implicações éticas dos sistemas de IA à medida que integram ou substituem estruturas sociais tradicionais;
- Ética para o Design: traduzido do termo *ethics for design* implica os códigos de conduta que tentam garantir a integridade de desenvolvedores e usuários na pesquisa, projeto, construção e emprego de sistemas dotados de IA.

O presente trabalho estará focado na dimensão que investiga soluções para implementação de alternativas capazes de dotar agentes artificiais de comportamentos considerados éticos, ou seja, em ética por design. Esta abordagem de estudo também recebe o nome de ética de máquina, que não deve ser confundida com ética computacional, que está mais focada no uso que o ser humano faz das máquinas (NATH; SAHU, 2020). Na literatura vigente ainda é possível encontrar outras denominações que caracterizam esta dimensão de análise, entre elas: “moralidade de máquina, moralidade artificial, moralidade computacional, roboética, e inteligência artificial amigável” (CERVANTES *et al.*, 2020, p.3, tradução nossa).

Este campo de investigação suscita novos desafios à pesquisadores de diferentes áreas do conhecimento. Entre estes desafios destacam-se: como transformar princípios éticos em modelos computacionais; como evitar enviesamento de dados que implicariam na replicação de preconceitos humanos e; como possibilitar que agentes inteligentes possam lidar com dilemas éticos (CóRDOVA *et al.*, 2021). Estas questões visam tornar os sistemas baseados em IA mais previsíveis, confiáveis e, consequentemente, mais seguros. Todavia, nas palavras de Dignum

[8] https://www.oecd.org/going-digital/ai/principles
[9] https://pt.unesco.org/courier/2018-3/em-direcao-um-codigo-etica-global-pesquisa-em-inteligencia- artificial

et al. (2018), para que estes sistemas possam ser explorados em todo o seu potencial de forma segura, é preciso mais do que implementar melhorias a nível de percepção, nos algoritmos de busca ou em seu poder computacional. É necessário que estes sistemas estejam alinhados com valores morais e princípios éticos humanos.

O uso do termo alinhado não é aleatório. Em 2014, Soares e Fallenstein (2014) propuseram o uso deste termo para fazer referência a uma IA construída de tal forma que fosse capaz de garantir que seu comportamento seja sempre benéfico ao ser humano, o que significa que a IA estaria alinhada aos interesses humanos. Nesse sentido é que, atualmente, o conjunto de esforços para construir soluções de IA capazes de considerar interesses humanos em seus processos de tomada de decisão está organizado sob o termo amplamente conhecido como Alinhamento de Valores (VA[10]) em inteligência artificial (KIM *et al.*, 2019).

Dessa forma, VA em IA pode ser definido como a "tentativa de implementar sistemas aderentes a valores éticos humanos" (ALIMAN; KESTER, 2019, p.1). Trata-se, pois, de uma área bastante complexa, uma vez que exige a abordagem de questões éticas, que, por sua vez, são também sempre complexas, principalmente devido à dependência de contexto, tempo e cultura, além da subjetividade de julgamento e do ponto de vista do observador (CóRDOVA *et al.*, 2021). Além disso, a falta de consenso entre os filósofos da moral sobre qual teoria ética deveria ser seguida, também pode ser considerado um obstáculo para o desenvolvimento de máquinas éticas (BOSTROM, 2014).

Nesse sentido, nos próximos tópicos serão apresentadas as principais doutrinas e fundamentos éticos considerados no campo do VA. Na sequência, será abordado como estas doutrinas são usadas pelas propostas de solução para o desenvolvimento de agentes inteligentes alinhados à valores éticos.

2.2.1 Principais Doutrinas Éticas para Alinhamento de Valores em IA

A partir de uma perspetiva histórico-filosófica, pode-se afirmar que a palavra ética tem sua origem na língua grega, vocábulo *ethos*, tendo sido usada primeiramente por Aristóteles em seu livro, Ética a Nicômaco, e significando costume ou prática comum (ANDINO, 2015). Além disso, é importante destacar que o vocábulo, moral, que tem sua origem no latim, mores, é equivalente à palavra grega ethos (ANDINO, 2015). Entretanto, para fins de organização deste estudo, será adotado como significado para ética, aquele definido por Cervantes *et al.* (2020), que

[10] Do termo em inglês: Value Alignment.

estabelece a ética como a disciplina da filosofia que estuda a dimensão moral dos seres humanos.

Nesse contexto, existem duas abordagens a serem consideradas ao tratar deste tema, a metaética e a ética normativa. Para Schroeder (2017), a metaética se preocupa com o significado dos julgamentos morais, tendo seu foco centrado em compreender a natureza das propriedades éticas, das declarações e dos julgamentos éticos, bem como em entender aquilo que os fundamenta. Por outro lado, a ética normativa está mais preocupada com a articulação e a explicação de princípios fundamentais sobre o que é certo ou errado, ou seja, em princípios que estabelecem diretivas sobre como as pessoas devem agir e o que moralmente devem fazer (SCHROEDER, 2017).

A identificação de doutrinas e modelos éticos para servir de guia ao comportamento humano tem sido um dos temas mais importantes do pensamento filosófico (VAMPLEW *et al.*, 2018). Esta abordagem caracteriza a ética normativa e é o ponto de maior relevância para o desenvolvimento da ética de máquina. Sendo assim, com o intuito de limitar o escopo deste trabalho às doutrinas éticas de maior relevância para o alinhamento de valores em IA, serão descritas a seguir, com maior profundidade, as doutrinas deontológicas e teleológicas, bem como algumas de suas variações.

A ética deontológica sustenta que uma dada ação deve ser julgada com base na sua compatibilidade com um conjunto de deveres reconhecidos como legítimos pelos tomadores de decisões racionais (VAMPLEW *et al.*, 2018). Estruturas deontológicas, portanto, são orientadas à deveres (PFORDTEN, 2012). O imperativo categórico de Immanuel Kant, por exemplo diz "aja conforme a uma máxima que possa valer ao mesmo tempo como uma lei universal" (KANT, 2013, p.30). Este imperativo estabelece um princípio geral a ser seguido, vinculando o julgamento moral à intenção de quem age. Por essa razão, esta abordagem também pode ser chamada de ética intencionalista.

Com um sentido um pouco diferente, Ross (1930) propõe uma estrutura baseada em uma lista de sete deveres *prima facie*, sendo eles: fidelidade, reparação, gratidão, não maleficência, justiça, beneficência e auto melhoramento (VAMPLEW *et al.*, 2018)). Nesse contexto, um tomador de decisão deve tentar satisfazer a todos os deveres, e, em caso de conflito, deve sopesar a importância de cada dever e decidir sobre qual a melhor ação. De todo modo, um modelo deontológico é focado sempre na ação em si e não em seus resultados.

Por outro lado, a ética teleológica é baseada na noção de que a moralidade de uma ação deve ser julgada pelas consequências que dela advém. "Assim, uma ação, em si não pode ser

negativa ou positiva, mas os resultados e os impactos que dela advém podem ser bons ou ruins" (BAUMANE-VITOLINA *et al.*, 2016, p.110, tradução nossa). O utilitarismo é um exemplo clássico de ética teleológica.

Na abordagem utilitarista, assume-se que a desejabilidade de uma ação pode ser medida por métricas de utilidade e que uma ação é julgada moralmente correta se as suas consequências levam à melhor utilidade possível (VAMPLEW *et al.*, 2018). Esta doutrina ética pode ser dividida em utilitarismo baseado na ação e utilitarismo baseado em regras. O primeiro visa selecionar uma entre várias ações possíveis, escolhendo aquela capaz de maximizar a utilidade em uma dada situação. O segundo permite identificar regras capazes de levar aos melhores resultados (VAMPLEW *et al.*, 2018).

Existem diferentes perspetivas quanto à definição de utilidade que se pretende maximizar. Bentham (1988), por exemplo, propôs que deve-se buscar a maximização da felicidade total de uma população de pessoas, em uma abordagem que ficou conhecida como utilitarismo hedonista. Wallach e Allen (2009), sugerem que devem ser consideradas e combinadas, múltiplas escalas de utilidade em uma só fórmula de ponderação. Contudo, não há consenso sobre qual é a maneira correta de ponderar diferentes fontes de utilidade, ou mesmo, se é conveniente combiná-las (VAMPLEW *et al.*, 2018).

Diferente das duas doutrinas anteriormente apresentadas, a ética da virtude, proposta por Platão e Aristóteles, baseia-se no desenvolvimento e aperfeiçoamento de certo conjunto de virtudes (CASAS-ROMA; ARNEDO-MORENO, 2019). Para Annas (2006), virtude é um estado ou disposição de uma pessoa. Nesse sentido, não pode ser confundida com o hábito, que pode ser inconsciente. A virtude é uma disposição para agir por alguma razão justificada e, portanto, exercida por meio do raciocínio prático. Nesse sentido, o julgamento moral não se dá com base na ação em si, nem nas suas consequências, mas no quanto esta ação representa ou responde por alguma virtude, como generosidade, bondade, justiça, compaixão, entre outras (CASAS-ROMA; ARNEDO-MORENO, 2019).

É importante ainda, no contexto da ética da virtude, citar um outro modelo dela derivado, conhecido como exemplarismo ou ética exemplarista. Este modelo foi proposto por Zagzebski (2010) e é fundamentada em exemplos de boa moral. Sendo assim, segundo esta autora, exemplos de boa moral são agentes admiráveis, cujos exemplos são dignos de serem seguidos. Nesse modelo, uma ação é julgada boa se é baseada em exemplos dignos de admiração moral.

Por fim, trazendo elementos das abordagens deontológica e utilitarista, a doutrina do

duplo efeito introduz o conceito de causalidade e proporcionalidade (McIntyre, 2019). Geralmente atribuída à São Tomás de Aquino, esta doutrina estabelece um conjunto de condições necessárias para lidar com situações com as quais um resultado moralmente questionável é previsível (QUINN, 1989). Estas condições são:

> (a) a intenção final pretendida deve ser boa; (b) os meios pretendidos para isso devem ser moralmente aceitáveis; (c) o desfecho ruim previsto não deve ser desejado (isto é, não deve ser, em algum sentido, pretendido); e (d) o bom final deve ser proporcional ao resultado ruim (ou seja, deve ser importante o suficiente para justificar o resultado ruim) (QUINN, 1989, p.334, tradução nossa).

Trata-se de uma estrutura de princípios bastante úteis para lidar com dilemas éticos, que serão melhor abordados posteriormente. Além disso, esta doutrina tem sido bastante usada em leis de guerra (BONNEMAINS *et al.*, 2018).

Estas são as doutrinas ou modelos éticos mais recorrentemente encontrados na literatura e relacionados ao alinhamento de valor em IA (VAMPLEW *et al.*, 2018). Contudo, devido à falta de consenso, adequação e padronização sobre qual seria a melhor abordagem, alguns pesquisadores propõem modelos mais pragmáticos. Estas alternativas são, em geral, mais limitadas e baseadas em restrições para um domínio específico ou adequadas a cenários éticos mais restritos. Este é o caso, por exemplo, de estruturas legais específicas, regras de segurança e códigos militares (VAMPLEW *et al.*, 2018). O próximo capítulo tratará com mais detalhes das iniciativas práticas que tem o intuito de orientar o desenvolvimento de tecnologias de IA considerando as implicações éticas da área.

2.2.2 Princípios e Regulamentações para Ética em IA

Conforme descrito no capítulo 2.2, dada a importância e necessidade de estabelecer limites éticos à utilização da IA, várias iniciativas foram sendo organizadas desde 2017 ao redor do mundo. Neste capítulo serão apresentadas algumas delas, buscando identificar os princípios éticos que devem nortear o desenvolvimento da área, sobretudo no que diz respeito à sua aplicação na educação.

Contudo, é importante introduzir, a fim de possibilitar uma melhor compreensão, alguns conceitos importantes que serão tratados dentro das diferentes propostas abordadas neste trabalho. O primeiro deles diz respeito ao **ciclo de vida de sistemas de IA**, que deve ser entendido como os estágios de pesquisa, projeto, desenvolvimento, entrega e uso, incluindo manutenção, operação,

negócios, financiamento, monitoramento, avaliação, fim de uso, descontinuação e desativação de soluções baseadas em tecnologias de IA (UNESCO, 2020). Outro conceito importante é o de **atores de IA**, grupo do qual fazem parte quaisquer atores envolvidos em pelo menos um dos estágios do ciclo de vida de sistemas de IA, podendo ser pessoas naturais ou jurídicas (UNESCO, 2020).

Por fim, cabe limitar o escopo de tecnologias ao qual se aplica o conjunto de princípios éticos apresentados no decorrer deste capítulo. Para isto, adotar-se-á a definição da UNESCO (2020) para sistemas de IA como sendo "sistemas tecnológicos que têm a capacidade de processar informações de uma forma que se assemelha ao comportamento inteligente e, normalmente, inclui aspectos de raciocínio, aprendizagem, percepção, previsão, planejamento ou controle" (UNESCO, 2020, p.4). Além disso, também se incluem neste escopo, os Sistemas Autônomos Inteligentes (A/IS[11]), tratados na proposta do IEEE (2019).

Nesse sentido, pode-se apontar o *Future of Life Institute* (FLI), que tem entre seus fundadores Stephan Hawking, Max Tegmark e Stuart J. Russel, como entidade autora de uma das iniciativas pioneiras neste campo. Em janeiro de 2017, o FLI organizou a *Beneficial AI Conference* em Asilomar, na Califórnia, que resultou em vinte e três princípios conhecidos como Asilomar *AI Principles* (ENGELS *et al.*, 2018). O documento resultante divide os vinte e três princípios em três categorias, sendo: questões de pesquisa; ética e valores; e questões de longo prazo.

Assim, visando tratar daqueles que tem maior relevância para os fins do presente trabalho, é possível destacar, da categoria questões de pesquisa, os seguintes itens:

- Objetivo da pesquisa: O objetivo da pesquisa em IA deve ser criar não inteligência não direcionada, mas inteligência benéfica.
- Vínculo Ciência-Política: Deve haver um intercâmbio construtivo e saudável entre pesquisadores de IA e formuladores de políticas.
- Cultura de pesquisa: Uma cultura de cooperação, confiança e transparência deve ser promovida entre pesquisadores e desenvolvedores de IA.
- Evitar corridas: As equipes que desenvolvem sistemas de IA devem cooperar ativamente para evitar falhas nos padrões de segurança (FLI, 2021).

[11] Neste trabalho foi dotada a sigla para o nome em inglês Autonomous Intelligent Systems (A/IS) por ser este mais amplamente conhecido.

Da categoria éticas e valores, por sua vez, podem ser destacados os princípios:

- Segurança: os sistemas de IA devem ser seguros e protegidos ao longo de sua vida operacional e verificável quando aplicável e viável.
- Transparência quanto a falhas: Se um sistema de IA causar danos, deve ser possível determinar o porquê.
- Responsabilidade: Designers e construtores de sistemas avançados de IA são partes interessadas nas implicações morais de seu uso, mau uso e ações, com a responsabilidade e a oportunidade de moldar essas implicações.
- Alinhamento de valores: sistemas de IA altamente autônomos devem ser projetados de forma que seus objetivos e comportamentos possam ter a garantia de estar alinhados com os valores humanos ao longo de sua operação.
- Valores humanos: os sistemas de IA devem ser projetados e operados de forma a serem compatíveis com os ideais de dignidade humana, direitos, liberdades e diversidade cultural.
- Privacidade pessoal: as pessoas devem ter o direito de acessar, gerenciar e controlar os dados que geram, dado o poder dos sistemas de IA de analisar e utilizar esses dados.
- Liberdade e privacidade: A aplicação de IA aos dados pessoais não deve restringir injustificadamente a liberdade real ou percebida das pessoas.
- Benefício compartilhado: as tecnologias de IA devem beneficiar e capacitar o maior número possível de pessoas.
- Prosperidade compartilhada: A prosperidade econômica criada pela IA deve ser amplamente compartilhada, para beneficiar toda a humanidade.
- Controle Humano: os humanos devem escolher como e se delegam decisões aos sistemas de IA, para cumprir os objetivos escolhidos pelos humanos (FLI, 2021).

Por fim, da categoria problemas de longo prazo, apresentam-se os itens:

- Cuidado de capacidade: Não havendo consenso, devemos evitar suposições fortes sobre os limites superiores das capacidades futuras de IA.

- Importância: IA avançada pode representar uma mudança profunda na história da vida na Terra e deve ser planejada e administrada com cuidado e recursos adequados.

- Riscos: Riscos apresentados por sistemas de IA, especialmente riscos catastróficos ou existenciais, devem estar sujeitos a planejamento e esforços de mitigação proporcionais ao seu impacto esperado.

- Autoaprimoramento recursivo: os sistemas de IA projetados para se auto aperfeiçoar recursivamente ou se autorreplicar de maneira que possa levar a um rápido aumento da qualidade ou quantidade devem estar sujeitos a medidas rígidas de segurança e controle.

- Bem comum: a superinteligência só deve ser desenvolvida a serviço de ideais éticos amplamente compartilhados e para o benefício de toda a humanidade, e não de um estado ou organização (FLI, 2021).

Tais princípios, segundo Engels *et al.* (2018), são fortemente comprometidos com o cumprimento da Convenção das Nações Unidas sobre Direitos Humanos. Entretanto, apesar de serem genéricos o bastante para possibilitar uma ampla aplicação, possuem termos que necessitam de definições jurídicas para que possam ser de fato aplicados.

Por sua vez, com a semelhante preocupação acerca dos impactos éticos da IA, o IEEE propõe um conjunto de oito princípios gerais de alto nível para o desenvolvimento e operação de tecnologias de IA (IEEE, 2019). Estes princípios devem, segundo seus criadores, promover os valores humanos e garantir a confiabilidade da IA. De forma sintética, os princípios propostos pelo IEEE (2019), são:

- Direitos humanos: os A/IS devem ser criados e operados para respeitar, promover e proteger os direitos humanos internacionalmente reconhecidos.

- Bem-estar: desenvolvedores de A/IS devem adotar a promoção do bem-estar humano como o principal critério de sucesso.

- Agência de Dados: desenvolvedores de A/IS devem capacitar os indivíduos para acessar e compartilhar seus dados com segurança, visando manter a capacidade humana de ter controle sobre sua identidade.

- Eficácia: desenvolvedores e operadores de A/IS devem fornecer evidências de que seus sistemas são eficazes e adequados aos propósitos para os quais foram criados.

- Transparência: a base de uma decisão tomada por um A/IS particular deve sempre ser detectável.
- Responsabilidade: A/IS devem ser criados e operados para fornecer uma justificativa inequívoca para todas as decisões por eles tomadas.
- Conscientização sobre o uso indevido: desenvolvedores de A/IS devem se proteger contra todos os possíveis usos indevidos e riscos de A/IS em operação.
- Os criadores da Competência: desenvolvedores de A/IS devem especificar e os operadores devem aderir aos conhecimentos e habilidades necessários para uma operação segura e eficaz.

Estes princípios foram publicados em um relatório do IEEE chamado *ETHICALLY ALIGNED DESIGN: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition*, em 2019. O objetivo deste relatório criado no âmbito do IEEE *Global Initiative on Ethics of Autonomous and Intelligent Systems ("The IEEE Global Initiative")*, é oferecer um documento que "forneça percepções e recomendações pragmáticas e direcionais, servindo como uma referência chave para o trabalho de tecnólogos, educadores e legisladores nos próximos anos" (IEEE, 2019, p.2).

Da mesma forma, com o objetivo de promover o uso inovador e confiável da IA, que respeite os direitos humanos e os valores democráticos, a *Organisation for Economic Cooperation and Development* (OECD) propôs em 2021 um conjunto de princípios e recomendações para uma gestão responsável e confiável de IA. Tais princípios, conforme descritos em OECD (2021), estão listados a seguir:

- Crescimento inclusivo, desenvolvimento sustentável e bem-estar: as partes interessadas no desenvolvimento de soluções de IA devem se engajar proativamente na gestão responsável da IA confiável, buscando aumentar as capacidades e a criatividade humana, promover a inclusão, reduzir as desigualdades econômicas, sociais, de gênero e outras e proteger os ambientes naturais, estimulando, com isso, o crescimento inclusivo, o desenvolvimento sustentável e o bem-estar.
- Valores centrados no ser humano e justiça: os atores de IA devem respeitar o estado de direito, os direitos humanos e os valores democráticos em todo o ciclo de vida do sistema de IA. Isso inclui liberdade, dignidade e autonomia, privacidade e proteção de dados,

não discriminação e igualdade, diversidade, justiça, justiça social e direitos trabalhistas internacionalmente reconhecidos.

- Transparência e explicabilidade: os atores de IA devem se comprometer com a transparência e a divulgação responsável em relação aos sistemas de IA, visando promover uma compreensão geral de tais sistemas, conscientizar as partes interessadas sobre suas interações com eles, inclusive no local de trabalho, permitir que aqueles afetados por um sistema de IA compreendam o resultado e possam desafia-lo com base em informações simples e fáceis de entender sobre os fatores e a lógica que serviu de base para a previsão, recomendação ou decisão.

- Robustez, segurança e proteção: os sistemas de IA devem ser robustos, seguros e protegidos durante todo o seu ciclo de vida, de modo que possam ser executados em condições normais e adversas sem que ofereçam riscos excessivos. Além disso os atores de IA devem garantir a rastreabilidade dos dados, processos e decisões tomadas e aplicar uma abordagem de gerenciamento de riscos relacionados aos sistemas de IA, incluindo privacidade segurança digital e proteção contra vieses;

- Responsabilidade: os atores de IA devem ser responsáveis pelo funcionamento adequado dos sistemas de IA e pelo respeito aos princípios acima descritos, com base em suas funções no contexto e de acordo com o estado da arte.

Por fim, entre as iniciativas transcontinentais de relevância para este trabalho no que diz respeito ao estabelecimento de princípios éticos para IA, pode-se citar o *First Draft of the Recommendation on the Ethics of Artificial Intelligence*, desenvolvido pela *United Nations Educational Scientific and Cultural* (UNESCO). Esta proposta "aborda a ética da IA como uma reflexão normativa sistemática, baseada em uma estrutura holística de valores que evoluem e princípios e ações interdependentes capazes de orientar as sociedades sobre como lidar com os impactos conhecidos e desconhecidos das tecnologias de IA nos seres humanos, nas sociedades, no meio ambiente e ecossistemas, com responsabilidade [...]" (UNESCO, 2020, p.5).

As recomendações da UNESCO são mais focadas nas implicações éticas da IA nas suas áreas de domínio, sendo elas: educação, ciência, cultura e comunicação e informação (UNESCO, 2020). Por esta razão, e por se situar em um contexto interdisciplinar entre a Ciência da Computação e a Educação, o presente trabalho buscará maior alinhamento com estas

recomendações. Sendo assim, a seguir apresentam-se os princípios éticos para IA recomendados pela UNESCO (2020):

- Proporcionalidade e Não Causar Danos: nenhum dos processos relacionados ao ciclo de vida dos sistemas de IA deve exceder o necessário para atingir objetivos ou metas legítimas e deve ser adequado ao contexto. A escolha de um método de IA deve ser justificada pelas seguintes maneiras: (a) O método de IA escolhido deve ser desejável e proporcional para atingir um determinado objetivo legítimo; (b) O método de IA escolhido não deve ter uma violação negativa dos valores fundamentais capturados neste documento; (c) O método de IA deve ser apropriado ao contexto e deve ser baseado em fundamentos científicos rigorosos. Na eventualidade da ocorrência de qualquer dano ao ser humano ou ao meio ambiente e aos ecossistemas, deve ser assegurada a implementação de procedimentos de avaliação de riscos e a adoção de medidas que evitem a ocorrência de tais danos. Por fim, em cenários que envolvem decisões de vida ou morte, a determinação humana final deve ser aplicada.

- Segurança e Proteção: danos indesejados (riscos de segurança) e vulnerabilidades a ataques (riscos de proteção) devem ser evitados ao longo do ciclo de vida dos sistemas de IA para garantir a segurança e proteção humana, ambiental e do ecossistema. A IA segura e protegida será possibilitada pelo desenvolvimento de estruturas de acesso a dados sustentáveis e com proteção à privacidade que promovem um melhor treinamento de modelos de IA utilizando dados de qualidade.

- Justiça e Não Discriminação: os atores de IA devem fazer todos os esforços para minimizar e evitar reforçar ou perpetuar preconceitos sociotécnicos inadequados com base no preconceito de identidade, ao longo do ciclo de vida dos sistemas de IA para garantir a justiça de tais sistemas. Deve haver a possibilidade de existir um remédio contra a determinação e discriminação algorítmica injusta. Além disso, também deverá ser promovida a equidade no que diz respeito ao acesso e participação no ciclo de vida de tecnologias de IA.

- Sustentabilidade: a avaliação contínua do impacto social, cultural, econômico e ambiental das tecnologias de IA deve ser realizada com pleno conhecimento das implicações das tecnologias de IA para a sustentabilidade.

- Privacidade: a privacidade, um direito essencial para a proteção da dignidade humana, da autonomia humana e da agência humana, deve ser respeitado, protegido e promovido ao longo do ciclo de vida dos sistemas de IA, tanto no nível pessoal quanto coletivo.

- Supervisão Humana e Determinação: Sempre deve ser possível atribuir a responsabilidade ética e legal por qualquer estágio do ciclo de vida dos sistemas de IA a pessoas naturais ou jurídicas existentes. A supervisão humana se refere, portanto, não apenas à supervisão humana individual, mas também à supervisão pública, conforme apropriado. Mesmo quando for necessário confiar em decisões de sistemas de IA por razões de eficácia, a decisão de ceder o controle em contextos limitados continua a ser dos humanos. É possível que seja necessário recorrer a sistemas de IA para a tomada de decisões e execução de ações, mas um sistema de IA nunca pode substituir a responsabilidade humana final.

- Transparência e Explicabilidade: a transparência dos sistemas de IA é frequentemente uma pré-condição crucial para garantir que os direitos humanos fundamentais e os princípios éticos sejam respeitados, protegidos e promovidos. As pessoas têm o direito de saber quando uma decisão está sendo tomada com base em algoritmos de IA e, nessas circunstâncias, exigir ou solicitar informações explicativas de empresas do setor privado ou instituições do setor público. A transparência pode contribuir para a confiança dos humanos nos sistemas de IA. Explicabilidade refere-se a tornar inteligível e fornecer informações sobre o resultado dos sistemas de IA. A explicabilidade dos sistemas de IA também se refere à compreensibilidade da entrada, saída e comportamento de cada bloco de construção algorítmico e como ele contribui para o resultado do processamento.

- Responsabilidade e Prestação de Contas: Os atores de IA devem respeitar, proteger e promover os direitos humanos e a proteção do meio ambiente e dos ecossistemas, assumindo responsabilidade ética e legal de acordo com a legislação nacional e internacional vigente, em particular a legislação internacional de direitos humanos, princípios e padrões éticos e orientações ao longo do ciclo de vida de sistemas de IA. A responsabilidade ética e a responsabilidade por decisões e ações baseadas de qualquer forma em um sistema de IA devem sempre ser atribuídas aos atores de IA. Devem ser desenvolvidos mecanismos adequados de supervisão, avaliação de impacto e devida diligência para garantir a responsabilidade pelos sistemas de IA e seu impacto ao longo de seu ciclo de vida.

- Conscientização e Alfabetização: A conscientização pública e a compreensão das tecnologias de IA e do valor dos dados devem ser promovidas por meio de educação aberta e acessível, engajamento cívico, habilidades digitais e treinamento de ética em IA, alfabetização em mídia e informação e treinamento liderado em conjunto por governos, organizações intergovernamentais, sociedade civil, academia, a mídia, líderes comunitários e setor privado. Tal iniciativa deve considerar a diversidade linguística, social e cultural existente, para garantir a participação pública efetiva para que todos os membros da sociedade possam tomar decisões informadas sobre o uso de sistemas de IA e serem protegidos de influências indevidas. Aprender sobre o impacto dos sistemas de IA deve incluir aprender sobre, por meio de e para os direitos humanos, o que significa que a abordagem e a compreensão dos sistemas de IA devem ser baseadas em seu impacto nos direitos humanos e no acesso a estes direitos.

- Governança Adaptativa e Colaboração com Múltiplas Partes Interessadas: o direito internacional e a soberania devem ser respeitados no uso de dados. Soberania de dados significa que os Estados, em conformidade com o direito internacional, regulam os dados gerados em seus territórios ou que transitam por eles, e tomam medidas para a regulação efetiva dos dados com base no respeito ao direito à privacidade e outros direitos humanos. A participação de diferentes partes interessadas ao longo do ciclo de vida do sistema de IA é necessária para a governança inclusiva da IA, compartilhamento dos benefícios da IA e avanço tecnológico justo e sua contribuição para os objetivos de desenvolvimento. A adoção de padrões abertos e interoperabilidade para facilitar a colaboração deve estar em vigor. Devem ser adotadas medidas para levar em conta as mudanças nas tecnologias, o surgimento de novos grupos de partes interessadas e para permitir uma intervenção significativa por grupos marginalizados.

Como é possível notar, há várias intersecções entre os princípios propostos por cada entidade. Além disso, é possível observar que algumas propostas tratam de pontos que outras não cobrem ou cobrem de forma indireta, contextualizando-os em questões de direitos humanos, por exemplo. Por esta razão, e com o objetivo de compreender melhor as intersecções e diferenças entre tais princípios, o quadro 1 apresenta uma relação entre eles, tendo como base aqueles propostos pela UNESCO (2020), que, conforme já mencionado, trata com mais ênfase os temas de interesse do presente trabalho.

Quadro 1 – Princípios Éticos Propostos para IA e Suas Relações.

UNESCO	PRINCÍPIOS DE ASILOMAR	IEEE	OECD
Proporcionalidade e não causar danos	1. Liberdade e privacidade 2. Cuidado de capacidade	1. Bem-estar 2. Eficácia	
Segurança e Proteção	3. Evitar corridas 4. Autoaprimoramento recursivo 5. Segurança 6. Riscos	3. Conscientização sobre o uso indevido	1. Robustez, segurança e proteção
Justiça e Não Discriminação	7. Alinhamento de valores 8. Valores humanos 9. Benefício compartilhado	4. Direitos Humanos	2. Valores centrados no ser humano e justiça
Sustentabilidade	10. Objetivo da pesquisa 11. Prosperidade compartilhada		
Privacidade	12. Privacidade pessoal		
Supervisão humana e determinação	13. Controle Humano		
Transparência e explicabilidade	14. Transparência quanto a falhas	5. Transparência 6. Responsabilidade	3. Transparência e explicabilidade
Responsabilidade e prestação de contas	15. Responsabilidade		4. Responsabilidade
Conscientização e alfabetização	16. Cultura de pesquisa	7. Agência de Dados 8. Criadores de Competência	
Governança adaptativa e colaboração com múltiplas partes interessadas	17. Vínculo Ciência-Política 18. Importância 19. Bem comum		5. Crescimento inclusivo, desenvolvimento sustentável e bem-estar

Fonte: Autoria própria.

O quadro 1 serve como uma referência para compreender as relações entre os princípios propostos por cada entidade, mas é preciso alguns cuidados ao analisá-lo. Não se pode inferir, por exemplo, que os princípios propostos pelo IEEE (2019) e pela OECD (2021) não abordam questões de privacidade ou de supervisão humana. Em ambos os casos, não há uma menção direta, mas pode-se extrair da preocupação com questões de direitos humanos, valores centrados no ser humano e segurança, tratados diretamente por estas entidades, e que trazem implícitos os devidos cuidados com a privacidade e a supervisão humana. Sendo assim, ao observar o quadro 1, deve-se ter conhecimento sobre o que trata cada princípio proposto por cada entidade, de modo a compreender estas aparentes deficiências.

Com relação às iniciativas legais especialmente empreendidas no Brasil, onde o presente trabalho é desenvolvido, podem-se citar a lei No 12.965, de 23 de abril de 2014, conhecida como Marco Civil da Internet (MCI), que “estabelece princípios, garantias, direitos e deveres para o uso da Internet no Brasil” (BRASIL, 2014) e a lei No 13.709, de 14 de agosto de 2018, chamada

de Lei Geral de Proteção de Dados (LGPD), que

> dispõe sobre o tratamento de dados pessoais, inclusive nos meios digitais, por pessoa natural ou por pessoa jurídica de direito público ou privado, com o objetivo de proteger os direitos fundamentais de liberdade e de privacidade e o livre desenvolvimento da personalidade da pessoa natural (BRASIL, 2018, Art.1º).

Tais leis não abordam diretamente questões relacionadas à inteligência artificial, mas tratam de pontos sensíveis a esta, que podem ser identificados em alguns dos princípios supracitados. Esta relação pode ajudar a estabelecer limites quanto à aplicação da IA em todo o seu ciclo de vida.

Nesse sentido, os pontos de maior intersecção entre as leis brasileiras e os princípios internacionais diz respeito à privacidade e aos direitos humanos. O MCI, por exemplo, traz entre seus princípios, no Art. 4º (BRASIL, 2014): a garantia da liberdade de expressão, comunicação e manifestação de pensamento; a proteção da privacidade; e a proteção dos dados, cujos mecanismos são mais bem especificados na LGPD. Tais princípios, bem como os mecanismos estabelecidos no MCI para a garantia do seu atendimento, estão em acordo com os princípios de segurança e proteção, e de justiça e não discriminação, constantes na proposta da UNESCO (2020) para IA.

A LGPD, por sua vez, regulamenta o tratamento de dados de pessoas naturais ou jurídicas, sendo o tratamento compreendido como:

> toda operação realizada com dados pessoais, como as que se referem a coleta, produção, recepção, classificação, utilização, acesso, reprodução, transmissão, distribuição, processamento, arquivamento, armazenamento, eliminação, avaliação ou controle da informação, modificação, comunicação, transferência, difusão ou extração (BRASIL, 2018, Art.5º).

Sendo assim, a LGPD estabelece entre seus princípios:

- O princípio da finalidade: que afirma que todo tratamento de dados deve atender a propósitos legítimos, específicos, explícitos e informados ao titular dos dados, não podendo ser usada para outra finalidade;
- O princípio da adequação: que diz que o tratamento dos dados deve ser compatível com as finalidades informadas ao titular dos dados;
- O princípio da necessidade: que estabelece que o tratamento dos dados de um usuário deve ser limitado ao mínimo necessário para a realização das suas finalidades;

- Princípio do livre acesso: que expressa que deve ser garantido, aos titulares dos dados, a consulta facilitada e gratuita sobre a forma e duração do tratamento e sobre a integralidade de seus dados;

- Princípio da qualidade dos dados: que diz que deve ser garantido, aos titulares, a exatidão, clareza, relevância e atualização dos seus dados;

- Princípio da transparência: que declara que, aos titulares, devem ser garantidas informações claras, precisas e facilmente acessíveis sobre os processos de tratamento, bem como sobre os agentes de tratamento de seus dados, observadas as questões de segredo comercial e industrial;

- Princípio da segurança: que prega a utilização de medidas técnicas e administrativas adequadas para a proteção dos dados pessoais contra acessos não autorizados e situações acidentais ou ilícitas de destruição, perda, alteração, comunicação ou difusão;

- Princípio da prevenção: que diz que devem ser adotadas medidas para prevenir a ocorrência de danos aos titulares dos dados, em virtude de seu tratamento;

- Princípio da não discriminação: que defende a impossibilidade de realização de tratamento de dados para fins discriminatórios, ilícitos ou abusivos;

- Princípio da responsabilização e prestação de contas: que expressa que o agente de tratamento deve demonstrar que adotou medidas eficazes e capazes de comprovar a observância e o cumprimento das normas de proteção de dados, bem como da eficácia de tais medidas.

A exemplo do que ocorre entre os princípios mostrados no quadro 1, é possível identificar muitas intersecções entre aqueles citados na LGPD e a proposta da UNESCO (2020). O princípio da finalidade, constante na LGPD, por exemplo, pode ser relacionado aos princípios de proporcionalidade e não causar danos e da privacidade, constante na proposta da UNESCO (2020). Da mesma forma, o princípio do livre acesso (BRASIL, 2018) pode ser relacionado ao princípio da transparência e explicabilidade (UNESCO, 2020).

Estas relações são importantes, pois, conforme já mencionado, ainda que as leis brasileiras não tenham sido criadas especificamente para regular o desenvolvimento da IA, elas permitem proteger os dados dos usuários de possíveis consequências negativas do uso destas

tecnologias. Desse modo, ao menos neste contexto, pode-se afirmar que há um grau de segurança e adequação legal aos princípios internacionais para o desenvolvimento de tecnologias de IA. O próximo capítulo abordará, com mais detalhes, as principais abordagens e tecnologias para o desenvolvimento de agentes artificiais capazes de considerar critérios éticos em seus mecanismos de raciocínio e decisão.

2.2.3 Agentes Morais Artificiais (AMA)

Visando delegar parte do seu poder de tomada de decisão para agentes artificiais (AA), a humanidade vem aumentando o escopo de atividades deste tipo de agente Cervantes *et al.* (2020). Por esta razão, VA tem sido um uma questão de importância crescente em diferentes áreas, principalmente devido à falta de consenso quanto à abordagem ética a ser seguida e às tecnologias para implementá-las (CORDOVA; VICARI, 2021). Nesse sentido, várias pesquisas têm sido conduzidas no intuito de propor soluções para o desenvolvimento de Agentes Morais Artificiais (CERVANTES *et al.*, 2020).

Com isso, no presente trabalho assumir-se-á a definição de Cervantes *et al.* (2020), que descreve AMAs da seguinte forma:

> Um AMA é um agente virtual (software) ou agente físico (robô) capaz de se envolver em um comportamento moral ou, pelo menos, de evitar um comportamento imoral. Esse comportamento moral pode ser baseado em teorias éticas, como ética teleológica, deontologia e ética da virtude, mas não necessariamente. Os AMAs podem ser classificados de acordo com sua abordagem de design (CERVANTES *et al.*, 2020, p.5, tradução nossa).

Esta definição implica uma ampla gama de tipos de agentes capazes de comportar-se de forma considerada ética. Além disso, deixa claro que estes AA podem ou não ser orientados por uma teoria da moral, o que permite incluir diferentes técnicas como as que serão apresentadas mais adiante.

Entretanto, é necessário, primeiramente, deixar claro o que pode ser considerado como um comportamento ético por parte dos AMAs. A este respeito, embora não haja consenso, Dignum *et al.* (2018) sugerem que decisões éticas tomadas por sistemas de IA são aquelas que estão relacionadas ou tem impacto direto sobre a dignidade e o bem-estar de seres humanos. Este tipo de decisão não é trivial e, segundo Dennis *et al.* (2016) pode-se dizer que, em geral, o raciocínio explicitamente baseado na ética só será necessário em circunstâncias especializadas que ocorrem quando o agente precisa raciocinar dentro de alguns limites pré-determinados como

éticos ou para resolver algum conflito entre princípios morais. O segundo caso é conhecido como dilema ético.

A resolução de dilemas éticos é um dos pontos mais importantes em pesquisas sobre AMAs (CERVANTES *et al.*, 2020). Pode-se definir um dilema ético como uma situação para a qual não há uma decisão satisfatória e, portanto, qualquer decisão irá infringir ou sobrepor-se à uma regra ou princípio (AROSKAR, 1980). Este tipo de situação, segundo Cervantes *et al.* (2020), pode acontecer em duas condições não exclusivas: a primeira é quando ocorre um conflito entre regras internas de um mesmo agente; a segunda pode ocorrer entre dois agentes quando divergem sobre qual a decisão ética cabível. Esta última pode envolver dois AA ou a interação entre um AA e um humano.

Situações adicionais podem surgir a partir das condições supracitadas, a depender da quantidade e do tipo de agentes envolvidos e impactados pela decisão, bem como da relação entre eles e do tipo de dilema (CERVANTES *et al.*, 2020). Com relação ao tipo de dilema, Cristani e Burato (2009) os classificam em dilemas de obrigação e dilemas de proibição. No primeiro caso, todas as opções de ação do agente são viáveis e, com base em suas regras internas, são também obrigatórias. Entretanto, o agente não pode executar mais do que uma e precisa decidir sobre qual deve ser executada. No segundo caso, todas as opções do agente são proibidas de acordo com suas regras internas. Entretanto, ele deve executar uma delas e precisa decidir sobre qual.

Quanto à classificação dos AMAs, atualmente existem duas propostas bem aceitas pela comunidade de pesquisa em VA. Uma delas foi proposta por Allen *et al.* (2005) e a outra por Moor (2006). Apesar de a proposta de Moor (2006) ser mais genérica, ambas devem ser vistas como complementares para que se possa compreender com mais detalhes estes agentes capazes de demonstrar comportamento ético.

Nesse sentido, em sua proposta Moor (2006) classifica os AMAs em agentes éticos implícitos, agentes éticos explícitos e agentes éticos plenos. Uma descrição mais detalhada destas classes pode ser dada conforme a seguir:

- Agentes éticos implícitos: segundo Moor (2006), uma forma de desenvolver agentes éticos implícitos é restringindo suas possibilidades de ação para evitar comportamentos ou resultados não éticos. Para este autor, computadores implicitamente éticos atendem a requisitos de segurança e confiabilidade sem implementar códigos éticos explicitamente. Segundo Cervantes *et al.* (2020), agentes desta categoria apresentam três características bem claras: não dispõem de mecanismos para diferenciar uma ação ética de uma ação não

ética; suas características qualitativas de adequação funcional e segurança são conhecidas e satisfatoriamente testadas; e eles não contêm código malicioso. Qualquer computador com estas características pode ser considerado um agente ético implícito.

- Agentes éticos explícitos: esta categoria, como o nome sugere, é formada por AMAs que implementam a ética de forma explícita. Estes agentes são classificados de acordo com a estratégia utilizada em seu processo de tomada de decisão ética (MOOR, 2006). Estas estratégias são aquelas propostas por Allen *et al.* (2005), a saber: a abordagem *top-down*, que faz uso de teorias éticas, como deontologia, utilitarismo e ética da virtude; a abordagem *bottom-up*, que faz uso de mecanismos de aprendizagem de máquina para desenvolver habilidades para tomada de decisão ética, mas não implementa nenhuma estratégia baseada em teorias éticas; e a abordagem híbrida, que faz uso de recursos e critérios éticos das abordagens *top-down* e *bottom-up*. A maior parte dos esforços para o desenvolvimento de AMAs se concentra nesta categoria de agentes.

- Agentes éticos plenos: segundo Moor (2006), as características desta categoria de agentes são semelhantes às dos agentes éticos explícitos. Contudo, agentes éticos plenos têm atributos que são geralmente atribuídos à humanos, como consciência, livre arbítrio e intencionalidade. Há um sério debate sobre se, em um futuro próximo, será possível desenvolver, de fato, um agente desta categoria Han e Pereira (2019), Wynsberghe e Robbins (2019), Malle (2016).

Conforme já mencionado, a maior parte das pesquisas em ética de máquina se concentra no desenvolvimento de agentes éticos explícitos. Esta categoria é o foco deste trabalho. Além disso, a classificação proposta por Allen *et al.* (2005) também é direcionada para organizar AMAs nesta categoria. Nesse sentido, Cervantes *et al.* (2020) propuseram uma taxonomia baseada nas propostas de Moor (2006) e Allen *et al.* (2005) para organizar e classificar os AMAs identificados em seu estudo sobre o estado da arte dos AMAs, publicado em 2019 e disponível em (CERVANTES *et al.*, 2020).

Com isso, a seguir será apresentada esta taxonomia e seus respectivos modelos de AMA, bem como outros modelos não previstos nela, mas considerados relevantes para esta revisão da literatura. Finalmente, para enriquecer mais a discussão sobre o tema, serão também apresentados os principais desafios enfrentados pelas abordagens *top-down*, *bottom-up* e híbrida.

As abordagens *top-down*, conforme já mencionado, trazem a ideia de que teorias éticas ou princípios morais podem ser usados como regras para a seleção de ações eticamente apropriadas (ALLEN *et al.*, 2005). Esta abordagem faz uso, fundamentalmente, de um sistema baseado em regras, o que já se traduz, por si só, em limitações relacionadas a restrições por contexto e grande probabilidade de ocorrência de conflitos entre diferentes regras. Gips (1995), entretanto, argumenta que há tentativas, por parte de alguns filósofos, de criar princípios gerais mais abstratos dos quais regras mais específicas podem ser derivadas, e que tanto a abordagem utilitarista, quanto a deontológica contam com princípios gerais que poderiam ser relevantes para o projeto de robôs éticos.

Dessa forma, que concerne às soluções propostas dentro desta abordagem, destacam-se os seguintes modelos:

- MoralDM: um modelo computacional que integra várias técnicas de IA, como PLN para produzir representações formais de problemas a serem resolvidos, algoritmos para calcular a utilidade e a consequência de ações, módulos contendo princípios morais, entre outros. Atualmente, este modelo é capaz de exibir comportamento utilitarista e deontológico dependendo do problema enfrentado (DEHGHANI *et al.*, 2008; GUERINI *et al.*, 2015; BLASS, 2016).

- Jeremy: É uma solução baseada em ética utilitarista, que implementa o algoritmo Hedonistic Act Utilitarianism (HAU), fundamentado na teoria de Jeremy Bentham (ANDERSON; ANDERSON, 2008).

- Um mecanismo para rejeitar, apropriadamente, diretivas nas interações entre humanos e robôs: Proposta para dotar AA com um mecanismo capaz de rejeitar diretivas e prover uma explicação associada. Este mecanismo foi implementado na arquitetura robótica cognitiva DIARC/ADE. Tal robô foi testado em um cenário simples de interação humano-robô. O seu processo de raciocínio de aceitação ou rejeição de diretivas – em forma de comandos dados por humanos – abrange cinco categorias de condições de felicidade que devem ser mantidas para aceitar explicitamente uma proposta: conhecimento, capacidade, prioridade e tempo do objetivo, papel social e obrigação e permissibilidade normativa. A permissibilidade normativa, neste caso, considera um conjunto de regras que indicam quais ações estão incorretas e, portanto, devem ser rejeitadas (BRIGGS; SCHEUTZ, 2015).

Existem ainda outros estudos mais recentes que apresentam propostas de soluções para alinhamento de valores em IA usando a abordagem *top-down*. Estas soluções abrangem a implementação de funções de utilidade puras ou estruturadas para lidar com problemas multiobjectivo (VAMPLEW *et al.*, 2018) e a combinação de normas morais, regras de julgamento moral, planos de Crença-Desejo-Intenção (BDI), emoções e outras informações sobre contexto do agente para lidar com questões relacionadas à reputação social (COSTA; COELHO, 2019).

Com relação aos desafios de implementação dos modelos top-down, destacam-se a dificuldade em limitar a quantidade de variáveis a serem analisadas ou a quantidade de possíveis consequências para uma dada ação antes de decidir executá-la (ALLEN *et al.*, 2005), a complexidade para transformar princípios morais e normas éticas em modelos computacionais (BONNEMAINS *et al.*, 2018) e, por fim, esforços para evitar cenários de instanciação perversa e problemas de mudança de contexto (ALIMAN; KESTER, 2019). Além disso, criar mecanismos para evitar que o sistema sofra tentativas de corrupção das suas regras é fundamental para garantir a sua confiabilidade.

A abordagem *bottom-up*, por sua vez, abrange estratégias que não impõe a aplicação de teorias morais específicas, mas que possibilitam que um comportamento apropriado por parte do agente seja selecionado ou recompensado (ALLEN *et al.*, 2005). Essas estratégias para o desenvolvimento de comportamento moral envolvem a aprendizagem gradativa por meio da experiência, seja recorrendo a ajustes realizados por programadores ou engenheiros quando encontram novos desafios, ou ao desenvolvimento educacional de uma máquina de aprendizagem.

Quanto às tentativas de implementação de AMAs bottom-up, podem ser destacados os seguintes modelos:

- *Causuist BDI-agent*: uma arquitetura estendida do modelo de estados mentais conhecido como BDI, incluindo na arquitetura o método de raciocínio baseado em casos (HONARVAR; GHASEM-AGHAEE, 2009). Este AMA é baseado em experiências prévias e não usa códigos de ética. O *Causuist BDI-agent* possui um módulo para recuperação de casos passados (*Case Retriever*), um módulo de avaliação de casos (*Case Evaluator*), um atualizador de casos (*Case Updater*) e uma memória de casos (*Case Memory*), além é claro, do agente BDI com toda sua arquitetura. Assim, ao se deparar com uma situação qualquer, o agente irá submeter a situação para o *Case Evaluator*, que contém uma função utilidade capaz de calcular o impacto de suas ações. Se o agente se deparar com uma situação para a qual haja algum caso similar em seu *Case Memory*, o *Case Retriever* irá recuperar a

solução e o agente irá aceitá-la, adaptá-la e aplicá-la. De outro modo, caso o agente se depare com uma situação nova, irá se comportar como um BDI clássico, processando a situação e decidindo por meio do seu *Case Evaluator*, que deverá calcular o melhor cenário e aplicar a ação. No final o *Case Updater* irá criar um novo caso em sua *Case Memory* (HONARVAR; GHASEM-AGHAEE, 2009).

- *GenEth*: este é um analisador geral de dilemas éticos. Usa programação lógica indutiva para inferir princípios a partir de preferências de ação ética. Desse modo, aprende com exemplos de comportamento humano. Este analisador foi usado para codificar princípios em diferentes domínios relacionados ao comportamento de sistemas autônomos (ANDERSON; ANDERSON, 2018). Um exemplo prático de sua aplicação foi em um protótipo do robô humanoide NAO, desenvolvido pela *SoftBank Robotics* . A tarefa deste protótipo era lembrar pacientes sobre horários de tomar medicamentos. Precisava, portanto, lidar com situações nas quais o paciente deixava de tomá-los. Diante disso, o NAO robô deveria lidar com três deveres: garantir que o paciente receba um possível benefício ao tomar a medicação; prevenir os malefícios que podem advir de não tomar o medicamento; e respeitar a autonomia do paciente. O NAO, então, recebeu informações iniciais que incluíam a hora de tomar um medicamento, a quantidade máxima de dano caso não fosse tomado, o tempo que levaria para que esse dano ocorresse, a quantidade máxima de bem esperado a partir do uso do medicamento e quanto tempo levaria para que esse benefício fosse perdido. O robô, então, calcula, a partir dessas entradas, seus níveis de satisfação ou violação do dever para cada um dos três deveres e executava ações diferentes dependendo de como esses níveis mudavam ao longo do tempo. Os testes foram considerados satisfatórios pelos pesquisadores (ANDERSON; ANDERSON, 2010; CERVANTES *et al.*, 2020).

Além destes, é importante citar também o trabalho de Haas (2020), que sugere o uso de uma abordagem de aprendizagem de máquina conhecida como aprendizagem por reforço – do inglês, *Reinforcement Learning* (RL) –, para possibilitar que agentes aprendam a valorizar e responder a contextos e conteúdos morais. O uso de ML para resolver questões de alinhamento de valores tem sido bastante especulado nos últimos anos. Nesse contexto, Russell *et al.* (2015), sugerem que o alinhamento de valores em IA é uma excelente oportunidade para o desenvolvimento de soluções envolvendo a aprendizagem por reforço inverso – do inglês, *Inverse Reinforcement Learning* (IRL).

Por outro lado, técnicas de ML têm sido também bastante questionadas por trazerem consigo alguns de seus problemas já conhecidos. A este respeito, Arnold *et al.* (2017) argumentam que IRL não é adequada para resolver problemas de VA e apontam alguns importantes desafios a serem superados por estas técnicas, entre os quais, destacam a necessidade de superar o problema dos dados ruins. Sobre este ponto, afirmam que um sistema " herda, às vezes para pior, os preconceitos e as características dos dados sobre os quais é treinado. Se um agente IRL aprender com um comportamento antiético, ele aprenderá a se comportar de forma antiética" (ARNOLD *et al.*, 2017, p.4, tradução nossa) Segundo os mesmos autores (2017), outro problema estaria relacionado à capacidade de generalização, própria da IRL. Esta capacidade poderia fazer com que o agente generalizasse certos comportamentos para estados nos quais não seriam apropriados.

Por fim, outra questão que precisa ser superada pelas técnicas de ML no contexto da aprendizagem de valores éticos é o risco de incorrer na falácia naturalista. Esta, por sua vez, é a falácia que "incorre em definir o certo a partir apenas de um estado de coisas, ou em suma, inferir um deve de um é" (KIM *et al.*, 2019, p.3, tradução nossa). Por exemplo, o fato de que um grupo de pessoas esteja envolvida em uma determinada ação ou acreditem que esta ação é ética, não significa que esta ação de fato o seja. Confiar, portanto, o aprendizado sobre o alinhamento de valores à técnicas de IRL ou outro método empírico, pode não trazer aprendizagens genuínas sobre princípios éticos (KIM *et al.*, 2019).

A abordagem híbrida para AMAs, como o nome sugere, faz uso de estratégias *top-down* e *bottom-up* simultaneamente. Nesse sentido, enquanto a primeira enfatiza a importância de regras que vêm de fora da entidade, restringindo seu comportamento, a segunda tem foco em regras implícitas que são construídas dentro dela, e que, de certa forma, flexibilizam e aumentam suas opções de ação (ALLEN *et al.*, 2005).

Algumas das implementações dentro desta categoria de AMA, são:

- LIDA: trata-se de uma arquitetura cognitiva geral que considera o aspecto moral como uma questão relevante (WALLACH *et al.*, 2010). Para este AMA, decisões de ordem moral podem ser tomadas em vários domínios usando o mesmo mecanismo que permite a tomada de decisão em situações gerais. LIDA foi parcialmente implementado em um *CareBot*, consistindo em um assistente robô móvel, operando em um ambiente 2D simulado (MADL; FRANKLIN, 2015). Sua arquitetura híbrida consiste em agrupar suas funções cognitivas em uma abordagem *top-down*, que implica na implementação de teorias éticas através de regras, enquanto seu aspecto *bottom-up* inclui mecanismos de aprendizagem envolvendo

preferências em forma de sentimentos e valores que influenciam a moralidade (WALLACH *et al.*, 2010).

- Modelo de Tomada de Decisão Ética: um modelo computacional baseado em neurociência, projetado para prover agentes autônomos com capacidades para tomar decisão ética (CERVANTES *et al.*, 2020). Este modelo considera quatro níveis de avaliação, sendo: avaliação primária, que se concentra em itens – pessoas ou coisas – que podem ser afetadas por cada ação provável. O resultado dessa avaliação inicial é definido como o nível de prazer ou desprazer relacionado a cada item no ambiente. Em seguida, o agente computa tanto a recompensa esperada quanto a provável punição relacionada às ações. O agente usa experiências anteriores relacionadas à situação atual para calcular a recompensa esperada. Essas experiências são classificadas como boas ou más para determinar se a recompensa esperada será favorável ou desfavorável. Por fim, o agente realiza uma avaliação com base em normas éticas, que são expressas como regras. Tais regras são estruturadas de forma a expressar o seu nível de concordância, o seu significado e informações emocionais relacionadas ao respeito ou violação de tais regras. Estas informações são usadas para decidir quais regras podem ser violadas ao se deparar com dilemas éticos (CERVANTES *et al.*, 2020). Este modelo foi implementado em um agente virtual e testado usando alguns cenários hipotéticos, mostrando como informações emocionais podem influenciar as ações do agente.

- MedEthEx: este é um agente ético de saúde, cuja arquitetura combina uma abordagem casuística bottom-up com uma arquitetura *top-down* de uma teoria ética que implementa os princípios de Ética Biomédica de Beauchamp e Childress (2001), usando aprendizagem de máquina e deveres *prima facie* para resolver dilemas éticos biomédicos (ANDERSON *et al.*, 2005). Tal arquitetura é dividida em três componentes: uma interface baseada em conhecimento que provê orientação na seleção de intensidades de dever para um caso particular, um módulo conselheiro, que determina a ação correta para um caso particular consultando o conhecimento aprendido, e um módulo de aprendizagem que abstrai os princípios gerais e orientadores de casos particulares fornecidos por um especialista (ANDERSON; ANDERSON, 2008; ANDERSON *et al.*, 2005) (ANDERSON; ANDERSON, 2008), (ANDERSON et al., 2005). O MedEthEx foi testado usando simulações de caso e foi considerado um programa de aprendizagem baseado em computador, compondo um curso obrigatório de Bioética (CERVANTES *et al.*, 2020).

- Sistema de múltiplos agentes éticos: este modelo lida com questões relacionadas ao raciocínio ético-preferencial (CRISTANI; BURATO, 2009). Os algoritmos propostos neste modelo dotam sistemas multiagentes com capacidade de tomar decisões sobre dilemas éticos. Para isto, o modelo considera que os agentes coexistem em um ambiente não cooperativo, onde ou competem para atingir objetivos individuais, ou colaboram para atingir objetivos coletivos, jamais cooperam[12]. Além disso, o sistema é especificado em termos de agentes com compromissos éticos, tanto como obrigações quanto como proibições. O algoritmo implementa a noção de coerência para compromissos e especifica a noção de grau de incoerência, que possibilita aos agentes lidar com dilemas morais. Este modelo computacional foi testado usando demonstrações formais e casos teóricos (CRISTANI; BURATO, 2009).

Além dos modelos apresentados, não se pode deixar de citar a proposta de Arnold *et al.* (2017), que combina os pontos fortes de reinforcemente learning (RL) e representações lógicas para restringir ou regular o comportamento do agente. Neste caso, as normas serviriam como heurísticas prontas, capazes de evitar comportamentos antiéticos, ainda que o agente os aprendesse por meio da RL (ARNOLD *et al.*, 2017). Por outro lado, em uma abordagem oposta, Kim *et al.* (2019) propõe o uso da observação empírica para validar princípios éticos previamente concebidos no agente para evitar a falácia naturalista, ao mesmo tempo que tenta resolver questões relacionadas a mudanças de contexto ao tratar de princípios e normas éticas.

Os maiores desafios ao implementar a abordagem híbrida são aqueles trazidos pelas abordagens *top-down* e *bottom-up*, que compõe estas arquiteturas. Outro importante desafio, está relacionado à seleção e combinação de abordagens e técnicas adequadas.

Dado o exposto, apesar de todas as iniciativas em busca de soluções para o alinhamento de valores em IA demonstradas, um estudo publicado por Zoshak e Dew (2021) mostra que a maior parte dos esforços neste campo são direcionados para a implementação de teorias éticas, com maior foco em abordagens deontológicas e utilitaristas. Segundo os autores, isto deve ocorrer devido à adequação destas teorias à cultura ocidental. Além disso, a maior parte dos estudos em VA está focado no desenvolvimento da própria IA. Poucos estudos apresentam aplicação prática em contextos do mundo real. Entretanto, é possível encontrar trabalhos de aplicação de AMAs na área de veículos autônomos (THORNTON *et al.*, 2017), cuidados de saúde (MISSELHORN, 2020; ANDERSON *et al.*, 2005), jogos digitais (CASAS-ROMA; ARNEDO-MORENO, 2019)

[12] No capítulo 2.2 são apresentadas definições que diferenciam colaboração de cooperação.

e educação (MABASO, 2020). Por fim, como foi mostrado ao longo deste capítulo, a falta de consenso evidencia a complexidade do desenvolvimento de AMAs.

2.3 APRENDIZAGEM COLABORATIVA

Desde a década de 1960 tem se observado um interesse crescente em compreender e desenvolver as diferentes abordagens e técnicas para aprendizagem em grupo (DAVIDSON; MAJOR, 2014). Entre as principais propostas encontradas na literatura para este fim, encontram-se trabalhos relacionados a aprendizagem colaborativa, a aprendizagem cooperativa, a aprendizagem baseada em problemas, a instrução por pares, a tutoria por pares, entre outras (DAVIDSON; MAJOR, 2014). Esta variedade de abordagens evidencia a complexidade da aprendizagem em grupo. Sendo assim, no intuito de delimitar o escopo de pesquisa e dedicar maior atenção à abordagem mais relevante para os fins do presente trabalho, especial destaque será dado à aprendizagem colaborativa.

Esta abordagem, entretanto, é frequentemente confundida, comparada e, algumas vezes, até vista como sinônimo de aprendizagem cooperativa (JACOBS, 2015). Nesse sentido, visando estabelecer conceitos, características e diretivas acerca das duas abordagens, apresentar-se-á, a seguir, uma visão geral sobre ambas.

Com isso, é importante compreender, primeiramente, que tanto a aprendizagem colaborativa quanto a aprendizagem cooperativa têm o propósito de promover o trabalho em conjunto entre os estudantes (BRUFFEE, 1995). Além disso, ambas as abordagens visam oferecer aos alunos mais controle sobre sua própria aprendizagem e promovem a ideia de que aprender por meio da interação e da discussão com outras pessoas pode levar a melhores resultados (JACOBS, 2015).

De acordo com o Dicionário Online de Português, o vocábulo colaborar tem sua origem no latim, *collaborare* (trabalhar com), significando, portanto, "trabalhar em comum com outrem", sendo considerado sinônimo de cooperar (COLABORAR, 2020). A palavra cooperar, por sua vez, tem também sua origem no latim, *cooperari* (operar com), trazendo o significado de "operar simultaneamente ou coletivamente" e sendo reciprocamente considerada sinônimo de colaborar (COOPERAR, 2020).

A partir destas definições não é possível determinar uma diferença significativa entre as duas abordagens. Contudo, conforme afirmam Davidson e Major (2014), tais definições não se limitam a questões semânticas, em vez disso, estão mais relacionadas às filosofias, objetivos

e métodos subjacentes das diferentes abordagens, além da própria história e contexto onde foram desenvolvidas. Sendo assim, as principais distinções vêm de suas diferentes origens e originadores, tendo sido tais abordagens desenvolvidas separadamente, em seus próprios campos de investigação, com diferentes publicações e diferentes conferências (DAVIDSON; MAJOR, 2014).

Desse modo, para Bruffee (1995) a aprendizagem colaborativa se desenvolveu, principalmente, em universidades e faculdades, dando maior ênfase à autonomia do aluno em relação ao professor, enquanto a aprendizagem cooperativa foi mais utilizada em escolas primárias, exigindo maior interação, regulação e intervenção do professor. Com relação às áreas de aplicação, Davidson e Major (2014) afirmam que:

> A aprendizagem colaborativa tem sido usada principalmente nas ciências humanas, algumas nas ciências sociais, mas raramente em outras ciências ou programas profissionais. A aprendizagem cooperativa tem sido usada principalmente nas ciências, matemática e engenharia, ciências sociais e programas profissionais (DAVIDSON; MAJOR, 2014, p.3, tradução nossa).

As razões para a decisão sobre uma ou outra abordagem podem estar profundamente enraizadas nas filosofias e nos propósitos dos métodos (DAVIDSON; MAJOR, 2014).

A este respeito, pode-se dizer que na aprendizagem cooperativa os procedimentos são projetados cuidadosamente para envolver ativamente todos os integrantes do grupo em um empreendimento cooperativamente compartilhado. As funções dentro dos grupos devem estar devidamente definidas, comunicadas e sendo exercidas. O professor, por sua vez, tem um papel ativo, circulando entre os grupos e fornecendo assistência, encorajamento e fazendo perguntas estimulantes conforme necessário (DAVIDSON; WORSHAM, 1992). Assim, "o foco desta abordagem é garantir que os alunos trabalhem juntos, não apenas no mesmo projeto" (DAVIDSON; MAJOR, 2014, p.14, tradução nossa).

A aprendizagem colaborativa, por outro lado, assume que a aprendizagem ocorre quando alunos e professores trabalham juntos para criar conhecimento (MATTHEWS, 1996). Para Dillenbourg (1999), trata-se de uma situação na qual duas ou mais pessoas tentam aprender alguma coisa em conjunto. Assim, ao contrário da aprendizagem cooperativa, onde o foco está em trabalhar juntos, de forma interdependente, na aprendizagem colaborativa busca-se promover o trabalho em grupo, mas não necessariamente de forma interdependente, ou seja, visa-se trabalhar uns com os outros no mesmo projeto, buscando atingir um mesmo fim, mas não necessariamente de forma cooperativa na mesma tarefa (DAVIDSON; MAJOR, 2014).

Com isso, é possível perceber que a aprendizagem colaborativa é mais flexível, menos estruturada e emprega menos atividades de facilitação e intervenção do professor (MATTHEWS *et al.*, 1995). Bruffee (1995) reforça este entendimento ao afirmar que na abordagem cooperativa, ao tentar evitar tanto a competição entre os indivíduos, quanto a dependência crônica de alguns em relação a outros, os professores acabam estruturando suas aulas sob um conjunto mais rígido de regras sociais e costumam realizar mais intervenções durante a condução da aprendizagem. Outra característica marcante desta abordagem é a maior exigência sobre os alunos para que prestem contas de suas atividades por meio de diferentes formas de avaliação. Por outro lado, adeptos da abordagem colaborativa tendem a valorizar mais a autonomia dos estudantes, buscando torná-los mais responsáveis por seu próprio aprendizado. Nesse contexto, uma das principais propostas desta abordagem é a substituição da estrutura social tradicional da sala de aula com a autoridade focada no professor, por uma outra estrutura de relações negociadas entre discentes e docente (BRUFFEE, 1995).

Dando um pouco mais de ênfase às diferenças entre estas abordagens, Davidson e Major (2014) afirmam, entre outras coisas, que grupos de aprendizagem cooperativa demandam com mais frequência a ajuda do instrutor, enquanto grupos colaborativos são, em sua maioria, mais auto gerenciados. Outro aspecto importante é que "a aprendizagem colaborativa nunca usa funções de grupo atribuídas, mas algumas abordagens de aprendizagem cooperativa fazem isso" (DAVIDSON; MAJOR, 2014, p.34, tradução nossa). Além disso, ainda segundo estes autores (2014), a maioria dos modelos de aprendizagem cooperativa admite intervenção do instrutor até para a formação dos grupos, que também pode sofrer atribuição aleatória. Na abordagem colaborativa, entretanto, é mais frequentemente estimulado que os alunos formem seus próprios grupos.

Estas definições reforçam a ideia de que a aprendizagem colaborativa é mais flexível e requer menos intervenção do professor do que a cooperativa. Embora ambas visem "tornar os alunos mais responsáveis por sua aprendizagem, levando-os a assimilar conceitos e a construir conhecimentos de uma maneira mais autônoma" (TORRES; Irala, 2014, p.61), na abordagem colaborativa os alunos, em geral, tem maior liberdade de auto-organização. Na sequência serão apresentados os fundamentos teóricos desta abordagem, bem como seus objetivos mais específicos, no intuito de compreender melhor o que justifica esta proposta que dá mais foco na autonomia dos estudantes.

2.3.1 Fundamentos Teóricos da Aprendizagem Colaborativa

Conforme abordado no capítulo 2.1, uma importante definição apresentada por Dillenbourg (1999) afirma que a aprendizagem colaborativa ocorre em situações nas quais duas ou mais pessoas aprendem ou tentam aprender algo juntas. Esta definição, segundo o próprio autor, traz em si alguns elementos que precisam ser considerados com mais cuidado. Duas ou mais pessoas, por exemplo, pode ser interpretado como um par, um pequeno grupo, uma sala de aula, uma comunidade ou sociedade inteira com milhões de pessoas. Aprender algo, por sua vez, pode ser interpretado como acompanhar um curso, estudar determinado material, realizar atividades de aula ou aprender com a prática profissional ao longo da vida. Por fim, juntos, pode significar diferentes formas de interação, como: mediada por computador ou face a face, síncrona ou assíncrona, com esforço conjunto em cada atividade ou com divisão sistemática do trabalho, entre tantas outras possíveis interpretações (DILLENBOURG, 1999).

Com isso, "a prática de aprendizagem colaborativa pode assumir múltiplas caracterizações, podendo haver dinâmicas e resultados de aprendizagem diferentes para cada contexto específico" (TORRES; Irala, 2014, p.61). Ao tratar deste tema, portanto, é preciso considerar, principalmente no que concerne às bases teóricas que o sustentam, diferentes contextos e dimensões da vivência social. Nesse contexto, Torres e Irala (2014) afirmam que a aprendizagem colaborativa é fundamentada por um conjunto de tendências pedagógicas e bases teóricas bastante difundidas no decorrer da história da educação, dentre as quais: o movimento Escola Nova, a epistemologia genética de Piaget, a teoria socio cultural de Vygostky e a pedagogia progressista.

Como exemplos de educadores do movimento Escola Nova, podem ser destacados John Dewey, Maria Montessori, Freinet, Cousinet e Edouard Claparède. Segundo Nunes (1998), o movimento Escola Nova superou a concepção de educação escolar, ao expandir o tradicional domínio oral e escrito e propor um novo sistema de produção de significados com base na interação comunicativa. A escola nova, com isso, pretendia transformar o aluno em um agente participativo da sua própria educação. Nesse contexto, Bruffee (1995) chama a atenção para o conceito de vida associada, usado por John Dewey para se referir às atividades nas quais as relações humanas eram a chave para alcançar o bem-estar e a maestria, e pondera que esta é uma boa base para sustentar todas as ideias e procedimentos educacionais nos quais as pessoas dependem umas das outras para aprender. Além disso, segundo Torres e Irala (2014):

> foi também implementada por Dewey a metodologia de trabalho em grupos. Tendo como base os desenvolvimentos teóricos da psicologia e sociologia de

> sua época e com ênfase na educação democrática, suas filosofias exerceram grande influência para importantes mudanças na sociedade, tais como: a relação de dependência entre a aprendizagem e as atividades sociais, a influência do ambiente físico no desenvolvimento da cultura e a necessidade de promoção das diferenças individuais a fim de se produzirem mudanças na sociedade. Duas importantes filosofias, implementadas por Dewey, contribuíram para o desenvolvimento da aprendizagem colaborativa: a democracia na educação e a aprendizagem socialmente interativa (TORRES; Irala, 2014, p.71).

A abordagem colaborativa valoriza a ação e o estabelecimento de um ambiente democrático de ensino e aprendizagem. Esta democratização da sala de aula desconstrói a hierarquia clássica das interações entre alunos e professores e valoriza o papel do discente como protagonista deste processo. Estes são alguns dos elementos que caracterizam os pressupostos da Escola Nova e das teorias de Dewey na aprendizagem colaborativa. (TORRES; Irala, 2014).

A Epistemologia Genética de Jean Piaget, por sua vez, tem o objetivo de compreender, por meio de uma perspectiva da biologia, como o sujeito passa de um estado de menor conhecimento para um estado de conhecimento maior. Segundo esta teoria, o conhecimento não pode ser transferido ou imposto do meio para o sujeito, nem se encontra a priori no próprio sujeito aguardando a maturação, mas é o resultado de uma construção a partir das interações do sujeito com o seu meio (CAETANO, 2010). Nesse processo, segundo Caetano (2010), as estruturas de inteligência vão se construindo na medida em que o sujeito se depara com novas situações, para as quais já possui esquemas que lhe permitem a assimilação, ou por meio de mecanismos de ampliação de conhecimento chamados de acomodação. Esta sucessão de assimilações e acomodações têm como resultado a equilibração, um conceito central na teoria construtivista.

Este modelo tem como um de seus objetivos, “a criação de comunidades de aprendizagem que se assemelhem ao máximo com a prática colaborativa do mundo real” (TORRES; Irala, 2014, p.72).Neste tipo de comunidade, espera-se que os alunos assumam a responsabilidade por sua própria aprendizagem. Uma das premissas é que cada indivíduo traz uma perspectiva diferente sobre o tema a ser estudado, viabilizando a criação de ambientes de negociação e geração de significados e solução de problemas por meio de entendimento compartilhado (TORRES; Irala, 2014). Além disso, para Caetano (2010), a relação entre os indivíduos possibilita a construção do conhecimento das regras que organizam a convivência.

Este aspecto da abordagem construtivista é bastante relevante, pois, há “um paralelismo entre o desenvolvimento intelectual e o desenvolvimento moral. O segundo depende do desenvolvimento do primeiro, tendo-o como uma condição necessária” (CAMARGO; BECKER, 2012, p.528). Nesse sentido, o desenvolvimento de relações de respeito mútuo, sem uma hierarquia

definida e que prioriza a autonomia em detrimento da heteronomia, em uma relação que Piaget chamou de cooperação – não confundir com a definição de aprendizagem cooperativa descrita no capítulo 1.2 – é também um conceito Piagetiano chave, embora menos estudado, se comparado com outros (CAMARGO; BECKER, 2012). Todas estas características influenciaram significativamente as teorias que fundamentam a aprendizagem colaborativa (TORRES; Irala, 2014).

A teoria sociocultural de Vygostky compartilha com a epistemologia genética, a ideia construtivista central de que "a única aprendizagem significativa é aquela que ocorre através da interação entre sujeito, objetos e outros sujeitos" (COELHO; PISONI, 201, p.146). Para Vygotsky (1998), existem certas ações propositais determinantes de todo o desenvolvimento e aprendizagem humanos. Estas ações são mediadas por diferentes ferramentas, dentre as quais, a mais importante é a linguagem, que "representa o sistema semiótico que é a base do intelecto humano" (TORRES; Irala, 2014, p.73).

Com isso, as interações sociais, mediadas pela linguagem são a base para o desenvolvimento das funções superiores – aquelas próprias do ser humano –, sendo elas: "o controle consciente do comportamento, a ação intencional e a liberdade do indivíduo em relação às características do momento e do espaço presente" (COELHO; PISONI, 201, p.146).

Desta forma, indivíduos são capazes de se apropriar dos modos de funcionamento psicológico, dos comportamentos e da cultura, por meio da interação social. Esta apropriação se dá em um processo de internalização do conhecimento (VYGOTSKY, 1998). Tal processo, por sua vez, pode ser explicado da seguinte forma:

> Um processo interpessoal é transformado num processo intrapessoal. Todas as funções no desenvolvimento da criança aparecem duas vezes: primeiro, no nível social, e, depois, no nível individual; primeiro, entre pessoas (inter,psicológica), e, depois, no interior da criança (intrapsicológica). Isso se aplica igualmente para a atenção voluntária, para a memória lógica e para a formação de conceitos. Todas as funções superiores originam-se das relações reais entre indivíduos humanos (VYGOTSKY, 1998, p.41).

Nesse contexto, o meio social influencia no aprendizado e no desenvolvimento dos sujeitos, de modo que ao chegarem à escola, já dispõem de uma série de conhecimentos adquiridos. Estes conhecimentos são chamados de conceitos cotidianos ou espontâneos, e são construídos por meio da vivência prática, observações e manipulações diretas. Os conhecimentos desenvolvidos na escola, por outro lado, são chamados de conceitos científicos e caracterizam aqueles não acessíveis à observação ou ação imediata do sujeito (COELHO; PISONI, 201, p.149).

Na teoria sociocultural, em suma, a aprendizagem é considerada um processo contínuo e a educação se caracteriza por saltos qualitativos de um nível de aprendizagem para outro. Com isso, segundo Coelho e Pisoni (201), dois tipos de desenvolvimento podem ser identificados:

- Desenvolvimento real: que se refere às capacidades ou funções já consolidadas e que o sujeito é capaz de realizar por si só;
- Desenvolvimento potencial: determinado pelas capacidades ou funções que para serem executadas necessitam da orientação ou colaboração de outro indivíduo mais capaz;

A distância entre estes dois níveis de desenvolvimento é chamada de Zona de Desenvolvimento Proximal (ZDP), um importante conceito que se refere ao período em que o sujeito necessita de um apoio até que seja capaz de realizar determinada ação por si só (COELHO; PISONI, 201). A partir destes conceitos chave, pode-se afirmar que a interação de um aprendiz com companheiros mais capazes pode conduzi-lo ao aprendizado por meio de um esforço colaborativo. A mudança cognitiva ocorre, nesse contexto, quando, dentro da ZDP, as interações culturalmente mediadas são internalizadas, transformando-se em novas funções do sujeito (TORRES; Irala, 2014).

A Pedagogia Progressista, por fim, abrange "as tendências que, partindo de uma análise crítica das realidades sociais, sustentam implicitamente as finalidades sociopolíticas da educação" (LIBANEO, 1983, p.9). Para Rodrigues *et al.* (2020), a pedagogia progressista se caracteriza por um processo de busca por transformação social, instigando o diálogo e a discussão coletiva como forças propulsoras de uma aprendizagem significativa, contemplando as parcerias, o trabalho coletivo e a crítica reflexiva dos alunos e professores.

Nesse sentido, adeptos desta tendência acreditam que o desenvolvimento individual acontece por meio do compartilhamento de ideias, informações, responsabilidades, decisões e cooperação entre os sujeitos (RODRIGUES *et al.*, 2020). Por orientar uma educação que prioriza a transformação social em detrimento das necessidades individuais, a pedagogia progressista promove a aprendizagem colaborativa ao inserir a educação em um papel sociopolítico, contrária ao autoritarismo e estimulando a aprendizagem grupal, que conduz à transformação intelectual e social por meio da negociação e do diálogo. Esta abordagem tende a valorizar e estimular a experiência de vida e a gerência do processo educacional pelos próprios alunos (TORRES; Irala, 2014).

Em suma, “a pedagogia da Escola Nova e a Pedagogia Progressista, juntamente com as teorias cognitivas formuladas por Piaget e Vygotsky, formam, indubitavelmente, as bases da aprendizagem colaborativa” (TORRES; Irala, 2014, p.74). Seja pela mudança de paradigma na relação entre professor e aluno e na necessidade de uma apreensão crítica e prática dos conteúdos, no caso das duas primeiras, seja por uma nova compreensão acerca dos processos de construção do conhecimento e do papel da interação entre sujeito e objeto, trazida por Vygostsky e Piaget. Para Torres e Irala (2014)), a aprendizagem colaborativa apresenta-se hoje como uma abordagem diferenciada por suas características próprias que representam um desdobramento teórico e metodológico das teorias e pedagogias que a embasam.

Assim, é importante considerar que o conhecimento é o resultado de uma construção social e que a aprendizagem colaborativa se constitui de métodos e técnicas de aprendizagem em grupo, sendo realizada em ambientes onde o processo educativo é favorecido pela interação social, estimulando a participação ativa e a interação entre os sujeitos (SANTOS *et al.*, 2020). Trata-se, pois, de uma alternativa adequada às necessidades contemporâneas. No próximo capítulo serão apresentadas algumas soluções baseadas em computador para apoiar esta abordagem.

2.3.2 Aprendizagem Colaborativa com Suporte Computacional

A aprendizagem mediada por Tecnologias de Informação e Comunicação (TIC) estabeleceu-se como um importante tópico, ganhando cada vez mais espaço nas discussões acadêmicas (CARNEIRO *et al.*, 2020). O cenário de afastamento social causado pela pandemia de COVID-19 fez com que o interesse pelo desenvolvimento de ferramentas educacionais se tornasse ainda mais evidente (ISOHäTäLä *et al.*, 2021) . Além disso, “a evolução do vínculo entre educação e tecnologia tem ampliado as possibilidades de proporcionar ambientes educacionais colaborativos” (SANTOS *et al.*, 2020, p.92).

Sendo assim, muitas perspectivas contribuem para a compreensão do computador no suporte à aprendizagem colaborativa, visando agregar recursos capazes de prover e facilitar as interações no processo de colaboração. O conjunto de esforços que visa produzir e organizar conhecimentos para esta finalidade está organizado em um campo chamado de Aprendizagem Colaborativa com Suporte Computacional – do inglês, *Computer Supported Collaborative Learning* (CSCL), que é um ramo emergente da ciência da aprendizagem preocupado em estudar como as pessoas podem aprender juntas com a ajuda de computadores (STAHL *et al.*, 2006).

A CSCL, segundo Stahl *et al.* (2006), surgiu no início da década de 1990 em reação ao

paradigma de software que forçava os alunos a aprenderem como indivíduos isolados. Somado a isto, “a internet e sua popularização para a sociedade através das tecnologias digitais norteou o surgimento de novas estruturas, metodologias e abordagens no ensino em todos os níveis educacionais” (CARNEIRO *et al.*, 2020, p.2). Sendo assim, para Lämsä *et al.* (2021), desde que foi criada, a CSCL vem recebendo contribuições de uma ampla variedade de teorias educacionais, tecnologias e metodologias. Esta variedade de diferentes abordagens e recursos a torna uma área bastante complexa e com pouco consenso, entre os pesquisadores que a integram, acerca das mais efetivas soluções.

Ainda assim, o campo da CSCL tem produzido ao longo dos anos, uma ampla evidência empírica para apoiar os princípios subjacentes ao projeto e uso efetivo de tecnologias capazes dar suporte a interações cognitivas e sociais, planejadas para facilitar a colaboração (JäRVELä; ROSé, 2021). As mesmas autoras (2021), afirmam também, que estas evidências sugerem que a aprendizagem colaborativa é bastante efetiva, inclusive em contextos de aprendizagem remota.

Entretanto, uma atividade colaborativa suportada por computador não é capaz de garantir que a aprendizagem seja produzida (LäMSä *et al.*, 2021). Costaguta *et al.* (2019), sustentam este posicionamento ao afirmar que o fato de organizar os alunos em grupos e fazer com que eles resolvam problemas ou atividades de forma colaborativa, também não garante que a experiência de aprendizagem seja exitosa. Do mesmo modo, é possível que um grupo de estudantes façam uso de tecnologias digitais como bate-papos, fóruns de discussão, e-mails e compartilhamento de arquivos como meios de facilitar as interações entre os membros e, ainda assim, não colaborem de forma efetiva.

Nesse sentido, é importante salientar que em processos de aprendizagem colaborativa existem intensas trocas de informações entre os sujeitos envolvidos. A CSCL precisa viabilizar e promover meios para que estas trocas sejam realizadas de modo a possibilitar a aprendizagem. Por esta razão, mais do que disponibilizar recursos tecnológicos diversos, é necessário refletir sobre em que momento didático cada tipo de sistema digital deve ser aplicado e quais os possíveis desdobramentos de sua utilização (CASTRO; MENEZES, 2011).

Para facilitar o processo de desenvolvimento de sistemas colaborativos, diferentes modelos e *frameworks* têm sido propostos e utilizados no campo da CSCL (COLLAZOS *et al.*, 2019). Entre estes modelos, se destaca o modelo 3C, que busca analisar a colaboração em três diferentes dimensões: comunicação, coordenação e cooperação (FUKS *et al.*, 2011). Estas dimensões podem ser explicadas da seguinte forma:

> A comunicação é caracterizada pela troca de mensagens, pela argumentação e pela negociação entre pessoas; a coordenação é caracterizada pelo gerenciamento de pessoas, atividades e recursos; e a cooperação é caracterizada pela atuação conjunta no espaço compartilhado para a produção de objetos ou informações (FUKS *et al.*, 2011, p.24).

É importante observar que alguns autores usam o termo **colaboração/cooperação** em substituição a **cooperação** (COLLAZOS *et al.*, 2019). As diferenças entre ambos foram discutidas no capítulo 2.2, entretanto, no contexto do modelo 3C, para os fins do presente trabalho, cooperação caracterizará atividades desenvolvidas em conjunto em um dado espaço compartilhado (FUKS *et al.*, 2008). A figura 4 apresenta um exemplo de como os sistemas colaborativos poderiam ser classificados no modelo 3C.

Figura 4 – Posicionamento dos sistemas colaborativos no espaço 3C.

Fonte: Fuks *et al.* (2011).

Esta classificação leva em conta o grau de suporte que cada sistema presta a cada dimensão. Entretanto, Collazos *et al.* (2019)) afirma que a dimensão de cooperação está situada em um nível mais alto de abstração e que o sucesso da cooperação depende de uma boa comunicação e coordenação. Fuks *et al.* (2011) reforça esta ideia afirmando que ainda que o objetivo principal de um sistema seja dar suporte a um determinado C, também é preciso dar suporte para os outros dois Cs. Isto porque a separação das atividades de comunicação, coordenação e cooperação durante a execução de um trabalho colaborativo não é tão óbvia. Além disso, muitas vezes estas atividades são interdependentes ou se confundem.

Sendo assim, Castro e Menezes (2011) apontam algumas necessidades específicas da CSCL e como atendê-las por meio do uso de tecnologias digitais. Segundo estes autores, a

necessidade mais emergente é a de comunicação entre os atores envolvidos. Para atender a esta demanda, podem ser usados recursos como e-mails, que possibilitam comunicação assíncrona, observando que a utilização de e-mails requer alguns cuidados, como acordo prévio quanto à frequência de envios e procedimentos de contingência, caso o serviço se torne indisponível, por exemplo.

Ainda para atender a necessidades de comunicação, podem ser usados fóruns de discussão, que podem servir como meios para promover discussões e construções coletivas de ideias. Ao utilizar fóruns, entretanto, é necessário considerar estratégias para evitar fugas aos temas propostos e promover a participação de todos. Por fim, os bate-papos e reuniões por vídeo ou áudio conferências podem ser usados para comunicação e troca de ideias de forma síncrona, mas seu sucesso depende de planejamento quanto à forma de participação (espontânea, induzida, aleatória), a negociação de agendas, entre outras variáveis (CASTRO; MENEZES, 2011).

Para além das necessidades de comunicação, entretanto, existem também demandas para atividades de coordenação e cooperação. A este respeito, destacam-se sistemas para suporte à formação de grupos, que auxiliam na organização de equipes de acordo com critérios preestabelecidos, sistemas para construção de sínteses, como glossários e mapas conceituais e sistemas de organização de conteúdos, como repositórios online (CASTRO; MENEZES, 2011).

Carneiro *et al.* (2020), nesse contexto, chamam a atenção para o uso emergente de redes sociais educativas usadas por comunidades virtuais de aprendizagem para construção e troca de conhecimento. Para estes autores (2020) as redes sociais educativas possibilitam a interação entre os sujeitos e a troca permanente de conhecimento por meio de atividades de aprendizagem e aperfeiçoamento. Uma pesquisa realizada por Jong *et al.* (2021) sobre comunidades virtuais formadas por professores usando uma plataforma de *Inquiry Learning Space* (ILS), mostrou que a troca de conhecimentos entre os docentes influenciou positivamente, tanto seu conhecimento pessoal, quanto suas práticas em sala de aula.

Além disso, com finalidade bastante específica, existe ainda uma classe de sistemas conhecida como *groupware*. Segundo Collazos *et al.* (2019) *groupwares* se diferenciam de outros tipos de software por tornar os seus usuários conscientes (*aware*) de que são partes de um grupo. Esta conscientização acontece ao introduzir mecanismos capazes de manter os membros do grupo atualizados sobre o estado e as mudanças do espaço virtual compartilhado, sobre as ações que outros membros estão realizando e, mais recentemente, sobre estados emocionais e movimentação dos olhos dos colegas na tela do computador usando inteligência artificial

(HAYASHI, 2020; CHANEL *et al.*, 2016). Estas características baseiam-se na premissa de que:

> Perceber, reconhecer e compreender as atividades de outras pessoas são requisitos básicos para a interação e comunicação humana em geral. O desenvolvimento de um comportamento humano adequado requer consciência sobre as pessoas e objetos de trabalho (COLLAZOS *et al.*, 2019, p.4792, tradução nossa).

Este tipo de informação é chamado de informação de conscientização (DOURISH; BELLOTTI, 1992).

Por fim, existem dois recursos bastante difundidos e igualmente importantes para o campo da CSCL: *scripts* e *prompts*. De acordo com Morris *et al.* (2010), *scripts* são instruções sobre como os membros do grupo devem colaborar e desenvolver tarefas por meio de suas respectivas funções. Em suma, são apenas dicas ou receitas que representam uma função específica ou coordenam diferentes funções. Os *scripts* podem ser divididos em sociais e epistêmicos, sendo os primeiros, responsáveis por descrever como estruturar e sequenciar o discurso e as atividades colaborativas. Os *scripts* epistêmicos, por sua vez, servem para descrever processos e estratégias cognitivas a serem usadas na resolução da tarefa (MORRIS *et al.*, 2010).

Prompts, por outro lado, são usados para apresentar dicas, sugestões ou lembretes durante a execução de alguma tarefa ou papel. São bastante usados para facilitar a interpretação dos papeis e a execução de *scripts* (MORRIS *et al.*, 2010). Um exemplo de utilização de *prompts* que merece destaque é a utilização de um agente pedagógico conversacional que trabalha de forma colaborativa com o aluno (HAYASHI, 2020). Este agente é capaz de analisar o texto de um aluno e fornecer automaticamente instruções explícitas, como: “Lembre-se de que a tarefa é explicar o tópico usando os dois conceitos” ou “Tente considerar o conceito que você está explicando usando outros exemplos.” (HAYASHI, 2020, p.480).

A CSCL é um campo multidisciplinar, que abriga múltiplas abordagens e diferentes tecnologias, tornando-se, com isso, tão rica quanto desafiadora (ISOHäTäLä *et al.*, 2021). Para Järvelä e Rosé (2021), é necessário que haja uma integração teórica mais intensiva das perspectivas cognitivas e sociais sobre a aprendizagem com o campo da CSCL para que seja possível construir soluções mais efetivas. Além disso, mais esforços no sentido de aplicar tecnologias de inteligência artificial também se fazem necessárias no sentido de melhorar a capacidade de produzir conscientização (*awareness*) (HAYASHI, 2020).

2.3.3 Engajamento na Aprendizagem Colaborativa

O engajamento é considerado uma estratégia de motivação importante na aprendizagem colaborativa (BRITO, 2010). No contexto da CSCL pode ser entendido como a participação dos estudantes nos processos de aprendizagem em ambientes interativamente sociais, por meio do compartilhamento de tarefas (BARBOZA; GIORDAN, 2008). Isto é importante porque alunos engajados na mesma tarefa têm a necessidade de se comunicar, compartilhar, coordenar e negociar significados, crenças, saberes e valores (BARBOZA; GIORDAN, 2008).

Para muitos autores, o engajamento escolar é um conceito complexo e multifacetado, tendo sido confundido com motivação (BORGES *et al.*, 2005). Entretanto tais noções referem-se a processos sociopsicológicos distintos, havendo o consenso de que o engajamento surge da interação do sujeito com o contexto, sendo influenciado pelas mudanças do ambiente (FINN; ROCK, 1997). Esta definição constitui a base para a crença de que alterações no contexto de ensino e nos ambientes de aprendizagem possam promover alterações nas diferentes facetas do engajamento dos estudantes (BORGES *et al.*, 2005).

Segundo Veiga (2013), existem quatro dimensões de engajamento: o engajamento comportamental, o engajamento cognitivo, o engajamento emocional e o engajamento agenciativo. A primeira, diz respeito ao envolvimento do estudante na realização da atividade, a segunda considera o esforço do estudante para compreender ideias complexas e deter habilidades difíceis. A dimensão emocional, por sua vez, está relacionada, como o nome sugere, às emoções positivas ou negativas do estudante quanto aos fatores e recursos de aprendizagem. Por fim, a dimensão agenciativa refere-se ao entendimento dos estudantes como agentes ativos, com iniciativas, intervenções, diálogos, sugestões e questões levantadas (VEIGA, 2013).

Existem atualmente, diversas formas de mensurar o engajamento em suas diferentes dimensões. Segundo um relatório de 2011, chamado *Measuring student engagement in upper elementary through high school: a description of 21 instruments*, existiam até então mais de cinco mil trabalhos publicados sobre o tema, dentre os quais, alguns pesquisadores se dedicam à criação de escalas para mensurar o engajamento estudantil (FREDRICKS *et al.*, 2011).

Trabalhos mais recentes, entretanto, mostram que questões relacionadas ao engajamento de estudantes são bastante atuais. Nesse contexto, Nascimento e Padilha (2020) apresentam resultados sobre o engajamento de estudantes de graduação na modalidade híbrida. Almulla (2020), por sua vez, busca demonstrar a efetividade da Aprendizagem Baseada em Projetos para

a promoção de engajamento entre os estudantes. Ansari e Khan (2020) apresentam o resultado de uma investigação sobre o papel das redes sociais na construção do engajamento em alunos de graduação. E por fim, um estudo de 2022 mostra como a aprendizagem colaborativa online é capaz de promover engajamento (NG *et al.*, 2022).

3 ETHOSCHOOL: UM SISTEMA MULTI AGENTE ÉTICO PARA APRENDIZAGEM COLABORATIVA

O presente capítulo apresenta uma proposta de solução para o problema de alinhamento de valores em IA, usando ética por design para auxiliar equipes de aprendizagem colaborativa. Sendo assim, no intuito de tornar tal proposta mais compreensível e objetiva, os próximos capítulos que a descrevem estão divididos em: proposta de arquitetura e requisitos funcionais; modelos estrutural e comportamental; e implementação e testes da solução proposta.

3.1 PROPOSTA DE ARQUITETURA E REQUISITOS DO ETHOSCOOL

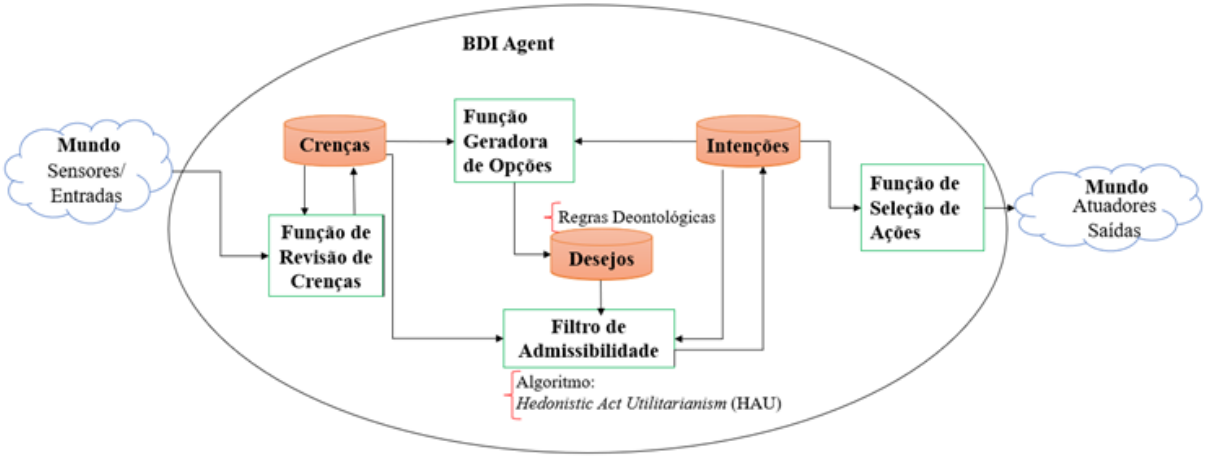
Conforme exposto no capítulo 2.2, existem três abordagens para a implementação de AMAs, sendo elas *top-down*, *bottom-up* ou híbrida (ALLEN *et al.*, 2005). Estas abordagens têm sido aplicadas em diferentes áreas no intuito de resolver problemas relacionados ao alinhamento de valores em IA. Contudo, quando se trata de contextos educacionais, em especial, em ambientes de sala de aula, é necessário ter um cuidado maior. Em tais casos, pelo fato de estar em jogo o processo educacional de pessoas, é muito importante evitar, o quanto possível, qualquer possibilidade de que AMAs aprendam comportamentos éticos enquanto perseguem seus objetivos pedagógicos, razão pela qual, a abordagem *bottom-up* ou híbrida não seriam adequadas (CORDOVA; VICARI, 2021).

O problema com estas abordagens é que elas abrem possibilidades para a ocorrência de enviesamento de dados e consequente replicação de preconceitos humanos (ARNOLD *et al.*, 2017). Em contextos educacionais esta é uma característica altamente indesejável, pois, dada à complexidade deste campo, que envolve as dimensões pedagógica, social e pessoal dos alunos, a tarefa de estabelecer a prioridade correta com relação a valores morais não deve ser delegado à sistemas de IA.

Por esta razão, propõe-se a aplicação da abordagem *top-down* para a implementação de AMAs de apoio à aprendizagem (CORDOVA; VICARI, 2021). AMAs desenvolvidos com base nesta abordagem implementam teorias éticas ou princípios morais que são usados como critérios para a seleção de ações eticamente apropriadas (ALLEN *et al.*, 2005), considerando capacidades morais como a aplicação de princípios éticos a casos particulares (MISSELHORN, 2020). Assim, a presente proposta adota o modelo BDI clássico com uma implementação híbrida das estruturas éticas deontológica e utilitarista.

Para tanto, em sua dimensão deontológica, o agente faz uso de princípios morais e critérios éticos para tomada de decisão, enquanto a sua dimensão utilitarista será usada para lidar com dilemas éticos. A figura 5 apresenta um modelo de alto nível, baseado na arquitetura BDI clássica para representar a ideia proposta.

Figura 5 – Arquitetura do AMA Proposto.



Fonte: CORDOVA e VICARI (2021).

Conforme é possível observar, as características que tornam esta proposta de agente especificamente um AMA são: a implementação de um agente, cujos desejos são guiados por princípios morais e regras deontológicas a serem perseguidas entre seus objetivos; e a implementação de um algoritmo utilitarista baseado na teoria de Jeremy Bentham, chamado de Utilitarismo de Ação Hedonista, par lidar com dilemas éticos.

Com relação à sua dimensão deontológica, não faz parte do escopo da presente pesquisa determinar quais princípios morais ou código de ética devem ser implementados em sala de aula. A solução aqui proposta deverá ser adequável à diferentes contextos e métodos de ensino. Neste trabalho, entretanto, de modo específico, o agente será modelado para atuar em ambientes de ensino colaborativo.

Assim, como não foi possível encontrar um conjunto de princípios morais ou regras éticas definidas especificamente para este tipo de ambiente, propõe-se alguns exemplos baseados nas características do ensino colaborativo; nos princípios para ética em IA, propostos pela UNESCO (2020); e nos princípios éticos de Beauchamp e Childress (2001) para Biomedicina, que já foram utilizados em um AMA para cuidados médicos (ANDERSON *et al.*, 2005).

Com isso, o quadro 2 apresenta algumas das regras a serem implementadas, bem como as referências que as embasam. No mesmo quadro, também é possível observar a relação de tais

regras com alguns dos princípios éticos para IA – os quais inspiram um cuidado mais direto – que constam na proposta da UNESCO (2020).

Quadro 2 – Exemplos de Regras e Princípios Éticos a Serem Implementados.

Princípios Éticos para IA Relacionados	Regra	Objetivo	Referência
Justiça e Não Discriminação	Todos os alunos deverão interagir	Maximizar a colaboração	(BRUFFEE, 1995); (DILLENBOURG, 1999); (DAVIDSON; MAJOR, 2014); (COLLAZOS, 2019); (UNESCO, 2020)
	Todos os alunos deverão interagir de forma equitativa		
Transparência e explicabilidade	Todos os alunos deverão ser informados sobre as interações dos seus colegas		
Justiça e Não Discriminação	Todos os alunos deverão entregar as atividades no prazo	Maximizar comportamento responsável	(UNESCO, 2020)
Privacidade Proporcionalidade e não causar danos	Não interferir na autonomia do aluno	Atender ao princípio do respeito à autonomia	(JACOBS, 2015); (BEAUCHAMP; CHILDRESS, 1979); (UNESCO, 2020)
Justiça e Não Discriminação Supervisão humana e determinação	Conduzir todos os alunos ao sucesso	Atender ao princípio da beneficência	(BEAUCHAMP; CHILDRESS, 1979); (UNESCO, 2020)
Proporcionalidade e não causar danos	Preservar a autoestima do aluno	Atender ao princípio da não maleficência	(BEAUCHAMP; CHILDRESS, 1979); (UNESCO, 2020)

Fonte: Autoria própria.

As regras apresentadas são fundamentadas em algumas das características esperadas de um grupo de aprendizagem colaborativa e em algumas das funcionalidades esperadas de um sistema no contexto da CSCL. Com relação aos princípios éticos para IA, o quadro 2 apresenta apenas aqueles que estão diretamente relacionados às regras do agente e que, por esta razão, precisam de uma maior atenção.

Entretanto, tais princípios, principalmente os não constantes no referido quadro, não se concretizam como requisitos funcionais, ou seja, não constituem funções explícitas a comporem o comportamento do agente. Entretanto, alguns deles são altamente desejáveis e, com isso, passam a ser requisitos não funcionais do modelo proposto.

Por exemplo, os princípios de segurança e proteção, assim como o de responsabilidade e prestação de contas, que não constam no quadro 2, são notoriamente relevantes e precisam ser considerados em todo o ciclo de vida do desenvolvimento do AMA proposto. O princípio de supervisão humana e determinação, por outro lado, mesmo constando no quadro 2, apenas se concretiza como a necessidade de atenção para que decisões críticas, como de aprovação

ou reprovação de alunos não sejam delegadas ao agente. Tal natureza de decisão está fora do escopo do agente, que, em razão disso, não estará sujeito à supervisão humana no escopo das suas decisões.

Além disso, alguns princípios morais deverão ser atendidos como requisitos funcionais, sendo eles: o princípio do respeito à autonomia – como a "capacidade de autogovernar-se" (FILHO *et al.*, 2018, p.89) –; o princípio da beneficência; transparência e explicabilidade; e o princípio da não maleficência.

3.1.1 Contribuição Pedagógica do Ethoscool

Embora a maior contribuição da solução aqui proposta seja a inclusão da ética por design e de discussões de teor mais prático sobre ética em IA no contexto da Educação, tratamos aqui de um SMA pedagógico. Por esta razão, é necessário que haja um objetivo pedagógico a ser perseguido pelo agente.

Nesse sentido, pode-se afirmar que o objetivo central do Ethoscool é buscar o engajamento dos estudantes que compõe um dado grupo de aprendizagem colaborativa. O engajamento, conforme descrito no capítulo 3.2.3, é uma estratégia de motivação importante, podendo ser entendido como a participação dos estudantes em processos de aprendizagem interativamente sociais (BARBOZA; GIORDAN, 2008). Portanto, a importância de buscar o engajamento está relacionada ao fato de que alunos engajados na mesma tarefa têm a necessidade de se comunicar, compartilhar, coordenar e negociar significados, crenças, saberes e valores (BRITO, 2010). Tais comportamentos constituem as bases da aprendizagem colaborativa.

Assim, considerando que o foco na inclusão de princípios deontológicos coletivamente estabelecidos, bem como na implementação de raciocínio utilitarista para lidar com dilemas éticos, em caráter ainda experimental, caracterizaram o maior esforço de pesquisa, o Ethoscool se limitará a buscar o engajamento comportamental dos estudantes. Para isso, o SMA deverá considerar a quantidade de interações realizadas pelos alunos como métrica, perseguindo um cenário no qual todos os participantes de um determinado grupo de alunos interajam de forma equitativa.

Ao perseguir tal cenário, o Ethoscool deverá monitorar as interações realizadas pelos estudantes em um fórum de discussões, por meio do qual o grupo deverá interagir para resolver atividades propostas por um professor. Com isso, analisando a quantidade de interações por aluno e por grupo e levando em consideração regras e princípios deontológicos, bem como o

bem comum em seu raciocínio utilitarista, o AMA decidirá sobre a conveniência ou não da intervenção por meio de mensagens para estimular a participação dos alunos.

Tal processo será mais bem detalhado nos capítulos seguintes. O objetivo aqui é deixar claro que o Ethoscool deverá buscar o engajamento comportamental dos membros de um grupo de aprendizagem colaborativa.

3.1.2 Descrição dos Requisitos Funcionais

O Ethoschool é um SMA concebido com o objetivo de operar em um ambiente CSCL, interagindo com os alunos para promover a coesão e o engajamento da equipe de estudos. Além disso, visa garantir uma participação equitativa de todos os estudantes ao longo da atividade. Para tanto, a solução proposta consiste em um ambiente que incorpora dois agentes distintos: o Monitor e o Tutor.

O agente Monitor é responsável por buscar e organizar as interações realizadas pelos alunos, registradas no banco de dados do fórum. Essa busca ocorre periodicamente, sendo possível configurar o intervalo de tempo entre as buscas. Uma vez obtidos os dados de interação, o agente Monitor tem como plano compilar e enviar essas informações ao agente Tutor (CóRDOVA *et al.*, 2023).

O agente Tutor, por sua vez, é um AMA que segue a abordagem *top-down* e visa analisar a necessidade de intervenção do agente nas interações dos alunos no fórum. Tal abordagem, conforme já mencionado, foi escolhida por entender-se que a aprendizagem do comportamento ético é inadequada para o ambiente de sala de aula.

3.1.2.1 Dimensão deontológica do Ethoscool

Ao adotar a abordagem *top-down*, o Ethoscool precisa atender a regras e princípios éticos já estabelecidos. Para este fim, um conjunto de regras derivadas do quadro 2 foram definidas e deverão servir como objetivos a serem perseguidos pelo agente Tutor.

A Tabela 1 exibe as regras definidas, juntamente com as prioridades estabelecidas entre elas, com o propósito de reduzir possíveis conflitos. Dessa maneira, apenas os conflitos inevitáveis e, em certa medida, desejáveis deverão ser resolvidos pelo algoritmo de raciocínio utilitarista, o qual será responsável por solucionar os dilemas decorrentes desses conflitos.

A característica impositiva das regras apresentadas na Tabela 1 é direcionada ao agente

Tabela 1 – Regras que o Ethoscool deverá seguir.

ID	Prioridade	Regras
1	1	Algum aluno do grupo deve interagir.
2	2	Todos os alunos do grupo devem interagir.
3	3	Todos os alunos devem interagir de forma equitativa.
4	2	Os alunos devem continuar a interagir durante toda a atividade.
5	1	Todos os grupos devem concluir as atividades no tempo determinado.
6	Universal	Não interferir nas decisões dos alunos.
7	1	Evitar enviar mensagens desestimulantes.
8	1	Evitar expor o comportamento do aluno a qualquer outro ser humano.

Fonte: Baseado em (CóRDOVA *et al.*, 2023)

Tutor, que deverá buscar satisfazê-las, mesmo que, dadas as variáveis ambientais, nem sempre seja possível. Por exemplo, considerando a regra: **algum aluno deverá interagir**, caso nenhum aluno o faça, o plano será considerado falho. As intenções do agente não serão movidas. Da mesma forma, pode-se também notar algumas contradições entre as regras estabelecidas. Um exemplo dessas contradições está entre a regra: **não interferir nas decisões dos alunos** e **todos os alunos devem interagir**. Neste caso, fazer com que o aluno interaja pode implicar em intervir em suas decisões, ferindo o princípio da autonomia – autodeterminação quanto a não interagir. Tal contradição resultará em um conflito entre as regras do próprio agente.

Entretanto, alguns conflitos podem ser mitigados. Para este fim, será adotada uma abordagem *prima facie* mediante a priorização de algumas regras em relação a outras, como é possível ver na Tabela 1, na coluna prioridade. Esta priorização também visa organizar a ordem dos estados a serem perseguidos pelo agente Tutor.

3.1.2.2 Dimensão utilitarista do Ethoscool

Para os casos, como o citado acima, onde não seja possível mitigar os conflitos e o agente Tutor acabe preso em um dilema, entrará em ação a sua dimensão utilitarista, que deverá, usando o algoritmo HAU, decidir sobre a melhor ação a ser tomada, considerando o estado atual do ambiente. Este algoritmo, segundo Anderson e Anderson (2008), escolhe, diante de um conjunto de opções possíveis, aquela ação que resulte no maior prazer líquido ou felicidade, considerando igualmente todos os afetados por tal ação. Desse modo, para decidir sobre a ação correta, o algoritmo HAU requer como entrada: o número de pessoas afetadas e, para cada pessoa, a intensidade de prazer/desprazer – podendo ser, por exemplo, em uma escala de 2 a -2; a duração do prazer/desprazer – em dias, por exemplo; e a probabilidade de ocorrência desse prazer/desprazer para cada ação possível. Com estas entradas, o algoritmo calcula, para cada

pessoa, o produto da intensidade, da duração e da probabilidade de obter tal prazer líquido. Em seguida, soma os prazeres líquidos individuais para obter o Prazer Líquido Total (PLT). A equação 1 apresenta a fórmula para o cálculo do PLT.

$$PrazerLquidoTotal = \sum_{i=1}^{n}(intensidade_i.duracao_i.probabilidade_i) \quad (1)$$

Onde, n é o número total de pessoas afetadas pela ação. Assim, a ação com o maior prazer líquido total será considerada a ação correta. Escolheu-se este algoritmo para compor o modelo porque ele leva em consideração o prazer ou o desprazer de todas as pessoas envolvidas. Nesse contexto, como a presente proposta é idealizada para ser utilizada em um ambiente de sala de aula, onde os interesses das pessoas devem ser seriamente considerados, o algoritmo baseado na teoria de Jeremy parece bastante adequado.

Com relação aos demais requisitos funcionais do agente Tutor, a sua percepção está limitada a perceber as mensagens enviadas pelo agente Monitor, contendo os dados compilados das interações realizadas pelos alunos no fórum. Em seu plano está incluída a análise desses dados, sendo que suas ações, decorrentes desta análise, deverão ser executadas de acordo com as condições ambientais.

3.1.2.3 Cenários hipotéticos e escopo de ação do Ethoschool

Com o fim de ilustrar com mais clareza o comportamento do SMA proposto, a seguir serão apresentados alguns cenários, descritos por Córdova e Vicari (2022), a partir dos quais será possível compreender alguns estados possíveis para o ambiente no qual o Ethoscool irá operar e como ele deverá agir diante destes estados:

Seja um grupo de aprendizagem colaborativa formado por três estudantes, **a**, **b** e **c**, interagindo por meio de um fórum para resolver uma dada situação problema definida pelo professor. Tal atividade tem duração de dois dias – 48 horas.

- No primeiro cenário, após 8 horas o agente verifica que nenhuma interação foi realizada pelos estudantes. Esse estado ativa o seu desejo correspondente à regra 1 da tabela 1, já considerando a priorização entre as regras, que diz que **algum estudante deve interagir**. Entretanto, tal desejo conflita com outro, relacionado a regra 6, que diz para o agente **não interferir nas decisões dos alunos**. Este conflito leva o agente a um dilema ético, exigindo

uma solução antes que o agente se comprometa com qualquer intenção. Neste cenário, se o agente deliberar pela intervenção, será executada a ação enviar uma mensagem motivadora ao grupo. Do contrário, nenhuma ação é executada. O Ethoscool salvará a decisão tomada, bem como os dados usados para tomá-la.

- No segundo cenário, após 26 horas, os alunos **a** e **b** registraram 15 interações e o aluno **c**, nenhuma. Este estado ativa o desejo do agente Tutor relacionado à regra 2 da tabela 1, já considerando a priorização entre as regras, que diz que **todos os alunos do grupo devem interagir**. Este desejo conflita com aquele relacionado à regra 6, que diz para o AMA **não interferir nas decisões dos alunos**, levando o agente a um conflito entre suas próprias regras. Este conflito leva o agente a um dilema ético, exigindo uma resolução, por meio do seu raciocínio utilitarista, antes que o agente se comprometa com qualquer intenção. Se, neste cenário, o agente deliberar pela intervenção, será executada a ação enviar uma mensagem motivadora individual. Do contrário, nenhuma ação é executada. O agente Tutor salvará a decisão tomada, bem como os dados usados para tomá-la.

- No terceiro cenário, após 36 horas o agente verifica 3 interações do aluno **c**, 23 do aluno **b** e 17 do aluno **a**. Uma vez satisfeito o desejo correspondente à regra 2 da tabela 1, que diz que **todos os alunos devem interagir**, o AMA verifica desproporcionalidade nas quantidades de interação, despertando o seu desejo relacionado às regras 3, que diz que **todos os alunos devem interagir de forma equitativa**. Entretanto, mais uma vez a regra 6 conflita com o desejo do agente por **defender a autonomia do aluno**, o que leva a necessidade de resolver o conflito. Nestas condições, se o agente deliberar pela intervenção, será executada a ação enviar uma mensagem para instigar a equidade na participação dos alunos. Do contrário, nenhuma ação é executada. O agente Tutor salvará a decisão tomada e os dados usados para tomá-la.

- No quarto cenário já se passaram 40 horas e o agente verificou que há 11 interações do aluno **a**, 10 do aluno **b** e 13 do aluno **c**. Sendo assim, os desejos 1, 2 e 3 estão satisfeitos. Entretanto, a última interação entre os alunos se deu após 8 horas de atividade, levando o agente a constatar que os alunos estão há muito tempo sem interagir no fórum. Este estado provoca o desejo relacionado à regra 4 da tabela 1, que diz que **todos os alunos devem continuar a interagir durante toda a atividade**, que, mais uma vez, conflitará com a regra 6, evocando, dessa forma, o resolvedor de dilemas. Neste cenário, se o agente

deliberar pela intervenção, será executada a ação enviar uma mensagem de incentivo ao grupo. Do contrário, nenhuma ação é executada. O agente Tutor salvará a decisão tomada e os dados usados para tomá-la.

Os cenários apresentados, descrevem situações extremas, que forçam o agente a entrar em um dilema ético. Tal dilema pode ser expressado, conforme mostra a equação 2.

$$\forall x[\neg P(x) \rightarrow I(a,x)] \wedge \forall x \forall y[I(a,x) \wedge P(a,y) \rightarrow x = y] \wedge \forall x[\neg I(a,x)] \quad (2)$$

Onde:

- P(x): significa que o aluno x está interagindo adequadamente.
- I(a, x): significa que o agente a intervém no comportamento do aluno x.

Como é possível observar, a equação 2 expressa que, se o aluno **x** não está interagindo adequadamente, então o agente *a* deve intervir no comportamento do aluno **x**. Além disso, a fórmula também determina que o agente *a* jamais deve intervir no comportamento de mais de um aluno **x** e **y** simultaneamente. E, por fim, declara que o agente *a* jamais deve intervir no comportamento do aluno **x**. Isso cria um dilema, já que a primeira parte da fórmula indica que que o agente deve intervir no comportamento do aluno, caso este não esteja interagindo adequadamente, mas a última parte da fórmula nega que o agente deve intervir no comportamento de qualquer aluno. Além disso, em uma situação normal de execução, o professor deverá informar ao SMA quando um grupo concluir sua atividade para que o Ethoscool leve este estado em conta e pare de fazer as leituras no banco de dados para este grupo.

Como é possível observar por meio dos cenários descritos, a adequação do HAU ao Ethoschool se dará no nível de calibração dos seus parâmetros de entrada em relação ao tempo de atividade. No entanto, os fatores determinantes para a definição de tais parâmetros serão a mestria e a experiência do professor. Por exemplo, a depender das características dos alunos, o professor pode considerar que a intensidade de prazer total do grupo, que uma intervenção do agente poderá trazer aos alunos, será maior, quanto mais próximo do final do tempo da atividade. Enquanto a duração deste prazer e a probabilidade de que ele ocorra serão maiores ou menores. Como o que será considerado é o prazer geral, o agente poderá considerar que sua intervenção, mesmo trazendo menor prazer ao aluno que a sofreu, pode trazer maior prazer geral ao grupo

por contribuir com o sucesso da equipe. Tal característica possibilita a aplicação do conceito de priorização do bem comum.

Admite-se, neste ponto, a inclusão de viés humano, por meio da figura do professor. Contudo, tal viés não deve trazer malefícios ou benefícios a grupos ou indivíduos específicos, uma vez que o mesmo conjunto de parâmetros se aplicará a todos os grupos de estudantes. Além disso, esta inclusão de viés trará mais previsibilidade e explicabilidade ao agente Tutor, que tomará suas decisões baseadas nos valores parametrizados por um humano. Os capítulos seguintes detalham como se dará esta parametrização e como os requisitos de explicabilidade serão atendidos.

Por fim, embora a aprendizagem de comportamento ético seja uma característica indesejada em AMAs no contexto de ensino e aprendizagem, outros escopos para o uso de recursos de aprendizagem podem ser úteis. O agente proposto no presente trabalho não contará com mecanismos com esta finalidade, mas poderá se adaptar a diferentes condições do ambiente mediante a revisão e atualização das suas crenças. Isso não significa que o AMA irá aprender com casos de conflitos éticos por ele já resolvidos, mas que terá flexibilidade suficiente para atender às dinâmicas de um processo de ensino colaborativo.

3.2 ESPECIFICAÇÃO DOS REQUISITOS FUNCIONAIS

O Ethoscool, conforme já mencionado, irá operar como um sistema multiagentes. Nesse sentido, adotou-se o uso da Engenharia de Software Baseada em Agentes (AOSE[1]) para as diferentes fases do presente projeto. A AOSE dispõe de diferentes soluções para "abordar a crescente complexidade dos sistemas de computação que geralmente devem operar em ambientes não previsíveis, abertos e que mudam rapidamente" (GIRARDI, 2004, p.913).

Além disso, ela combina características e recursos da Inteligência Artificial e da Engenharia de Software (ES) e surge da percepção de que o desenvolvimento de sistemas multiagentes necessita da aplicação de práticas de ES, em razão da sua complexidade (GUEDES; VICARI, 2022). Assim, a AOSE dispõe de diferentes metodologias para o desenvolvimento de Sistemas Multiagentes, como: GAIA, Tropos, PASSI, O-MaSE, entre outros (ABDALLA; MISHRA, 2021). Entretanto, tais metodologias, em sua maioria, não são capazes de cobrir todo o Ciclo de Vida de Desenvolvimento de Software (SDLC[2]) (JAZAYERI; BASS, 2020), que deve incluir

[1] Da sigla em inglês *Agent-oriented Software Engineering*.
[2] Da sigla em inglês *Software Development Life Cycle*.

todos os processos necessários para especificar e transformar requisitos em produto de software entregável (BOURQUE; FARLEY, 2014).

Para ilustrar melhor esta questão, pode-se dizer que apesar das diferenças com relação ao nome dado a cada uma das etapas pelos diferentes modelos de processos existentes na AOSE, o SLDC, em geral, é composto por processos de Engenharia de Requisitos (RE[3]), análise, projeto, implementação e implantação (JAZAYERI; BASS, 2020). As atividades de RE procuram orientar a elicitação, análise, especificação e validação de requisitos de software, bem como o gerenciamento dos requisitos durante todo o SLDC (BOURQUE; FARLEY, 2014). Na etapa de análise, usando os requisitos identificados, são analisadas e criadas descrições abstratas de diferentes componentes do sistema.

As tarefas de projeto, por sua vez, visam transformar os modelos analíticos e abstratos em modelos concretos e implementáveis do sistema. A definição da arquitetura, dos componentes, interfaces entre outras características do sistema ou de componentes fazem parte deste processo. A implementação corresponde à criação detalhada de um software funcional por meio da codificação, verificação e testes unitários, de integração e depuração. (BOURQUE; FARLEY, 2014). Por fim, a implantação refere-se à operacionalização do sistema de software, que pode ser efetuada com o auxílio de modelos de implantação desenvolvidos na fase de projeto (JAZAYERI; BASS, 2020).

Assim, conforme já mencionado, as metodologias da AOSE disponíveis atualmente, além de serem mais focadas nos processos de requisitos, projeto e implementação, não permitem uma adequada integração entre as fases do SLDC e necessitam ainda serem aprimoradas para atender melhor às necessidades existentes (ABDALLA; MISHRA, 2021).

Por esta razão, optou-se por não aplicar nenhuma das referidas metodologias aos processos previstos para o presente trabalho. Ao invés disso, foram utilizadas diferentes metodologias, linguagens e ferramentas para diferentes etapas do projeto.

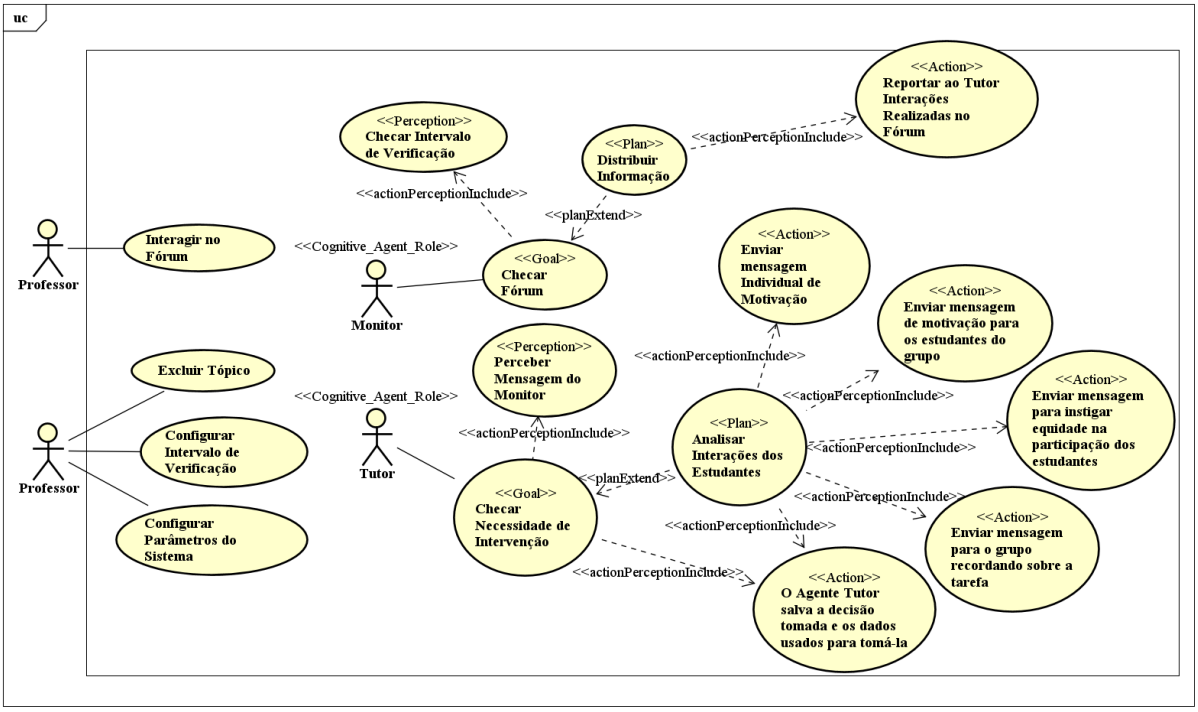
Para a especificação dos requisitos foi aplicada a *Multi-Agent Systems Requirements Modeling Language* (MASRML), uma linguagem específica de domínio baseada na UML para a modelagem de requisitos em projetos de SMA (GUEDES, 2020). Esta linguagem permite identificar os requisitos funcionais internos para SMAs, tais como: quais papeis de agente devem ser suportados, bem como objetivos associados a estes papeis; as condições que devem ser satisfeitas para que tais objetivos se tornem intenções; quais planos devem existir para que

[3] Da sigla em inglês *Requirement Engineering*

os objetivos sejam alcançados; quais as percepções de cada papel de agente; e quais as ações externas permitidas para uma dada função.

Assim, com base nesta linguagem, foi produzido um Diagrama de Caso de Uso (UCD), conforme mostra a Figura 6. As especificações de requisitos apresentadas a seguir foram adaptadas do artigo ETHOSCHOOL: *An Artificial Moral Agent Model for Collaborative Learning*, publicado nos anais do *19TH International Conference on Intelligent Tutoring Systems* (ITS), por Córdova *et al.* (2023).

Figura 6 – Representação dos Requisitos Funcionais Usando MASRML.



Fonte: Córdova *et al.* (2023).

Dito isso, é possível observar a partir do UCD mostrado na figura 6, a representação dos comportamentos do sistema. Neste UCD estão presentes atores e Casos de Uso (UC) UML padrão – sem estereótipos –, e também atores e casos de uso internos MASRML – sendo aos atores, aplicado o estereótipo *Cognitive_Agent_Role* e aos casos de uso, os estereótipos *Perception*, *Goal*, *Plan* e *Action*. Nesse sentido, o estereótipo *Cognitive_Agent_Role* representa o papel assumido pelos agentes no sistema, enquanto os casos de uso internos representam os comportamentos internos dos agentes.

Assim, os Casos de Uso Internos (IUC[4]) com o estereótipo *Goal* representam os objetivos (desejos) associados a um determinado *Cognitive_Agent_Role*, ou seja, representa

[4] Da sigla em inglês *Internal Use Case*

o comportamento de um agente que assume esse papel para determinar se o objetivo pode se tornar uma intenção. Para isso, em geral, esse tipo de UC precisa de uma ou mais formas de perceber o ambiente – *Perceptions* –, que possam mudar as crenças do agente e permitir que este determine quando o objetivo pode se tornar uma intenção. Com isso, um IUC com o estereótipo de *Perception* demonstra um comportamento no qual um agente sondará o ambiente para determinar se alguma de suas crenças deve ser alterada.

Por outro lado, IUCs com o estereótipo *Plan* representam o comportamento realizado por um agente quando um objetivo se torna uma intenção, ou seja, quando um agente passa a acreditar – muitas vezes devido à percepção de alterações no ambiente – que um objetivo pode ser alcançado e começa a agir para atingir tal objetivo. Por fim, IUCs com o estereótipo *Action* representam o comportamento que um agente executa para produzir uma ação que afeta o ambiente e pode ser percebida por outros agentes.

Portanto, ao examinar a figura 6, pode-se observar que os atores **Aluno** e **Professor** são atores normais da UML, representando papéis assumidos por usuários externos que operam o sistema. Enquanto **Monitor** e **Tutor** são *Cognitive_Agent_Role*, conforme mostram seus estereótipos, representando papéis que podem ser assumidos por agentes artificiais comprometidos com seus objetivos.

Ainda examinando a figura 6, pode-se notar que os casos de uso **Interagir no Fórum**, **Excluir Tópico**, **Configurar Intervalo de Verificação** e **Configurar Parâmetros do Sistema** são casos de uso normais, ou seja, comportamentos que podem ser executados pelos usuários que assumem o papel de professor ou aluno no sistema. Por outro lado, Os IUCs **Checar Fórum** e **Checar Necessidade de Intervenção** são objetivos, como se depreende de seus estereótipos. Esses IUCs estão associados aos *Cognitive_Agent_Role* **Monitor** e **Tutor**, respectivamente. Isso significa que os agentes que assumem esses papéis devem verificar se esses objetivos podem ser alcançados e agir sobre eles, quando for o caso.

Para que esses desejos (*Goals*) se tornem intenções, muitas vezes é necessário que os agentes recebam do ambiente, por meio de sua capacidade de percepção, informações que os façam acreditar que os objetivos podem ser alcançados. Para este fim, existem IUCs que representam percepções – usando o estereótipo *Perception* – associados a cada IUC *Goal*.

É importante notar que essas IUCs do tipo *Perception* estão associadas às IUCs do tipo *Goal* por meio de associações *actionPerceptionInclude*, o que é semelhante à associação de *include* na UML padrão, em que a execução da IUC *Goal* implica a execução da IUC *Perception*

a ela associada. Assim, sempre que os objetivos são executados pelos agentes, suas percepções também devem ser executadas para determinar se o objetivo pode ou não se tornar uma intenção.

No mesmo UCD, ainda é possível observar que existem IUCs com o estereótipo *Plan* associado à IUCs do tipo *Goal*. Esses IUCs representam os planos que serão executados por um agente se ele acreditar que seu objetivo poderá ser alcançado, tornando-se uma intenção. Sendo assim, os IUCs do tipo *Plan* são associados aos IUCs com o estereótipo *Goal* por meio de relacionamentos *PlanExtend*. Isso significa que os planos só serão executados mediante o cumprimento de determinada condição. Neste caso, a condição necessária é que o agente passe a acreditar que o objetivo em questão se tornou uma intenção. Para exemplificar o funcionamento da documentação MASRML, o Quadro 3, o Quadro 4 e o Quadro 5 apresentam a documentação referente ao Agente Tutor.

Quadro 3 – Descrição do Objetivo Checar Necessidade de Intervenção

Nome do IUC:	Checar Necessidade de Intervenção
Estereótipo:	*Goal.*
Cognitive_Agent_Role:	Tutor.
Descrição:	Este caso de uso interno descreve os passos seguidos pelo agente que assume a função de Tutor ao verificar a necessidade de intervenção e o tipo de intervenção em cada aluno ou grupo.
Crenças Iniciais:	Estudantes estão interagindo de forma apropriada.
Percepções:	Sondar novas mensagens do monitor.
Cenário Principal.	
Ações do *Cognitive_Agent_Role*	
1. Executar o caso de uso interno "Percebe mensagem do Monitor".	
Cenário Alternativo - Estudantes NÃO interagindo de forma apropriada.	
Ações do *Cognitive_Agent_Role*	
1. Evidencia que há alunos que não estão interagindo.	
2. Mudar a crença “Estudantes estão interagindo de forma apropriada” = falso.	
3. Avaliar necessidade de intervenção com base nas regras do agente.	
4. Se necessário, resolver o dilema ético usando o *Hedonistic Act Utilitarianism* (HAU).	
5. Transformar o objetivo em uma intenção.	
Cenário Alternativo 2 - Estudantes interagindo de forma apropriada.	
Ações do *Cognitive_Agent_Role*	
1. Evidencia que os estudantes estão interagindo adequadamente.	
2. Salvar a decisão e os dados usados para tomá-la.	

Fonte: Córdova *et al.* (2023).

Os quadros 3 e 4 descrevem os passos comportamentais relativos à percepção e plano associados ao objetivo **Checar Necessidade de Intervenção**. Este é o plano mais importante para a abordagem proposta neste trabalho, além de fazer parte da implementação da prova de conceito, que será apresentada no capítulo 3.4. Por esta razão, recebeu todo o foco na especificação.

Os UCDs e a domumentação dos IUCs, entretanto, não dispõe de recursos para repre-

Quadro 4 – Descrição da Percepção Perceber Mensagem do Monitor

Nome do IUC:	Perceber Mensagem do Monitor
Estereótipo:	*Perception.*
Cognitive_Agent_Role:	Tutor.
Descrição:	Agente Tutor sonda o ambiente em busca de mensagens do Agente Monitor.
Pré-condições:	*Goal* **Checar necessidade de intervenção** deve estar em execução.
Crenças Iniciais	Mensagem recebida do Agente Monitor = falso.
Cenário Principal.	
Ações do *Cognitive_Agent_Role*	
1. Sondar o ambiente em busca de novas mensagens do agente Monitor com dados das interações dos alunos.	
Cenário Alternativo - Mensagem enviada pelo Agente Monitor.	
Ações do *Cognitive_Agent_Role*	
1. Passa a crer que os dados das interações foram enviados.	
2. Recebe os dados das interações dos alunos.	

Fonte: Córdova *et al.* (2023).

Quadro 5 – Descrição do Plano Analisar Interações dos Estudantes

Nome do IUC:	Analisar Interações dos Estudantes
Estereótipo:	*Plan.*
Cognitive_Agent_Role:	Tutor.
	Agente Tutor analisa os dados de interações dos alunos no fórum.
Cenário Principal.	
Ações do *Cognitive_Agent_Role*	
1. Determinar o número de alunos que não está interagindo.	
Cenário Alternativo -Identificado que um estudante em determinado grupo não está interagindo.	
Ações do *Cognitive_Agent_Role*	
1. Enviar mensagem de motivação individual.	
2. O Agente Tutor salva a decisão tomada e os dados usados para tomá-la.	
Cenário Alternativo 2 - Identificado que alguns estudantes em determinado grupo não estão interagindo de forma equitativa.	
Ações do *Cognitive_Agent_Role*	
1. Enviar mensagem para instigar equidade na participação dos estudantes.	
2. Salvar a decisão e os dados usados para tomá-la.	
Cenário Alternativo 3 - Identificado que determinado grupo de estudantes não está interagindo	
Ações do *Cognitive_Agent_Role*	
1. Enviar mensagem para todos os estudantes do grupo visando promover a participação na atividade.	
2. Salvar a decisão e os dados usados para tomá-la.	
Cenário Alternativo 4 - Identificado grande intervalo de tempo desde a última interação de determinado grupo.	
Ações do *Cognitive_Agent_Role*	
1. Enviar mensagem para o grupo recordando sobre a tarefa.	
2. Salvar a decisão e os dados usados para tomá-la.	
Cenário Alternativo 5 - Raciocínio utilitarista delibera pela não intervenção	
Ações do *Cognitive_Agent_Role*	
1. Determinar inadequação da intervenção.	
2. Salvar a decisão e os dados usados para tomá-la.	

Fonte: Córdova *et al.* (2023).

sentar todos os aspectos técnicos do Ethoscool e do Agente Tutor. Sendo assim, a seguir, serão discutidos os detalhes das funcionalidades internas, que não podem ser representadas por estes componentes do projeto.

3.3 MODELAGEM DA SOLUÇÃO PROPOSTA

Seguindo a abordagem usada para a especificação de requisitos, não serão aplicadas metodologias específicas da AOSE para a modelagem estrutural e comportamental do SMA proposto. Para este fim, será aplicada a *Multi-Agent System Modeling Language* (MAS-ML). Esta linguagem é capaz de representar os aspectos estáticos e dinâmicos essenciais dos SMAs, enfatizando uma clara representação dos seus conceitos e relacionamentos (SILVA; LUCENA, 2003).

Além disso, as linguagens MASRML e MAS-ML complementam uma a outra, sendo, conforme mostrado, a primeira aplicada à etapa de especificação de requisitos e a segunda à etapa de modelagem (CóRDOVA *et al.*, 2023). Nesse sentido, nos próximos capítulos serão apresentados os modelos estrutural e comportamental do Ethoscool.

3.3.1 Modelo Estrutural

Modelos estruturais ajudam a ilustrar a composição física ou lógica do software a partir de seus componentes. Sendo assim, a modelagem estrutural estabelece a fronteira entre o software que está sendo implementado ou modelado e o ambiente no qual ele deve operar (BOURQUE; FARLEY, 2014).

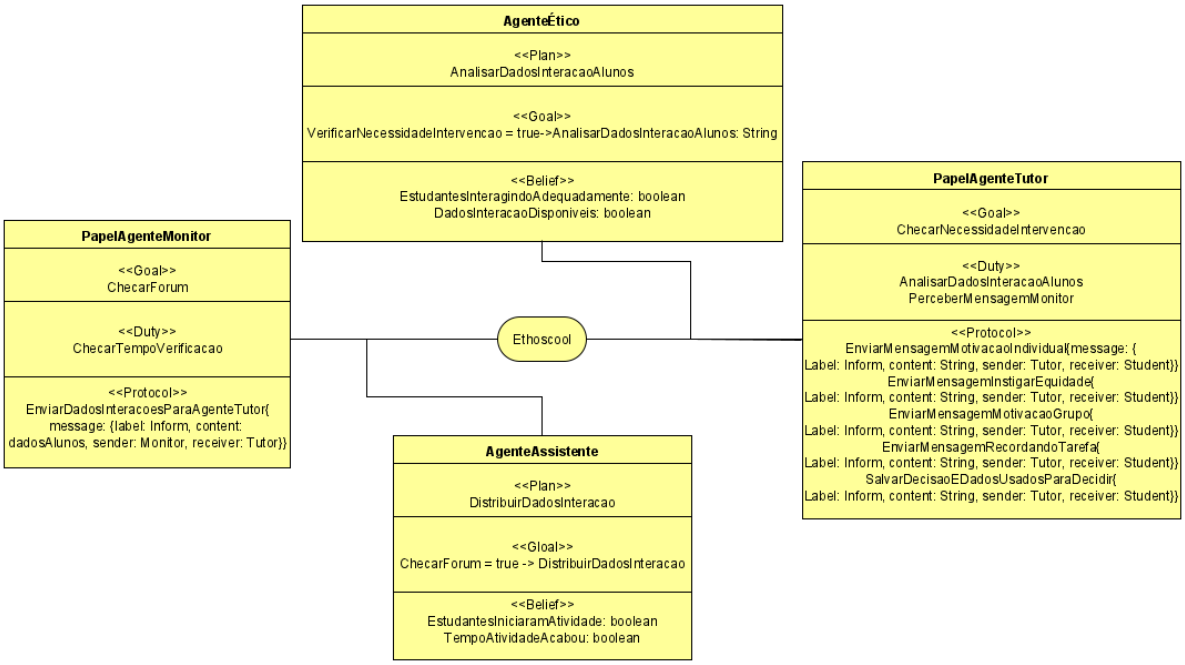
Com isso, a Figura 7 apresenta o modelo estrutural na forma de um diagrama organizacional da MAS-ML, que permite identificar os tipos de agentes suportados pelo sistema e os papéis que eles podem assumir em sua organização. A organização, por sua vez, representa o próprio sistema Ethoschool.

Desse modo, o diagrama apresentado na Figura 7 apresenta dois tipos de agentes e dois papéis de agente que podem ser assumidos. Isto pode ser observado por meio do relacionamento – linha contínua – que conecta os papeis **PapelAgenteTutor** e **PapelAgenteMonitor** à organização. Este relacionamento é chamado de relação de propriedade. Os agentes **AgenteÉtico** e **AgenteAssistente**, por sua vez, tem os papéis que podem assumir representados por meio do relacionamento chamado de relação de atuação – linha contínua – que os conecta à sua respectiva associação de propriedade.

Compreende-se, desse modo, que o Ethoschool suporta os papéis de agente **PapelAgenteMonitor** e **PapelAgenteTutor**, que são interpretados pelos agentes **AgenteAssistente** e **AgenteÉtico**, respectivamente, sendo estes últimos, classes de agentes. Nesse sentido, o **PapelA-**

genteMonitor tem o propósito de consultar o fórum para buscar o histórico de participação dos estudantes. Este objetivo será executado de acordo com o intervalo parametrizado pelo professor. Além disso, ele também usa o protocolo da *Foundation for Intelligent Physical Agents* (FIPA) para enviar os dados das interações realizadas pelos estudantes ao **PapelAgenteTutor**.

Figura 7 – Diagrama Organizacional do Ethoschool.



Fonte: Córdova *et al.* (2023).

O **AgenteAssistente**, por sua vez, tem o plano para distribuir os dados recuperados do fórum para o **PapelAgenteTutor**. Esse plano, entretanto, somente será executado quando o objetivo **ChecarForum** se tornar uma intenção. Além disso, o **AgenteAssistente** também possui crenças nas quais acredita, ou não, que os estudantes iniciaram uma nova tarefa no fórum e que esta atividade atingiu, ou não, o seu tempo limite.

Por outro lado, o **AgenteÉtico**, que assume o papel de Tutor tem como objetivo verificar a necessidade de intervenção para estimular a participação dos estudantes, analisando os dados de interação recuperados do fórum e enviados pelo **AgenteAssistente**, que assume o papel de Monitor. Também cabe ao **AgenteÉtico**, reconhecer, por meio de sua percepção, o recebimento de novas mensagens contendo estes dados. Além disso, o **PapelAgenteTutor** possui protocolos para enviar mensagens encorajadoras tanto para um indivíduo quanto para toda a equipe de estudantes, bem como o dever de salvar a decisão tomada e os dados utilizados para tomá-la.

Para isso, a análise dos dados das interações no fórum se materializa na forma de um plano de agente, que é acionado quando o objetivo de verificar a necessidade de intervenção do

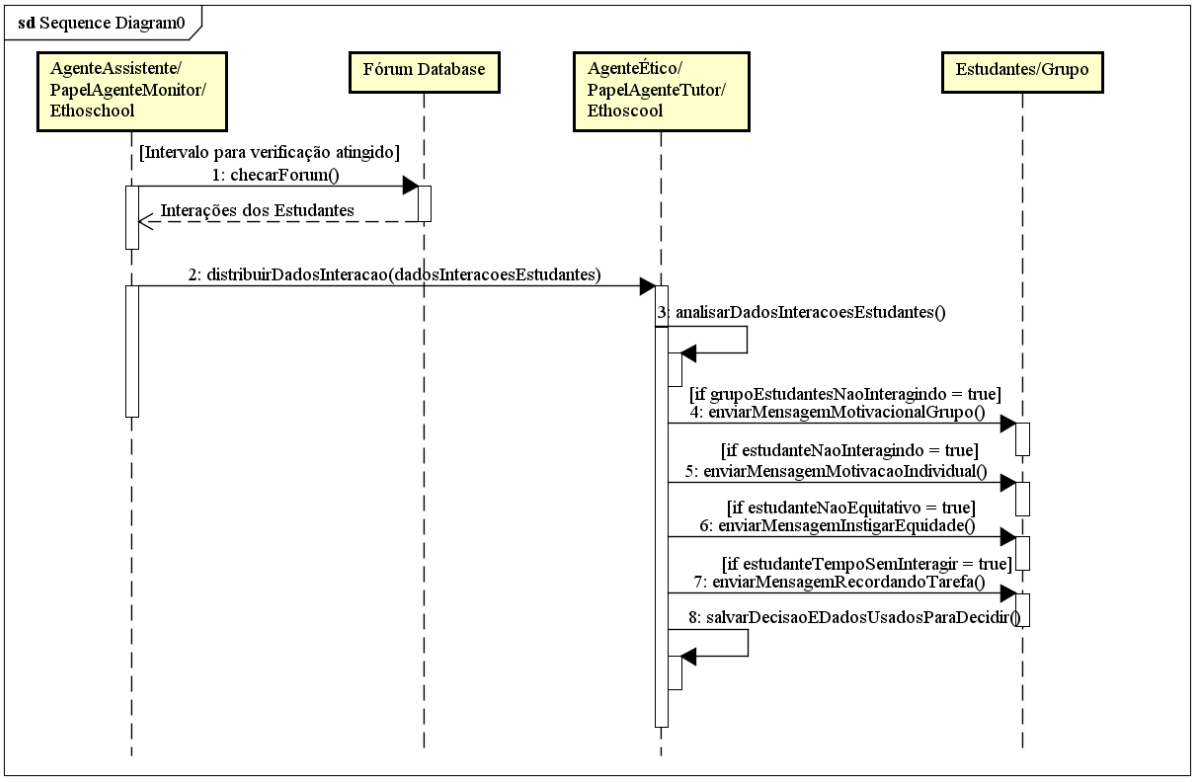
PapelAgenteTutor se torna uma intenção. Esse agente também possui crenças que determinam se é necessário ou não realizar essa análise, percebendo o recebimento de novas mensagens do agente que assume o papel de Monitor e se os alunos estão interagindo adequadamente por meio do fórum.

3.3.2 Modelo Comportamental

Segundo Bourque e Farley (2014), modelos comportamentais servem para identificar e definir as funções de um software. O modelo comportamental a ser apresentado neste capítulo assume a forma de um fluxo de controle, por meio do qual será possível retratar como uma sequência de eventos causa a ativação ou desativação de processos pré-estabelecidos.

Com isso, o diagrama escolhido para representar o modelo comportamental do funcionamento do agente Tutor é o de sequência, conforme apresentado na figura 8. Foi optado por mostrar o comportamento do plano do agente **Monitor** para buscar e distribuir os dados das interações dos alunos e do plano do agente **Tutor** para analisar tais dados. Buscando uma visão geral do funcionamento do agente, ambos os planos foram modelados em um único diagrama.

Figura 8 – Diagrama de Sequência do Ethoscool.



Fonte: Autoria própria.

Desse modo, como pode ser visto na figura 8, quando o intervalo entre as consultas ao banco de dados do fórum é atingido, o **AgenteAssistente** faz uma consulta no banco de dados do fórum, organiza estes dados e os envia para o **AgenteÉtico**. Conforme descrito no capítulo anterior, esta comunicação entre os agentes se dará por meio do formato FIPA. Na sequência, o **AgenteÉtico** realiza uma análise dos dados recebidos, verificando se há estudantes ou grupos que se encaixem em alguma das condições para ser considerado candidato a intervenção do agente.

Nesse contexto, o **AgenteÉtico**, conforme suas especificações, deverá ser capaz de lidar com quatro diferentes cenários:

- Caso seja identificado que em determinado grupo de estudantes nenhum aluno está interagindo – não consta, para o grupo, nenhuma interação desde o início da atividade –, o **AgenteÉtico** deverá enviar uma mensagem a todos os integrantes do grupo com o intuito de promover a participação na atividade. Este comportamento atende à regra 1 da Tabela 1, bem como sua ordem de prioridade na busca pela participação e consequente engajamento dos alunos.

- Caso seja identificado que em determinado grupo, um ou mais estudantes não estejam interagindo – não consta, para eles, nenhuma interação desde o início da atividade –, o **AgenteÉtico** deverá enviar mensagens individuais para este(s) estudante(s) com o objetivo de promover a sua participação. Este comportamento atende à regra 2 da Tabela 1, bem como sua ordem de prioridade.

- Caso seja identificado que em determinado grupo todos os estudantes estejam interagindo, mas um ou alguns deles não tenha realizado uma quantidade de interações considerada equitativa em relação às outras, o **AgenteÉtico** deverá enviar mensagens individuais para este(s) estudante(s) com o objetivo de promover a equidade na quantidade de interações. Este comportamento atende à regra 3 da Tabela 1, bem como a sua ordem de prioridade. A quantidade mínima a ser considerada equitativa será parametrizada pelo professor, atendendo aos requisitos não funcionais relacionados à supervisão humana e transparência.

- Caso seja identificado que em determinado grupo todos os estudantes estejam interagindo de forma equitativa, mas que um ou mais deles realizou sua última interação há um tempo considerado maior que o admissível em relação ao tempo atual da atividade, o agente deverá enviar uma mensagem individual a este(s) estudante(s) com o objetivo de encorajar

sua participação no decorrer da atividade. Este comportamento atende à regra 4 da Tabela 1, bem como sua respectiva prioridade.

- Por fim, o **AgenteÉtico** deverá salvar em um arquivo *.txt todas as decisões tomadas, bem como os dados usados para tomá-las. Este comportamento atende aos princípios de explicabilidade e transparência para IA.

Uma vez descritos os comportamentos mais básicos do AgenteÉtico e como eles atendem às regras estabelecidas na Tabela 1 e aos princípios éticos para IA, sobretudo àqueles propostos pela UNESCO (2020), é imprescindível citar que tais comportamentos não são automáticos. Devido à existência da regra 8 constante na Tabela 1, o agente não deverá intervir nas decisões dos alunos. Tal regra está em acordo com o princípio do respeito à autonomia, e, apesar de aparecer na tabela como tendo prioridade 1, deve ser observada em todas as decisões do agente, conflitando, por tanto, com qualquer objetivo do **AgenteÉtico** que vise interferir nas decisões dos alunos, mesmo àquelas que impliquem em não participar da atividade por meio de interações no fórum.

Estas características levam o agente a um dilema ético a cada vez que identifica algum dos cenários que exijam sua intervenção por meio de mensagens. Com isso, visando resolver tais dilemas por meio do atendimento ao bem comum e, trazendo maior prazer líquido ao maior número de pessoas possível, o **AgenteÉtico** fará uso de seu raciocínio ético utilitarista, conforme descrito no capítulo 3.1.2.2.

3.4 PROVA DE CONCEITO

Com o objetivo de avaliar a efetividade da solução proposta, foi implementado o agente Tutor do Ethoscool, pois este é o AMA que compõe o SMA descrito nos capítulos anteriores. Nesse sentido, sendo a ética por design o foco da presente pesquisa, a construção deste componente é considerada suficiente para avaliar o comportamento ético demonstrado pelo agente.

Sendo assim, para a codificação do agente Tutor, foi utilizado o JADEX, um framework que, entre outras características, aplica o paradigma orientado a objetos para a criação de agentes orientados a objetivos. Além disso, o Jadex fornece um ambiente de execução e uma Interface de Programação de Aplicativos (API[5]) para o desenvolvimento de agentes BDI (DELJOO *et al.*,

[5] Neste trabalho usamos a sigla do termo em inglês: Application Programming Interface.

2018).

A razão para a escolha desta plataforma foi a sua ampla utilização, acesso aberto e robustez (CARDOSO; FERRANDO, 2021). Além disso, o JADEX dispõe de vários recursos, como uma infraestrutura de tempo de execução para agentes, vários estilos de interação, suporte de simulação, formação de rede de sobreposição automática e um amplo conjunto de ferramentas de tempo de execução (POKAHR *et al.*, 2005). Os próximos capítulos descrevem a implementação do agente Tutor.

3.4.1 Implementação do Agente Tutor

Conforme mencionado no capítulo anterior, a implementação de agentes BDI utilizando o JADEX se dá por meio da aplicação do paradigma orientado a objetos. Com isso, neste capítulo serão apresentados os trechos mais relevantes do código implementado durante a construção do agente Tutor, usando a linguagem de programação JAVA.

Sendo assim, na Figura 9 é possível observar a assinatura da classe **AgenteTutor**, do pacote ethoscool. Também é notável o uso de *Annotations* do JADEX para a identificação de elementos da programação de agentes. A *Annotation @Agent*, por exemplo, indica que a classe **AgenteTutor** será um agente, sendo que, por meio dos parâmetros informados, evidencia-se que tal agente será baseado na arquitetura BDI.

Figura 9 – Assinatura da Classe AgenteTutor.

AgenteTutor.java

```
1  package ethoschool;
2
3  import java.io.BufferedReader;
26
27 @Agent(type=BDIAgentFactory.TYPE)
28 public class AgenteTutor {
29     private double      reqPercInteracoes; //Percentual de interações mínimo para ser considerado eqüitativo
30     private double      reqTempoMaxSemInteracoes; //Tempo máximo (minutos) tolerável do aluno sem interação
31     private Date        inicioAtividade; //Data Hora do início da atividade
32     private Date        fimAtividade; //Data hora do fim da atividade
33     private String[][]  decisaoUtilitarista; //Matriz que recebe o resultado da função HAU.
34
35     /*************************************************************************************************************
36      *
37      * CRENÇAS DO AGENTE BDI @Belief Annotation
38      *
39      *************************************************************************************************************/
40     @Belief
41     protected boolean   dadosInteracoesDisponivel; //Inicia acreditando que não há dados a analisar
42
43     @Belief
44     protected boolean   estudantesInteragindoAdeq = true; //Inicia acreditando que a interação dos estudantes é adequada
```

Fonte: Código fonte do AgenteTutor.

Também se faz importante salientar, a presença dos atributos que permitirão a inclusão de importantes parâmetros com os quais o Tutor precisará lidar ao perseguir seu objetivo. Nesse contexto, além dos atributos **inicioAtividade** e **fimAtividade**, que caracterizam, como o nome

sugere, a data e hora do início e fim da atividade a ser executada pelos alunos, respectivamente, também é possível notar os atributos **reqTempoMaxSemInteracoes** e **reqPercInteracoes**.

O primeiro materializa a visão do professor sobre o que pode ser considerado o máximo tolerável no que diz respeito ao tempo que um aluno pode ficar sem interagir com o grupo por meio do fórum. O segundo, indica, também a partir da perspectiva do professor, qual é o percentual mínimo de interações, que um aluno deve realizar para ser considerado equitativo, em relação ao aluno do mesmo grupo que mais interagiu.

Por meio destes dois últimos atributos é que o agente Tutor terá condições de atender às regras 3 e 4 da Tabela 1. Podendo identificar e determinar se um ou mais alunos estão há muito tempo sem interagir ou se as interações no fórum estão sendo realizadas de forma equitativa.

Estas parametrizações permitirão ao professor inserir suas concepções e entendimentos com relação à aprendizagem colaborativa, tornando o agente mais previsível, pois é com base nesses parâmetros que o agente Tutor irá se orientar e orientar os alunos. Por fim, o atributo **decisaoUtilitarista** servirá para armazenar a decisão do agente sobre se deve intervir ou não nas dinâmicas dos alunos ou grupos, por meio de mensagens.

Com relação às crenças do agente Tutor, também mostradas na Figura 9, pode-se observar a partir da *Annotation @Belief* que estão presentes as crenças **dadosInteracoesDisponivel** e **estudantesInteragindoAdeq**. A primeira faz com que o agente acredite que os dados das interações estão disponíveis para análise. Tais dados, no modelo original, seriam enviados pelo agente Monitor, entretanto, nos presentes testes serão carregados de um arquivo *.txt. Com isso, esta crença será alterada na iniciação do agente, para garantir que os dados sejam analisados.

A análise dos dados, por sua vez, se dá por meio do objetivo (*goal*) do agente, que se materializa pelo nome de **ChecarNecessidadeIntervencao**. A Figura 10 apresenta este objetivo e o gatilho que o aciona.

Figura 10 – Assinatura do Objetivo ChecarNecessidadeIntervenção.

```
*AgenteTutor.java ×
56
57    /*************************************************************************************************************************
58     *
59     * OBJETIVO DO AGENTE BDI
60     * Este objetivo checka a necessidade de intervenção a cada vez que a crença dadosInteraçõesDisponíveis é alterada
61     * para true.
62     *************************************************************************************************************************/
63    @Goal
64    public class ChecarNecessidaIntervencao {
65
66        @GoalCreationCondition(beliefs="dadosInteracoesDisponivel") //Ativa o objetivo ao alterar a crença dadosInteracoesDisponivel
67        public ChecarNecessidaIntervencao(){
68
69        }
70    }
```

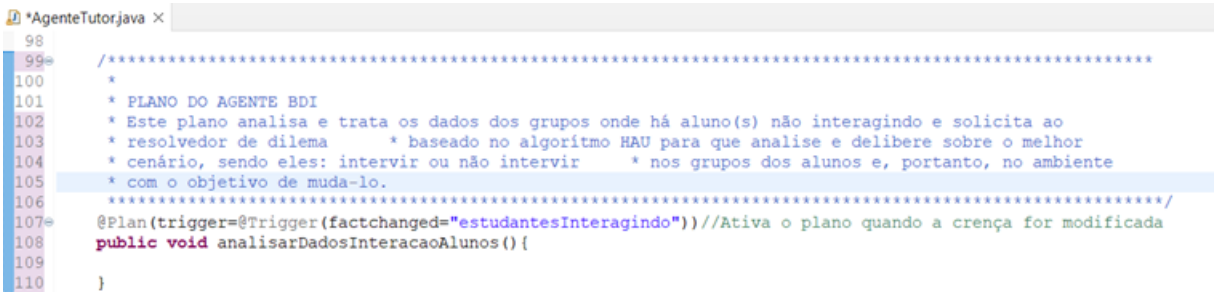
Fonte: Código fonte do AgenteTutor.

Por meio deste objetivo é que o agente irá identificar ou não a presença de aluno(s) ou

grupo(s) que não estejam interagindo de forma adequada. No âmbito da presente pesquisa, a forma adequada se dá pelo atendimento às regras previstas na Tabela 1.

Sendo assim, toda vez que forem identificados casos nos quais as interações sejam classificadas como inadequadas, o agente passará a acreditar que os alunos não estão interagindo adequadamente, mudando o valor da sua crença **estudantesInteragindoAdeq** para false. Esta condição ativa o plano **analisarDadosInteracaoAlunos**, que executará os procedimentos, inclusive a resolução de dilemas éticos, se for o caso, para deliberar sobre a viabilidade da intervenção. A Figura 11 mostra a assinatura do referido plano.

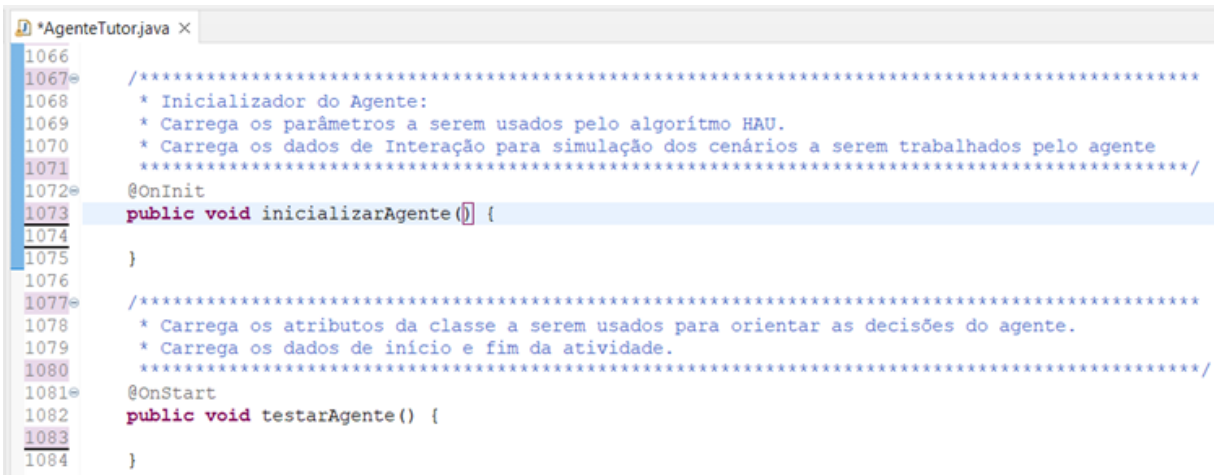
Figura 11 – Assinatura do Plano AnalisarDadosInteracaoAlunos.



Fonte: Código fonte do AgenteTutor.

Dito isso, uma vez que a implementação do agente Tutor aqui descrito foi elaborada com o fim de avaliar a viabilidade do modelo, os demais parâmetros para a sua acurada execução se darão por meio das suas funções de iniciação. Estas funções se caracterizam pelas *Annotations* @*OnInit*, que deve marcar métodos a serem executados durante a criação do agente na memória e @*OnStart*, que deve marcar métodos a serem executados durante a iniciação do agente, logo após métodos marcados com @*OnInit*. A Figura 12 apresenta a assinatura destes métodos.

Figura 12 – Assinatura dos Métodos de Inicialização do Agente.



Fonte: Código fonte do AgenteTutor.

Por fim, a Figura 13 mostra a função **resolverDilema**(), que implementa o algoritmo HAU, responsável por resolver os dilemas éticos com os quais o agente Tutor venha a se deparar. Assim, por ser uma função chave no código do agente, será mostrada toda a codificação da função. Como é possível observar, a função recebe como parâmetros:

- Uma matriz do tipo String denominada dadosParaDecisao, contendo os dados usados pelo agente para tomada de decisão. Tais dados deverão ser retornados junto com a decisão do algoritmo utilitarista sobre realizar ou não a intervenção, para ser impressa, posteriormente, em um arquivo *.txt;
- Um ArrayList contendo um cenário possível, considerando que o agente deliberou pela não intervenção;
- Um ArrayList contendo um cenário possível, considerando que o agente deliberou pela intervenção.

Figura 13 – Função que Implementa o Algoritmo HAU.

*AgenteTutor.java ×

```
1175    public static String[][] resolverDilema(String[][]dadosParaDecisao, ArrayList<DadosDilema> cenarioIntervencao,
1176                                            ArrayList<DadosDilema> cenarioNaoIntervencao) {
1177        double pLTIntervencao = 0, pLTNaoIntervencao = 0; //Prazer Líquido Total
1178        String[][] decisao = new String[1][8];
1179        //Cenário no qual o agente intervem
1180        for(DadosDilema dado : (ArrayList<DadosDilema>)cenarioIntervencao) {
1181            pLTIntervencao = pLTIntervencao + dado.getIntensidadePrazer()*dado.getDuracaoPrazer()*dado.getProbabilidadeOcorrencia();
1182        }
1183        //Cenário no qual o agente NÃO intervem
1184        for(DadosDilema dado : (ArrayList<DadosDilema>)cenarioNaoIntervencao) {
1185            pLTNaoIntervencao = pLTNaoIntervencao + dado.getIntensidadePrazer()*dado.getDuracaoPrazer()*dado.getProbabilidadeOcorrencia();
1186        }
1187        if(pLTIntervencao > pLTNaoIntervencao) {
1188            decisao[0][0] = "true"; //Decisão do agente pela Intervenção
1189            decisao[0][1] = String.valueOf(pLTIntervencao); //PLT considerando a Intervenção
1190            decisao[0][2] = String.valueOf(pLTNaoIntervencao); //PLT considerando a não intervenção
1191            decisao[0][3] = dadosParaDecisao[0][0]; //Aluno Analisado
1192            decisao[0][4] = dadosParaDecisao[0][1]; //Intensidade do Prazer Individiual
1193            decisao[0][5] = dadosParaDecisao[0][2]; //Probabilidade do Prazer Individual
1194            decisao[0][6] = dadosParaDecisao[0][3]; //Duração do Prazer Individual
1195            decisao[0][7] = dadosParaDecisao[0][4]; //Percentual do Tempo Decorrido
1196        }else if(pLTIntervencao == pLTNaoIntervencao) {
1197            decisao[0][0] = "false"; //Decisão do agente pela NÃO intervenção
1198            decisao[0][1] = String.valueOf(pLTIntervencao); //PLT considerando a Intervenção
1199            decisao[0][2] = String.valueOf(pLTNaoIntervencao); //PLT considerando a não intervenção
1200            decisao[0][3] = dadosParaDecisao[0][0]; //Aluno Analisado
1201            decisao[0][4] = dadosParaDecisao[0][1]; //Intensidade do Prazer Individiual
1202            decisao[0][5] = dadosParaDecisao[0][2]; //Probabilidade do Prazer Individual
1203            decisao[0][6] = dadosParaDecisao[0][3]; //Duração do Prazer Individual
1204            decisao[0][7] = dadosParaDecisao[0][4]; //Percentual do Tempo Decorrido
1205        }else {
1206            decisao[0][0] = "false"; //Decisão do agente pela NÃO intervenção
1207            decisao[0][1] = String.valueOf(pLTIntervencao); //PLT considerando a Intervenção
1208            decisao[0][2] = String.valueOf(pLTNaoIntervencao); //PLT considerando a não intervenção
1209            decisao[0][3] = dadosParaDecisao[0][0]; //Aluno Analisado
1210            decisao[0][4] = dadosParaDecisao[0][1]; //Intensidade do Prazer Individiual
1211            decisao[0][5] = dadosParaDecisao[0][2]; //Probabilidade do Prazer Individual
1212            decisao[0][6] = dadosParaDecisao[0][3]; //Duração do Prazer Individual
1213            decisao[0][7] = dadosParaDecisao[0][4]; //Percentual do Tempo Decorrido
1214        }
1215        return decisao;
1216    }
1217 }
```

Fonte: Código fonte do AgenteTutor.

Com base nestes parâmetros, a função **resolverDilema**() irá avaliar os dois cenários e verificar qual deles apresenta um prazer líquido total maior. A partir disso, será tomada a decisão por realizar ou não a intervenção. O retorno da função será uma matriz contendo a decisão e os

dados usados para tomá-la. O próximo capítulo descreverá com mais detalhes os arquivos usados como parâmetros, os cenários usados para testar o agente Tutor e os resultados obtidos a partir dos testes.

3.4.2 Parametrização do Agente Tutor

Para que haja uma maior compreensão acerca dos resultados gerados pelo agente durante a sua execução, considera-se imprescindível, primeiramente, a apresentação detalhada dos seus parâmetros de entrada. Isto possibilitará avaliar com maior clareza as decisões tomadas, bem como estar ciente da importância das variáveis envolvidas.

Sendo assim, em primeiro lugar deverão ser parametrizados os dados relativos à atividade. Nos exemplos a seguir serão usadas como cenário as datas e horários de início e fim da atividade, conforme apresentado na Tabela 2.

Tabela 2 – Parâmetros de Início e Fim da Atividade Usado nos Testes.

Data Início	Hora Início	Data Fim	Hora Fim
20/06/2023	08:00:00	22/06/2023	08:00:00

Fonte: Produzido pelo autor

Nos testes a serem realizados, pretende-se simular os cenários apresentados no capítulo 3.1.2.3. Por esta razão, o tempo total da atividade será de quarenta e oito horas. Isto permitirá avaliar como o agente se comporta naqueles cenários.

Outros dois atributos determinantes para as decisões do agente Tutor são os atributos **reqPercInteracoes** e **reqTempoMaxSemInteracoes**. Tais atributos, conforme descrito no 3.4.1, informam ao agente o percentual mínimo requerido de interações que um aluno deve realizar, em relação ao aluno que mais interagiu, para ser considerado equitativo e o tempo máximo que deve ser tolerado de um aluno sem interagir no fórum. A Tabela 3 apresenta os valores de ambos os atributos para os testes a serem realizados.

Tabela 3 – Valores dos Atributos para Uso do Agente Tutor.

Percentual Mínimo de Interações Requerido	Tempo Máximo Sem Interações Tolerado
70%	1080 (minutos)

Fonte: Produzido pelo autor

Como é possível notar, o percentual mínimo de interações requerido para que um aluno seja considerado equitativo será de setenta por cento, em relação ao aluno que mais interagiu. Não se pretende afirmar, neste trabalho, que este é um valor ótimo, pois isto pode variar de

acordo com o grupo de trabalho ou entendimento do professor. Entretanto, para efeito dos testes aqui propostos, entende-se que setenta por cento poderá garantir uma boa participação geral do grupo.

O tempo máximo sem interações tolerado, por sua vez, será de dezoito horas. Isto porque, considerando que o tempo total da atividade é de quarenta e oito horas, julga-se, para fins de testagem, que tolerar que os estudantes permaneçam sem interagir por mais tempo, implicaria em quase cinquenta porcento do tempo da atividade sem interações, o que poderia inviabilizar a aprendizagem colaborativa, pela falta de colaboração.

Com relação aos parâmetros a serem processados pelo algoritmo utilitarista, há uma complexidade maior em sua definição. Não por causa dos parâmetros em si, que são apenas quatro: o número de pessoas afetadas e para cada uma delas: a intensidade de prazer/desprazer; a duração do prazer/desprazer; e a probabilidade de ocorrência desse prazer/desprazer para cada ação possível. Mas sim, porque no ambiente de aprendizagem colaborativa, consideramos que os valores destes parâmetros mudam na medida em que o final da atividade se aproxima. Além disso, cada tipo de abordagem que o agente possa fazer, pode gerar um nível de prazer/desprazer diferente.

Por exemplo, realizar uma intervenção para lembrar o aluno acerca da atividade, quando ele já está há muito tempo sem interagir com o grupo, pode ser percebido como positivo se o tempo da atividade já estiver perto do fim e o aluno houver esquecido de concluir alguma tarefa. Por outro lado, a mesma natureza de intervenção nas primeiras horas de atividade pode ser percebida como desnecessária e inconveniente. A mesma situação pode se dar com relação aos outros tipos de intervenção.

Por outro lado, mesmo que um determinado aluno sinta um grande desprazer com a intervenção do agente Tutor, o grupo como um todo pode se beneficiar, experimentando prazer com os resultados. Isto pode ocorrer, por exemplo, em uma situação na qual um aluno que não estava participando seja levado a fazê-lo, e com isso, contribuir positivamente para o desenvolvimento da atividade.

É claro que estas são situações hipotéticas, mas o agente precisa estar preparado para o máximo de estados possíveis. Por esta razão, a experiência do professor ao parametrizar o sistema pode ser determinante. Além disso, também há o entendimento de que o algoritmo leva em conta um prazer ou desprazer suposto pelo professor. No entanto, esta inclusão de viés é que torna o agente adaptável a diferentes contextos e permite que ele permaneça alinhado aos valores

neles vigentes, da mesma forma que a ética também pode variar de acordo com o tempo, com o espaço e com o contexto.

Com isso, foi elaborada uma planilha que prevê esta variabilidade nos parâmetros, tornando-os adaptáveis quanto ao tipo de intervenção, tempo decorrido da atividade e se ela se aplica ao grupo ou individualmente. A Tabela 4 apresenta uma amostra destas possibilidades aplicadas ao cenário onde um ou mais alunos não estão interagindo de forma equitativa.

Dessa forma, ao observar a Tabela 4, nota-se a presença de sete colunas. As quatro primeiras, da esquerda para a direita, determinam um possível cenário, enquanto as demais determinam os parâmetros a serem usados pelo algoritmo HAU. Nesse sentido, o agente deverá identificar o cenário ao qual uma eventual intervenção se faça necessária, antes de solicitar ao algoritmo utilitarista que tome a decisão final sobre qual ação executar. Sendo assim, a seguir serão descritos maiores detalhes sobre estes cenários e parâmetros.

Tabela 4 – Amostra de Parâmetros Possíveis para Decisão Utilitarista.

Tipo Intervenção	Cenário Intervenção	Escopo	Tempo Decorrido	Prazer	Probabilidade	Duração
ANE	*true*	G	0	0	0.5	0
ANE	*true*	G	0.25	1	0.5	1
ANE	*true*	G	0.5	1	0.7	2
ANE	*true*	G	0.75	1	0.8	2
ANE	*true*	G	0.9	2	0.95	2
ANE	*false*	G	0	0	0.5	0
ANE	*false*	G	0.25	0	0.5	0
ANE	*false*	G	0.5	0	0.7	0
ANE	*false*	G	0.75	-1	0.8	1
ANE	*false*	G	0.9	-1	0.9	2
ANE	*true*	A	0	-2	0.6	2
ANE	*true*	A	0.25	-1	0.6	2
ANE	*true*	A	0.5	1	0.5	2
ANE	*true*	A	0.75	1	0.6	2
ANE	*true*	A	0.9	-1	0.7	2
ANE	*false*	A	0	0	0.5	0
ANE	*false*	A	0.25	0	0.5	0
ANE	*false*	A	0.5	0	0.5	0
ANE	*false*	A	0.75	-1	0.6	1
ANE	*false*	A	0.9	-1	0.9	1

Fonte: Produzido pelo autor

As colunas constantes na Tabela 4, da esquerda para a direita trazem as seguintes informações:

- Tipo Intervenção: conforme o modelo projetado, o agente divide os tipos de intervenção possível em cinco: AIN, que identifica os alunos que estão interagindo de acordo com as regras; ANE, que identifica os alunos não equitativos, com relação ao número de intervenções; ANI, que identifica os alunos que não estão interagindo; ANT, que identifica

os alunos que estão a um tempo maior que o tolerado sem interagir; e GNI, que identifica os grupos nos quais nenhum aluno está interagindo.

- Cenário: os cenários, neste contexto, são apenas dois: ou o agente intervém, com valor *true*; ou o agente não intervém, com valor *false*.

- Escopo: esta coluna mostra a quem se aplicam os parâmetros, sendo G para o grupo todo ou A, para um aluno individualmente. Isto é importante porque, algumas intervenções podem ser menos prazerosas do ponto de vista individual e mais prazerosas do ponto de vista do grupo, ou vice e versa. E estas características devem ser levadas em conta pelo algoritmo HAU.

- Tempo Decorrido: o valor dos parâmetros pode variar também de acordo com o tempo decorrido da atividade. Algumas intervenções, por exemplo, podem gerar mais prazer ou desprazer, para o indivíduo ou para o grupo, conforme o tempo da atividade se aproxime do fim. Os valores que podem ser assumidos por este campo se dão em percentuais do tempo decorrido, podendo ser: 0, para um tempo menor do que vinte e cinco porcento do total da atividade; 0,25 para um tempo que varia de vinte e cinco porcento até quarenta e nove por cento; 0,5 para valores de tempo entre cinquenta por cento e setenta e quatro porcento; 0,75 para valores entre setenta e cinco por cento e oitenta e nove por cento; e 0,9 para valores que variam de noventa a cem porcento.

- Prazer: esta coluna, assim como as próximas duas, será usada diretamente pelo algoritmo HAU para tomada de decisão. Ela indica quanto prazer ou desprazer o aluno pode experimentar no cenário avaliado pelo algoritmo. Seus valores variam de -2, para muito desprazer, até 2 para muito prazer.

- Probabilidade: em ambientes de ensino e aprendizagem é preciso lidar com variáveis nem sempre controláveis. As emoções humanas estão entre elas. Por esta razão é que não se pode determinar, a priori, quanto prazer um indivíduo irá experimentar. Nesse sentido, o algoritmo trabalha com a probabilidade de ocorrência do prazer ou desprazer previsto na coluna Prazer. Os valores deste campo podem variar de zero a noventa e nove porcento.

- Duração: o algoritmo também leva em conta o quanto espera-se que este prazer ou desprazer dure. No presente trabalho usou-se uma parametrização por horas. Os valores desta coluna, para fins de teste variam de uma hora até quatro horas.

Conforme já mencionado, a Tabela 4 apresenta apenas uma amostra dos parâmetros usados nos testes do agente Tutor. A tabela completa, entretanto, pode ser vista no Anexo A. Para os fins da realização da prova de conceito aqui descrita, optou-se por carregar estes parâmetros na memória para uso do agente, por meio de sua função de iniciação, na qual os dados serão lidos de um arquivo *.txt, juntamente com os dados simulando as interações dos estudantes.

Estes últimos, por sua vez, conforme já mencionado, foram modelados para simular os cenários descritos no capítulo 3.1.2.3. Dessa forma, a Tabela 5 apresenta o modelo de dados a ser importado e tratado pelo agente Tutor. Nesse modelo é possível evidenciar a existência de quatro colunas, sendo: grupo, que deve caracterizar o grupo ao qual o aluno pertence; aluno, que traz a identificação única do aluno no grupo; a quantidade de interações realizada por cada aluno; e, por fim, a data e a hora da última interação realizada por cada aluno.

Tabela 5 – Relação de Dados das Interações dos Alunos.

Grupo	Aluno	Quantidade Interações	Última Interação
1	1	5	2023-09-05 - 16:30:11
1	3	7	2023-09-05 - 13:27:24
1	5	4	2023-09-04 - 23:28:57

Fonte: Produzido pelo autor

Dessa forma, considerando o formato dos dados apresentados na Tabela 5, foram realizados vários testes com o objetivo de avaliar o funcionamento do agente. Para isso, foram elaborados Casos de Teste (CT) para dar mais transparência e previsibilidade ao processo. O próximo capítulo apresenta a execução e os resultados dos testes realizados.

3.4.3 Execução e Avaliação do Agente Tutor

No contexto da Engenharia de Software, CTs podem ser descritos como conjuntos específicos de condições e dados usados para verificar se um determinado componente, sistema ou funcionalidade de software está funcionando corretamente (ISO/IEC/IEEE, 2022). Nesse sentido, o Quadro 6 apresenta o caso de teste para o primeiro cenário descrito no capítulo 3.1.2.3. Em tal cenário, é importante lembrar que a atividade terá quarenta e oito horas de duração e os demais parâmetros se darão conforme a Tabela 3.

Dito isso, ao executar esse caso de teste, o agente deliberou por não realizar qualquer intervenção no grupo de estudantes. Tal intervenção implicaria em enviar uma mensagem para todos os integrantes do grupo, com o intuito de incentivar a participação dos estudantes por meio de interações no fórum.

Sendo assim, após deliberar sobre a necessidade de intervenção, conforme especificado em seus requisitos, o agente salvou sua decisão e os dados usados para tomá-la. Esta característica visa possibilitar auditorias, e dotar o Ethoscool de meios que lhe assegurem maior explicabilidade.

Quadro 6 – Caso de Teste do Primeiro Cenário

Caso de Teste	Grupo Não Interagindo
Descrição	Após 8 horas de atividade o agente evidencia que nenhuma interação foi realizada pelos estudantes de um determinado grupo.
Passos:	1. Parametrizar os dados de entrada, referentes às interações dos estudantes, de modo que não contenham nenhuma interação de nenhum aluno. 2. Executar o Agente Tutor.
Resultados Esperados:	1. O agente deve avaliar a necessidade e a pertinência de realizar a intervenção por meio de mensagens. Os resultados podem ser: a. Positivo: a intervenção foi considerada necessária e pertinente. b. Negativo: a intervenção não foi considerada necessária e pertinente. 2. O agente deve salvar a decisão e os dados usados para tomá-la.

Fonte: Produzido pelo autor.

Nesse sentido, Figura 14 apresenta o relatório criado pelo agente Tutor, contendo a sua decisão e os dados que a embasaram. No relatório é possível observar a data de quando a decisão foi tomada, o grupo e o aluno ao qual a decisão se aplica, se for o caso. Sobre este aspecto, mais detalhes serão dados na análise dos outros cenários.

Figura 14 – Dados Sobre a Decisão do Agente para o Primeiro Cenário.

```
----------------------------------------------------------------
Data: Mon Jul 24 20:05:23 BRT 2023
Dados usados para tomada de decisão do agente.
----------------------------------------------------------------
----------------------------------------------------------------
Analise para o Grupo: 1  -  Aluno: Não se Aplica
Grupo Aluno  Tipo   Q.Interacoes  Tempo U.Interacao  U.Interacao
  1     1    GNI       0              0               null
  1     3    GNI       0              0               null
  1     5    GNI       0              0               null

Não Intervir.
Dados Individuais para Cenario de Não Intervencao:
(%)Tempo Decorrido: 17%
Prazer Líquido Total ao Intervir: 0.0
Prazer Líquido Total ao Não Intervir: 0.0
Decisão do agente: Não Intervir.
----------------------------------------------------------------
```

Fonte: Relatório gerado pelo AgenteTutor.

No relatório também é possível observar, por meio das colunas, da esquerda para a direita: o(s) grupo(s) – pode haver mais de um na mesma análise; o aluno; o tipo de intervenção; a quantidade de interações realizadas pelo aluno, o tempo desde a última interação – em minutos; e a data e hora da última interação. Por fim, o relatório apresenta dados sobre o percentual do tempo de atividade decorrido no momento da análise, o PLT calculado para o cenário de intervenção e não intervenção e a decisão final do agente.

No primeiro caso analisado, o tipo de intervenção se daria no grupo inteiro – GNI. Isto ocorre quando nenhum integrante do grupo está interagindo. Neste caso, a análise individual por aluno não se aplica. Considerando a parametrização realizada, mesmo que seja de suma importância que ao menos um integrante do grupo interaja para atender à primeira regra do agente, constante na Tabela 1, os resultados obtidos estão em conformidade com o esperado.

O próximo cenário, descrito no Quadro 7, analisa o caso de um estudante que não está interagindo com o grupo por meio do fórum. Com exceção daqueles específicos usados pelo algoritmo HAU, os demais parâmetros são os mesmos aplicados para o primeiro cenário.

Quadro 7 – Caso de Teste do Segundo Cenário

Caso de Teste	Aluno Não Interagindo com o Grupo
Descrição	Após 26 horas de atividade o agente evidencia que um dos integrantes de um determinado grupo ainda não realizou nenhuma interação com os colegas.
Passos:	1. Parametrizar os dados de entrada, referentes às interações dos estudantes, de modo que um integrante do grupo não apresente nenhuma interação e os demais estejam equitativos. 2. Executar o Agente Tutor.
Resultados Esperados:	1. O agente deve avaliar a necessidade e a pertinência de realizar a intervenção por meio de mensagens. Os resultados podem ser: a. Positivo: a intervenção foi considerada necessária e pertinente. b. Negativo: a intervenção não foi considerada necessária e pertinente. 2. O agente deve salvar a decisão e os dados usados para tomá-la.

Fonte: Produzido pelo autor.

Nesse cenário, dadas as condições extremas às quais o agente foi exposto, a sua decisão pela intervenção também foi tomada conforme o esperado. A Figura 15 apresenta os dados que serviram de subsídio para tal decisão. Como é possível observar, a análise agora não foi realizada para o grupo, mas para o aluno 5 – linha 7 do relatório. Este, por sua vez, está classificado como aluno não interagindo – ANI –, constando para ele zero interações.

Figura 15 – Dados Sobre a Decisão do Agente para o Segundo Cenário.

```
--------------------------------------------------------------------------------
Data: Mon Jul 24 20:20:35 BRT 2023
Dados usados para tomada de decisão do agente.
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
Analise para o Grupo: 1  -  Aluno: 5
Grupo Aluno  Tipo   Q.Interacoes  Tempo U.Interacao  U.Interacao
  1     1    AIN       15           373              Mon Jul 24 14:07:00 BRT 2023
  1     3    AIN       18           484              Mon Jul 24 12:16:00 BRT 2023
  1     5    ANI       0            0                null

Dados Individuais para Cenario de Intervencao:
(%)Tempo Decorrido: 54%
Prazer Líquido Total ao Intervir: 5.0
Prazer Líquido Total ao Não Intervir: -9.0
Decisão do agente: Intervir.
--------------------------------------------------------------------------------
```

Fonte: Relatório gerado pelo AgenteTutor.

Os demais membros do grupo estão classificados como alunos interagindo normalmente

– AIN. Esta classificação dispensa quaisquer outras análises no sentido de decidir sobre a necessidade ou não de uma intervenção por parte do agente.

Sendo assim, ao prosseguir com os testes, o Quadro 8 traz um CT baseado no cenário 3, no qual um dos alunos do grupo não está interagindo de forma equitativa. Nesse contexto, é importante lembrar que, para ser considerado equitativo, um aluno precisa ter, ao menos, setenta porcento da quantidade de interações do aluno que mais interagiu no grupo.

Quadro 8 – Caso de Teste do Terceiro Cenário

Caso de Teste	Aluno Não Interagindo de Forma Equitativa
Descrição	Após 36 horas de atividade o agente evidencia que um dos integrantes de determinado grupo não está interagindo com os colegas de forma equitativa.
Passos:	1. Parametrizar os dados de entrada, referentes às interações dos estudantes, de modo que um integrante de determinado grupo apresente uma quantidade de interações não equitativa em relação aos outros membros. 2. Executar o Agente Tutor.
Resultados Esperados:	1. O agente deve avaliar a necessidade e a pertinência de realizar a intervenção por meio de mensagens. Os resultados podem ser: a. Positivo: a intervenção foi considerada necessária e pertinente. b. Negativo: a intervenção não foi considerada necessária e pertinente. 2. O agente deve salvar a decisão e os dados usados para tomá-la.

Fonte: Produzido pelo autor.

Os resultados da execução deste cenário, apresentados na Figura 16, também de acordo com o esperado, mostram que o agente deliberou pela intervenção. Como é possível evidenciar pela classificação do aluno – na coluna **Tipo** –, o membro número 5 foi considerado não equitativo – ANE.

Figura 16 – Dados Sobre a Decisão do Agente para o Terceiro Cenário.

```
--------------------------------------------------------------------------------
Data: Mon Jul 24 20:17:24 BRT 2023
Dados usados para tomada de decisão do agente.
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
Analise para o Grupo: 1  -  Aluno: 5
Grupo Aluno  Tipo   Q.Interacoes  Tempo U.Interacao  U.Interacao
  1     1    AIN       15            370              Mon Jul 24 14:07:00 BRT 2023
  1     3    AIN       18            481              Mon Jul 24 12:16:00 BRT 2023
  1     5    ANE       3             241              Mon Jul 24 16:16:00 BRT 2023

Dados Individuais para Cenario de Intervencao:
(%)Tempo Decorrido: 75%
Prazer Líquido Total ao Intervir: 6.19
Prazer Líquido Total ao Não Intervir: -4.5
Decisão do agente: Intervir.
--------------------------------------------------------------------------------
```

Fonte: Relatório gerado pelo AgenteTutor.

Para esta natureza de análise, na qual o agente decide acerca de intervir ou não sobre um aluno considerado não equitativo, fica mais clara a natureza quantitativa dos dados que o agente considera para tomar sua decisão. Nesses casos, considerando a calibração dos parâmetros para o

algoritmo HAU, o que mais pesa na decisão é o tempo decorrido da atividade, sendo que, quanto mais próximo do seu início, maior a probabilidade de o agente deliberar pela não intervenção.

Por fim, o Quadro 9 analisa o quarto cenário, onde um ou mais alunos podem estar há mais tempo que o tolerado, sem interagir no fórum. Nesse contexto, é importante mencionar que, de acordo com os parâmetros, este tempo é de mil e oitenta minutos, ou dezoito horas.

Quadro 9 – Caso de Teste do Quarto Cenário

Caso de Teste	Aluno Há Muito Tempo sem Interagir
Descrição	Após 40 horas de atividade o agente evidencia que um dos integrantes de determinado grupo está há mais tempo do que o tolerado sem interagir.
Passos:	1. Parametrizar os dados de entrada, referentes às interações dos estudantes, de modo que um integrante do grupo tenha realizado a sua última interação há um tempo maior do que o tolerado para que um estudante fique sem interagir. 2. Executar o Agente Tutor.
Resultados Esperados:	1. O agente deve avaliar a necessidade e a pertinência de realizar a intervenção por meio de mensagens. Os resultados podem ser: a. Positivo: a intervenção foi considerada necessária e pertinente. b. Negativo: a intervenção não foi considerada necessária e pertinente. 2. O agente deve salvar a decisão e os dados usados para tomá-la.

Fonte: Produzido pelo autor.

A Figura 17 permite evidenciar que o aluno 5, agora equitativo, realizou sua última interação há 1690 minutos, ou seja, há um tempo maior que o tolerado. Esta condição levou o agente a ter que deliberar sobre a intervenção a ser realizada individualmente para este aluno. A decisão do agente, conforme o esperado, foi pela intervenção.

Figura 17 – Dados Sobre a Decisão do Agente para o Quarto Cenário.

```
--------------------------------------------------------------------------------
Analise para o Grupo: 1  -  Aluno: 5
Grupo Aluno  Tipo   Q.Interacoes  Tempo U.Interacao  U.Interacao
  1     1    AIN       15            379              Mon Jul 24 14:07:00 BRT 2023
  1     3    AIN       18            490              Mon Jul 24 12:16:00 BRT 2023
  1     5    ANT       10            1690             Sun Jul 23 16:16:00 BRT 2023

Dados Individuais para Cenario de Intervencao:
(%)Tempo Decorrido: 83%
Prazer Líquido Total ao Intervir: -1.2
Prazer Líquido Total ao Não Intervir: -5.0
Decisão do agente: Intervir.
--------------------------------------------------------------------------------
```

Fonte: Relatório gerado pelo AgenteTutor.

Os cenários usados para avaliar o funcionamento do agente até este ponto, são baseados em exemplos que aplicam condições extremas, nas quais as decisões tomadas ocorreram conforme o esperado. Entretanto, é necessário enfatizar que em cenários nos quais os alunos estejam interagindo de forma adequada – do ponto de vista do agente Tutor –, o Ethoscool não irá ter acionado nenhum de seus desejos (objetivos), e, deste modo, nenhuma ação será executada por ele. Por esta razão, não faz sentido demonstrar testes sendo realizados em cenários assim.

Por outro lado, a previsibilidade dos cenários de testes apresentados também não permite uma avaliação mais abrangente do comportamento do agente em diferentes condições. Nesse sentido, a Tabela 7 apresenta um conjunto de testes executados para o mesmo cenário, mas em tempos de atividade diferentes. Com isso, espera-se demonstrar o comportamento do agente ao longo do tempo para o mesmo conjunto de estados. Sendo assim, para tornar mais compreensível a análise dos dados apresentados na Tabela 7, a Tabela 6 resume os possíveis tipos de intervenção tratados pelo agente Tutor.

Tabela 6 – Classificações Usadas pelo AgenteTutor.

Tipo	Descrição
GNI	Não foram identificadas interações para nenhum membro do grupo.
ANI	Não foram identificadas interações para o aluno.
ANT	Aluno sem interagir há mais tempo do que o tolerado.
ANE	Aluno com interações insuficientes para ser considerado equitativo.
AIN	Aluno interagindo adequadamente.

Fonte: Produzido pelo autor

Tabela 7 – Resultados da Execução do AgenteTutor em Diferentes Condições.

Grupo	Estudante	Tipo	Quantidade Interações	Tempo Decorrido	PLT Intervenção	PLT Não Intervenção	Decisão
5	NA	GNI	0	90.5%	2.8	-5.6	I
1	5	ANI	0	90.5%	5.0	-9.0	I
2	7	ANT	18	90.5%	2.3	-2.7	I
3	6	ANE	3	90.5%	6.1	-4.5	I
5	NA	GNI	0	70.0%	9.6	0.0	I
1	5	ANI	0	70.0%	4.4	-6.8	I
2	7	ANT	18	70.0%	3.7	-1.1	I
3	6	ANE	3	70.0%	7.6	-2.2	I
5	NA	GNI	0	51.9%	2.2	0.0	I
1	5	ANI	0	51.9%	3.6	-0.1	I
2	7	ANT	18	51.9%	0.7	-1.2	I
3	6	ANE	3	51.9%	3.8	0.0	I
5	NA	GNI	0	28.1%	-1.5	0.0	NI
1	5	ANI	0	28.1%	0.0	0.0	NI
2	7	ANT	18	28.1%	-1.5	0.6	NI
3	6	ANE	3	28.1%	0.0	-0.9	I
5	NA	GNI	0	12.3%	0.0	0.0	NI
1	5	ANI	0	12.3%	1.1	2.0	NI
2	7	ANT	18	12.3%	0.0	0.0	NI
3	6	ANE	3	12.3%	2.3	3.7	NI

Fonte: Produzido pelo autor

Como é possível observar, da esquerda para a direita são mostrados os campos: grupo, que corresponde ao identificador único do grupo de estudantes; aluno, que mostra o identificador único de um aluno – NA indica que a análise não se aplica a um aluno específico, mas ao grupo; tipo, que indica o tipo de intervenção necessária identificada; o número de interações;

o percentual do tempo decorrido da atividade; o PLT calculado em caso de intervenção; o PLT calculado em caso de não intervenção; e a decisão do agente.

A responsabilidade pelos resultados apresentados, conforme recomendado pelo documento publicado pela UNESCO (2020) para ética em sistemas de IA, podem e devem ser atribuídos aos humanos que inseriram os parâmetros usados pelo agente Tutor, tanto para realizar a classificação dos estudantes no que se refere ao tipo de intervenção necessária, quanto para calibrar os pesos e medidas usados pelo algoritmo utilitarista. Uma alteração em tais parâmetros teriam resultado em diferentes decisões relacionadas aos mesmos dados.

O presente capítulo teve a pretensão de apenas descrever a solução proposta e os resultados obtidos durante sua implementação e execução. A seguir serão realizadas discussões mais amplas acerca da ética em IA e da adequação da solução aqui descrita.

4 RESULTADOS E DISCUSSÃO

O presente capítulo apresenta discussões acerca da solução proposta e descrita no capítulo 3. Para isso, pretende-se, primeiramente, discorrer sobre as características que tornam o Ethoscool um agente moral artificial. Posteriormente, serão demonstrados recursos e comportamentos que explicam como o agente Tutor se alinha e atende aos princípios éticos, com ênfase àqueles propostos pela UNESCO (2020), para IA. Na sequência, serão expostas algumas das contribuições deixadas pela presente proposta para a IA na Educação e, por fim, fragilidades e possibilidades futuras para pesquisas na área de ética em IA aplicada ao ensino terão, também, um espaço de discussão.

4.1 PORQUE O ETHOSCOOL PODE SER CONSIDERADO UM AMA

Em concordância com o que foi apresentado no decorrer desta pesquisa, sobretudo considerando as definições e pressupostos mostrados no referencial teórico e os resultados expostos no capítulo 3, pode-se afirmar que a ética se apresenta, em diferentes tempos da história humana, como um tema de grande complexidade. Nesse sentido, dedicar-se à busca por respostas no campo da ética normativa, que tem como maior preocupação a articulação e explicação de princípios fundamentais e diretivas sobre como as pessoas devem agir (SCHROEDER, 2017), implica em levantar importantes e desafiadoras discussões.

Nesse contexto, sendo a busca pela identificação de doutrinas e modelos éticos que possam servir de guia para o comportamento humano, um dos temas mais importantes do pensamento filosófico (VAMPLEW *et al.*, 2018), pode-se inferir que limitar as fronteiras do que é certo ou errado não é uma tarefa trivial. A complexidade das questões relacionadas à ética, nascem principalmente da sua dependência de contexto, cultura e tempo. Além disso, a subjetividade do julgamento, tanto do executor, quanto daquele que observa a ação a ser julgada, são fatores igualmente determinantes e devem ser, ambas, levadas em conta (CóRDOVA *et al.*, 2021).

A presente pesquisa precisa, também, lidar com tais questões, apesar de não fazer parte do seu escopo, nem a proposição de princípios para o julgamento de ações com base em uma lógica intencionalista, nem o estabelecimento de pesos e contrapesos para análises utilitaristas acerca da moral. Isto porque, aplicar valores morais ou raciocínios cujas decisões

possam ser consideradas éticas à modelos computacionais, implica em definir quais estruturas éticas, princípios e formas de raciocínio pré-definidas e estabelecidas por uma comunidade de práticas formada por filósofos e pesquisadores, podem ou devem ser aplicadas ao contexto de investigação. No presente caso, ética em IA para o ensino colaborativo.

Desse modo, tanto o contexto teórico, quanto o contexto empírico dos trabalhos executados conduziram ao desenvolvimento de uma solução na qual está contida um AMA. Isto pode ser evidenciado por meio de uma descrição da solução proposta e de sua analogia com as definições disponíveis na literatura vigente.

Ao refletir sobre a definição trazida por Cervantes *et al.* (2020), por exemplo, tem-se que um agente moral artificial deve ser capaz de se engajar em comportamentos considerados éticos ou, ao menos, de evitar comportamentos considerados não éticos. Tal comportamento pode – mas não necessariamente precisa – ser baseado em teorias ou estruturas éticas existentes.

Sobre esta abordagem, é válido destacar dois aspectos: o primeiro diz respeito ao objetivo do agente, que é buscar o engajamento dos estudantes no estudo colaborativo, por meio de uma participação equitativa; o segundo está relacionado à forma como o agente faz isso, que é seguindo princípios deontológicos, priorizados por uma ética *prima facie* e lidando com dilemas éticos aplicando raciocínio utilitarista.

Nesse contexto, é possível observar comportamentos éticos do agente Tutor quando ele busca promover comportamentos éticos nos estudantes – contribuir com a equipe de forma equitativa, não deixar de participar das discussões etc. –, e pela forma com a qual ele faz isso: buscando o bem-estar humano, ao não intervir na autonomia do aluno, a menos que julgue ser pelo bem comum; sendo transparente e explicável, ao deixar explícitos os seus critérios de decisão para ser consultado por humanos; e previsível, no sentido de que não tomará decisões fora do seu escopo de atuação. O próximo capítulo abordará com mais detalhes estes quesitos.

Além disso, o Ethoscool também pode ser classificado de acordo com a literatura vigente como um AMA explícito, pois implementa a ética de forma explícita em sua estrutura algorítmica (MOOR, 2006). Evidencia-se isto por meio do algoritmo HAU, que implementa a teoria utilitarista de ação hedônica de Jeremy Benthan (ANDERSON; ANDERSON, 2008). Para isto, faz uso de uma abordagem também amplamente conhecida para a construção de AMAs, chamada *top-down* (ALLEN *et al.*, 2005), implementando as estruturas éticas deontológica e utilitarista.

Por fim, um aspecto de fundamental importância do agente Tutor é o seu alinhamento

com os valores humanos. Tal alinhamento se verifica observando suas duas dimensões éticas: a dimensão deontológica se mostra alinhada por meio do comportamento do agente, que obedece a regras baseadas em princípios éticos amplamente difundidos; a dimensão utilitarista, por sua vez, se alinha aos valores humanos por meio dos seus parâmetros, que incluem pesos e medidas para tomada de decisão.

Nesse sentido, as crenças, valores e julgamentos do professor acerca do que seria mais ou menos satisfatório para os seus alunos, serão passados ao AMA que irá refletir tais crenças e valores ao tomar suas decisões. Não se trata de um alinhamento a valores humanos universais, mas esta característica é que torna o agente adaptável a diferentes contextos, tempos e entendimentos sobre o que seria o bem comum em um ambiente de aprendizagem colaborativa.

Dito isso, é possível concluir que as características apresentadas tornam o Ethoscool um SMA composto por um agente, denominado Tutor, dotado de capacidade de raciocínio ético. Tais características permitem classificá-lo como um AMA explícito, que implementa uma abordagem *top-down*. O próximo capítulo irá mostrar como a solução proposta atende aos princípios da ética em IA.

4.2 COMO O ETHOSCOOL ATENDE AOS PRINCÍPIOS DA ÉTICA PARA IA

Conforme exposto no capítulo 2.2.2, o presente trabalho é orientado majoritariamente pelos princípios para ética em IA propostos pela UNESCO (2020). Além disso, o Quadro 1 oferece uma visão onde é possível evidenciar as intersecções entre os princípios propostos por esta entidade e outras propostas de relevância internacional.

Sendo assim, no corrente capítulo serão apresentadas características do Ethoscool que permitem afirmar a sua conformidade com alguns dos princípios propostos pela UNESCO (2020) e também com algumas leis brasileiras. Não será realizada a análise de todos os princípios, porque alguns deles não se aplicam ao contexto do Ethoscool. Entretanto, maior foco será dado àqueles que dizem respeito à "equidade/não-discriminação (*fairness*), responsabilidade/prestação de contas (*accountability*) e transparência (*transparency*), conhecidas como a matriz FAT." (BRASIL, 2021).

Tais princípios, além de serem de grande relevância para a solução apresentada neste trabalho, também são motivos de preocupação nas principais propostas para regulação da IA, como é possível observar no Quadro 1. Além disso, são mencionadas também no primeiro item do tópico Ações Estratégicas, da Estratégia Brasileira de Inteligência Artificial (EBIA),

publicada em abril de 2021 pelo Ministério da Ciência, Tecnologia e Inovação (BRASIL, 2021). Sendo assim, nos próximos tópicos serão analisados alguns princípios e como o Ethoscool é capaz de atendê-los.

4.2.1 Proporcionalidade e Não Causar Danos

As preocupações ligadas a este princípio estão relacionadas, principalmente, aos meios empregados para resolver problemas usando IA. Nesse sentido, é estabelecido que os processos do ciclo de vida da IA não devem exceder o necessário para atingir os objetivos legítimos de projeto, e devem ser adequados ao contexto. Estes cuidados visam evitar riscos ao ser humano, ao meio ambiente e aos ecossistemas (UNESCO, 2020).

O Ethoscool se alinha a este princípio por ter mantido, desde a sua concepção, o cuidado necessário para a adequada definição do escopo de ação do agente, que não inclui decisões críticas, como aprovação ou reprovação de alunos, por exemplo. Ademais, ao impedir que o agente Tutor possa aprender comportamento ético, suprime-se a possibilidade de discriminação ou cometimento de erros no que diz respeito à ética, por parte do agente.

Por fim, as regras que regem o comportamento do agente preveem que ele não possa causar, aos alunos, danos conhecidos, como: expor decisões individuais dos estudantes ao grupo; agir de forma desproporcional ou discriminatória; e, somente intervir em sua autonomia se for pelo bem comum do grupo. Além disso, por ter sido construído seguindo a arquitetura BDI, o agente Tutor é orientado por desejos (objetivos). Estes objetivos são claros e bem definidos, assim como os planos e ações necessários para atingi-los.

Estas características conferem ao agente comportamentos mais previsíveis, impedindo que decisões fora do seu escopo de atuação sejam tomadas. Este uso contextualizado, estrito e bem definido da IA, com a aplicação de uma tecnologia adequada, contribuem para que o resultado final esteja alinhado ao princípio da proporcionalidade e não causar danos.

4.2.2 Justiça e Não Discriminação

Este princípio visa evitar o reforço e a perpetuação de preconceitos sociotécnicos inadequados, bem como a discriminação algorítmica injusta (UNESCO, 2020). Trata-se, dessa forma, de uma das questões mais discutidas no que diz respeito à ética em IA (CASAS-ROMA *et al.*, 2021).

Nesse sentido, por meio das regras que orientam o seu comportamento – conforme a Tabela 1 –, o Ethoscool busca agir com isonomia ao tomar decisões que afetam as pessoas com as quais interage. Além disso, conforme mostrado no capítulo 3, nenhum dado pessoal dos alunos é usado pelo agente, tornando nula qualquer possibilidade de que uma decisão seja tomada tendo por base o sexo, gênero, faixa etária, renda, cor ou crença das pessoas envolvidas.

Estas características, somadas ao modo como o algoritmo HAU toma suas decisões, baseado no maior prazer líquido e considerando igualmente todos os membros do grupo, fazem com que as decisões tomadas estejam isentas de preconceitos discriminatórios. Os vieses humanos incluídos pelo professor por meio da parametrização do SMA, não afetam a individualidade e a dignidade humana.

4.2.3 Privacidade

A privacidade é um direito essencial para a proteção da dignidade, da autonomia e da agência humana. Por esta razão, deve ser respeitado, protegido e promovido tanto no nível pessoal, quanto no coletivo (UNESCO, 2020).

O Ethoscool, ao não processar, nem armazenar dados pessoais dos estudantes, não está sujeito ao risco de cometer violações de privacidade. Sendo assim, além de atender ao princípio da privacidade, também está isento de qualquer preocupação com relação à LGPD.

Aos sistemas de fórum com os quais o SMA poderá ser integrado, caberá o atendimento a estes requisitos, principalmente no que diz respeito aos aspectos legais. Isto dependerá, em suma, dos dados coletados dos alunos no momento do cadastro e de como estes sistemas lidarão com estes e outros dados gerados pelos estudantes no decorrer das atividades. Contudo, ainda nestes casos, o Ethoscool não irá fazer uso de qualquer dado pessoal.

4.2.4 Supervisão Humana e Determinação

Este princípio orienta no sentido de que todas as responsabilidades éticas e legais relacionadas a qualquer estágio do ciclo de vida dos sistemas de IA possam ser atribuídas a pessoas naturais ou jurídicas existentes (UNESCO, 2020). Por isso, desde a sua concepção, o presente projeto teve a preocupação de documentar todas as etapas do seu desenvolvimento, como é possível observar no capítulo 3. Desse modo, as responsabilidades legais e éticas pelo bom funcionamento do Ethoscool podem ser imputadas ao pesquisador responsável pelo projeto.

Entretanto, no contexto de execução e uso do SMA, considerando o conjunto de parametrizações e atributos que determinarão as decisões chave do agente Tutor, tais responsabilidades, desde que não estejam relacionadas a erros ou falhas de software, poderão ser atribuídas aos responsáveis pela definição e inclusão de tais parâmetros no sistema, conforme mostrado no capítulo 3.4.2. Estas características possibilitam sempre responsabilizar um agente humano pelas decisões tomadas pelo sistema.

Além disso, não faz parte do escopo de decisões do Ethoscool, qualquer deliberação que implique em risco à vida, à dignidade ou ao bem-estar de seres humanos ou animais. Da mesma forma, as decisões do Ethoscool não apresentam perigos ao meio ambiente ou ecossistemas. Sendo assim, conforme orienta a UNESCO (2020) a delimitação explícita do contexto de atuação do SMA proposto, provê maior segurança ao ceder a ele a responsabilidade pelas decisões de que é encarregado, dispensando, com isso, a necessidade de supervisão humana explícita. Estes fatores permitem afirmar que o Ethoscool atende aos princípios da supervisão humana e determinação.

4.2.5 Transparência e Explicabilidade

Sendo uma pré-condição para garantir que os direitos humanos fundamentais e os princípios éticos sejam respeitados, protegidos e promovidos, transparência e explicabilidade dizem respeito ao direito que as pessoas têm de saber quando uma decisão está sendo tomada por algoritmos de IA e poder exigir ou solicitar informações explicativas sobre tais decisões (UNESCO, 2020). Nesse sentido, o propósito de um Sistema de Inteligência Artificial Explicável (XAI[1]) é tornar o seu comportamento mais inteligível para os seres humanos, fornecendo explicações sobre as suas decisões, previsões ou ações, bem como revelar as informações nas quais está baseando os seus resultados (GUNNING; AHA, 2019).

Dessa forma, a transparência desempenha um papel crucial no fortalecimento da confiança dos indivíduos nos sistemas de Inteligência Artificial. Simultaneamente, a explicabilidade torna esses sistemas compreensíveis, permitindo que forneçam informações sobre seus resultados (UNESCO, 2020). No entanto, cada explicação é inserida em um contexto que depende da tarefa e das habilidades e expectativas do usuário do sistema de IA. As definições de interpretabilidade e explicabilidade são, portanto, dependentes do domínio e não podem ser definidas independentemente deste (GUNNING; AHA, 2019).

[1] Da sigla em inglês: *Explainable Artificial Intelligence*.

Sendo assim, o Ethoscool, em seu domínio de atuação, é capaz de fornecer explicabilidade explícita ao professor, que será o operador do sistema, por meio do seu relatório, que contém as decisões tomadas e os dados usados para tomá-la. Da mesma forma, o uso de estruturas éticas e princípios conhecidos e já validados por uma comunidade de práticas, bem como a aplicação de raciocínio capaz de solucionar dilemas éticos, cujos pesos e medidas são definidos pelo próprio professor, contribuem para tornar o sistema mais inteligível, previsível e explicável (CóRDOVA; VICARI, 2022). Além disso, o modelo BDI – usado como base para o agente Tutor –, sendo orientado por objetivos, torna mais intuitivo para o usuário final, compreender por que um agente tomou determinada decisão (CARDOSO; FERRANDO, 2021).

Esta intuitividade, contribui também para atender ao princípio da transparência do SMA, que não tem processos de decisão que não podem ser explicados, seja pelos parâmetros definidos pelo professor, seja pelos princípios que o guiam ou por meio do seu relatório de decisões. Por fim, neste mesmo relatório deverá constar a informação de que a decisão do Ethoscool foi tomada por um agente deliberativo, deixando claro ao operador que tais decisões são tomadas usando IA.

Com relação aos estudantes, o Ethoscool atende ao princípio da transparência e explicabilidade, permitindo que, a qualquer tempo, possa ser solicitado ao professor o relatório das decisões que dizem respeito ao aluno solicitante. Dessa forma, qualquer decisão tomada pode tornar-se auditável e explicável a qualquer pessoa envolvida nas decisões do sistema, contribuindo, desse modo, para aumentar a confiança dos humanos no Ethoscool. Para concluir, é importante destacar que, embora menos exploradas, a XAI e a IA responsável que, diferente das abordagens tradicionais, se preocuparam em construir uma IA ética e em que possamos confiar, demonstram que nunca devemos falar apenas de IA, mas sempre dos princípios, das metodologias e das técnicas que suportam cada um dos seus subdomínios.

4.2.6 Responsabilidade e Prestação de Contas

Bastante similar ao princípio da supervisão humana e determinação, a responsabilidade e prestação de contas defende que os atores de IA devem respeitar, proteger e promover os direitos humanos e a proteção do meio ambiente e dos ecossistemas, além de assumir a responsabilidade ética e legal de acordo com a legislação nacional e internacional vigente, em particular a legislação internacional de direitos humanos (UNESCO, 2020). Nesse caso, entretanto, há uma preocupação maior sobre a responsabilização dos atores de IA, enquanto no caso da supervisão humana e determinação, a preocupação é mais evidente no que se refere à delegação de responsabilidades

a sistemas de IA, que sempre deve ser supervisionada e autorizada pelos atores de IA.

Desse modo, conforme descrito no capítulo 4.2.4, todas as responsabilidades pelos resultados gerados pelo Ethoscool podem ser atribuídas ao pesquisador e desenvolvedor da solução ou aos operadores do sistema. A atribuição da responsabilidade dependerá unicamente do tipo de resultado, sendo: erros ou falhas de software, de responsabilidade do pesquisador desenvolvedor; e comportamentos e decisões do sistema baseados nos parâmetros fornecidos, de responsabilidade dos operadores.

4.3 CONTRIBUIÇÕES PARA IA NA EDUCAÇÃO

A presente pesquisa traz como solução a proposta um modelo de AMA capaz de suportar processos de aprendizagem colaborativa. Tal proposição se evidencia importante pela necessidade de investigar soluções para ética em IA no contexto do ensino e aprendizagem (CASAS-ROMA *et al.*, 2021), principalmente, em um contexto no qual se observa um aumento contínuo do uso de tecnologias de IA aplicadas a este fim (VICARI, 2021).

Entretanto, as contribuições resultantes das investigações realizadas no decorrer deste projeto vão além da entrega de um modelo de AMA para a aprendizagem colaborativa e da sua respectiva implementação como prova de conceito. As discussões levantadas sugerem soluções de alcance mais abrangente.

Nesse sentido, ao definir e fundamentar a abordagem *top-down* como a mais adequada para o desenvolvimento de AMAs capazes de suportar processos de ensino e aprendizagem, delineiam-se diretivas e orientações para futuros pesquisadores na área. Da mesma forma, a instituição de princípios deontológicos e critérios utilitaristas para tomadas de decisão, significa um pequeno avanço no sentido de uma IA mais segura no campo da Educação.

Além disso, este trabalho traz para o âmbito da ética por design, questões que, até o momento, se davam apenas no âmbito da ética em design, da ética para o design e das especulações e previsões desastrosas sobre o mal uso da IA. É certo que estas últimas são tão importantes quanto a primeira, mas é por meio da ética por design que podem surgir soluções concretas para problemas de alinhamento de valores em IA.

Assim, considerando a abrangência das soluções apresentadas, o modelo proposto pode ser aplicado a outras técnicas de ensino e aprendizagem, tanto abordagens individuais quanto colaborativas. As escolhas feitas no presente trabalho em relação à abordagem de ensino e aos critérios éticos não são limitações do modelo proposto, mas escolhas pontuais que serviram

como diretivas para a implementação da prova de conceito, necessária para a sua validação. Tais escolhas possibilitaram testar a inclusão de critérios éticos na arquitetura de estados mentais, como é o caso do modelo BDI, e colocá-los em prática em um ambiente de testes.

Nesse contexto, com relação à proposta específica para o apoio ao ensino colaborativo, pode-se afirmar que o Ethoscool executou com sucesso o que era esperado nos testes, conforme apresentado no capítulo 3.4. Assim, referindo-se especialmente aos resultados mostrados na Tabela 7, pode-se observar uma tendência maior do agente em realizar intervenções, quanto mais o tempo avança em direção ao final da atividade.

Não se pode afirmar que este comportamento é esperado, mas sem dúvida é previsível, considerando a parametrização dos pesos e contrapesos para o algoritmo HAU, inserida no SMA. Tal resultado reflete o entendimento de que no início da atividade é possível ser mais tolerante com relação à autonomia do aluno, quando este não interage no fórum. Entretanto, na medida em que o tempo da atividade se aproxima do fim, torna-se mais preocupante o fato de haver estudantes ou grupos interagindo de forma insuficiente.

Em um cenário como este, considera-se que intervenções são necessárias para estimular o engajamento comportamental dos alunos. Entretanto, um entendimento diferente poderia ter influenciado e implicado em diferentes decisões por parte do agente, caso tivesse sido inserido no Ethoscool por meio dos seus parâmetros.

Além disso, também é importante notar que na Tabela 7 existem alguns registros com igual valor para o PLT. Nestes casos, a teoria utilitarista de Benthan prega que ambas as decisões sejam consideradas corretas (ANDERSON; ANDERSON, 2008). Contudo, se ambas podem ser consideradas corretas e, levando em conta o princípio de não interferir na autonomia do aluno, projetou-se o agente para que não realizasse intervenções nestes casos.

Estas características visam tornar o Ethoscool adequável a diferentes contextos, como a ética deve ser. Além disso, uma das suas maiores contribuições é a possibilidade de usar o modelo proposto para diferentes abordagens de ensino colaborativo. Por exemplo, pode-se utilizá-lo para trabalhos de pesquisa colaborativa no ensino a distância, no ensino híbrido, em tutorias de *Problem-based Learning*, entre outras. Por fim, é importante citar a significativa contribuição desta pesquisa, ao possibilitar um pequeno passo em direção a uma IA mais previsível, explicável e confiável para ser usada em processos de ensino.

4.4 FRAGILIDADES E POSSIBLIDADES FUTURAS PARA PESQUISA

Apesar das contribuições descritas no capítulo anterior, reconhece-se que a solução apresentada e descrita neste trabalho apresenta fragilidades. Entretanto, tais fragilidades podem ser transformadas em oportunidades de avanço no campo da ética em IA voltada à educação.

Dessa forma, apesar de ter servido bem ao propósito de possibilitar os primeiros testes envolvendo a inclusão de critérios éticos em ferramentas que usam IA aplicada ao ensino, a análise puramente quantitativa das interações pode não expressar com precisão o desempenho de um integrante em uma equipe de aprendizagem colaborativa. O modelo proposto, porém, é flexível o bastante para permitir que novas regras possam ser incluídas, assim como novos pesos e contrapesos de decisão para o raciocínio utilitarista.

Nesse sentido, a incorporação de outras técnicas de IA como: PLN para interpretar o conteúdo das interações dos alunos; mecanismos que explorem as emoções dos estudantes; e até ML, desde que não inclua em seus processos a aprendizagem de comportamento ético, pode ser bastante útil. A inclusão de novos recursos como os citados poderia contribuir significativamente para enriquecer a lista de critérios usada pelo Ethoscool para tomar suas decisões (CóRDOVA; VICARI, 2022).

A análise quantitativa das interações, tal qual apresentada, traz um viés baseado no preconceito de que a quantidade de interações determina a qualidade da colaboração e das contribuições realizadas pelos alunos. Por esta razão, recursos como os citados, capazes de avaliar qualitativamente as interações realizadas, podem tornar mais assertivas as decisões do agente Tutor.

Além disso, a completa implementação do Ethoscool, assim como a avaliação do seu desempenho ao perseguir o objetivo de estimular o engajamento em equipes de aprendizagem colaborativa, precisa ser testado em sala de aula. A presente pesquisa limitou-se a avaliar as possibilidades e os resultados da inclusão do raciocínio ético e da capacidade das tecnologias de IA de lidar com dilemas em processos de ensino.

Considera-se que os objetivos de pesquisa foram alcançados, mas a aplicação prática dos resultados apresentados representa avanços tão significativos quanto a proposição do modelo descrito. Dessa forma, não se considera, nem se poderia considerar de forma alguma, que pesquisas nesta área estejam encerradas. Oportunidades para pesquisas futuras se abrem a partir dos resultados apresentados no presente trabalho.

5 CONSIDERAÇÕES FINAIS

As discussões em torno do uso ético da IA tem ganhado espaço nas mais diferentes áreas que a empregam. Os riscos, intencionais ou não, decorrentes da aplicação de tais tecnologias, conforme apresentado no decorrer do presente trabalho, podem influenciar, entre ouras coisas, o a dignidade e o bem-estar humano, incorrendo, deste modo, em ações consideradas não éticas (DIGNUM, 2018). Dessa forma, áreas nas quais seres humanos estão mais diretamente expostos a influências de sistemas que usam IA, como é o caso da Educação, precisam buscar soluções capazes de aumentar a confiança das pessoas nesses sistemas.

Há uma ampla gama de evidências apontando para os riscos que a expansão da IA sem os devidos cuidados pode representar. Entre eles, destacam-se: riscos de viés, que podem implicar na replicação de discriminação e preconceitos; de instanciação perversa, que pode levar a um comportamento indesejado para atingir determinado fim programado; e riscos aos direitos humanos, como violações de privacidade, autonomia e segurança.

Nesse sentido, buscando contribuir para a mitigação de tais riscos no campo da Educação, mais especificamente em processos de ensino colaborativo, propôs-se o Ethoscool, um SMA no qual está inserido um AMA capaz de se orientar e tomar decisões levando em conta critérios considerados éticos. Para isso, recorreu-se a uma proposta baseada no modelo BDI para implementar uma solução usando a abordagem *top-down* para AMAs.

A abordagem *top-down*, nesse contexto, prega que a implementação de comportamentos considerados éticos em sistemas de IA, mais especificamente em AMAs, deve ser guiada por estruturas éticas existentes, como utilitarismo, deontologia, ética da virtude, entre outras. Esta abordagem não prevê a aprendizagem de comportamento ético por meio de mecanismos de ML. Por esta razão, considerou-se que seria a mais adequada para ambientes de sala de aula, uma vez que a aprendizagem de comportamentos considerados éticos poderia resultar na replicação de comportamentos considerados não éticos. Com isso, propor um modelo guiado por estruturas éticas estabelecidas e amplamente reconhecidas por uma comunidade de práticas, pode conduzir ao desenvolvimento de soluções capazes de atender, com maior efetividade, aos preceitos e requisitos de uma IA alinhada aos valores humanos.

Além disso, a escolha de um enfoque híbrido, capaz de aplicar em um só modelo a estrutura ética deontológica e a utilitarista, permitiu que se pudesse lidar com duas questões chave na área: a prevenção de comportamentos considerados não éticos, por meio da primeira e

o tratamento de dilemas éticos, por meio da segunda. Assim, no que diz respeito a princípios e regras deontológicas, procurou-se implementar um sistema alinhado a propostas de ética para IA apresentadas por importantes entidades de abrangência e relevância internacional, como UNESCO, IEEE e FLI. A dimensão utilitarista do Ethoscool, por sua vez, foi viabilizada por meio da implementação da teoria de Jeremy Bentham (1988), materializada no algoritmo HAU, que possibilita considerar o bem-estar de todos os envolvidos na decisão do agente.

Por fim, a escolha do modelo BDI para a implementação do agente se deu em função da sua característica de ser guiado por objetivos (desejos). Esta característica torna seus comportamentos mais previsíveis e explicáveis ao usuário final. Somando isso ao fato de o agente salvar suas decisões, bem como os dados usados para tomá-las, pode-se afirmar que se dispõe de meios para prover maior segurança e confiança aos operadores e afetados pelo Ethoscool, contribuindo para uma IA mais explicável e confiável no campo do ensino.

Com relação à viabilidade do modelo proposto, foi possível evidenciá-la por meio da implementação e teste do agente Tutor, responsável pelo comportamento e pelas decisões de cunho ético dentro do Ethoscool. Os resultados mostraram que o agente Tutor foi capaz de tomar decisões conforme o esperado em ambientes e estados criados para forçar a execução do seu raciocínio utilitarista.

A previsibilidade das decisões do agente, neste caso, está relacionada ao alinhamento do agente Tutor aos valores humanos. Este alinhamento, por um lado, se verifica nas regras – Tabela 1 – que orientam o seu comportamento, e que tem como base valores expressos em publicações no âmbito científico, como é o caso das publicações de Beauchamp e Childress (2001), e no âmbito político, como é o caso das publicações para ética em IA propostas por organizações internacionais. Por outro lado, o alinhamento a valores humanos se evidencia em sua dimensão utilitarista, cujos parâmetros de execução materializam o entendimento e os valores do professor quanto ao que deve ser levado em conta pelo agente ao executar suas avaliações e resoluções de dilemas éticos.

Além disso, conforme apresentado no capítulo 4, o Ethoscool pode ser considerado um AMA alinhado aos princípios para ética em IA. E, nesse contexto, apesar das suas fragilidades, que incluem uma forma estritamente quantitativa de avaliar a participação dos alunos em grupos de aprendizagem colaborativa, o Ethoscool abre várias possibilidades para pesquisas futuras, visando, precisamente, melhorar sua capacidade de avaliação. Para isso, diferentes técnicas de IA, como PLN, ML e recursos para avaliar as emoções dos estudantes podem ser empregadas.

Nos últimos anos têm sido observados diferentes estudos e propostas no âmbito dos AMAs. Alguns deles se assemelhando ao Ethoscool, seja quanto à sua abordagem *top-down* híbrida que aplica doutrinas éticas deontológica e utilitarista, seja quanto às tecnologias empregadas, como BDI e o algoritmo HAU.

Com relação ao primeiro conjunto, pode-se destacar o *MoralDM*, que opera de forma diferente do agente Tutor, analisando o problema em questão e identificando se há princípios fundamentais envolvidos na decisão a ser tomada. Caso não haja, seu processo decisório será executado com base em funções de utilidade, sendo, portanto, de ordem utilitarista. Do contrário, irá operar sobre uma base deontológica, preferindo a inação, se for o caso (DEHGHANI *et al.*, 2008; BLASS, 2016). Para o segundo conjunto, podem-se destacar o Jeremy e o *Casuist BDI-agent*. O primeiro é um sistema criado por Anderson *et al.* (2005) para prover conselhos baseados na doutrina utilitarista criada por Jeremy Bentham (1988). Esta implementação, entretanto, apesar de materializar com precisão uma estrutura ética completa, consistente e prática (ANDERSON *et al.*, 2005) não havia sido testada em outros contextos. O *Casuist BDI-agent*, por sua vez, é uma arquitetura que estende o modelo BDI (HONARVAR; GHASEM-AGHAEE, 2009), mas diferente do Ethoscool, faz uso de uma base de casos passados, que é atualizada sempre que o agente se depara com um novo caso, para tomar decisões. Nesse sentido, o *Casuist BDI-agent* é capaz de aprender com novos casos.

No caso do Ethoscool, optou-se por não implementar nenhum tipo de aprendizagem, nem usar uma base de casos, porque uma das variáveis mais importantes para o agente Tutor é o tempo da atividade, sendo difícil que dois cenários sejam suficientemente similares para que uma decisão seja baseada em um caso passado. Além disso, o Ethoscool prioriza o seu alinhamento com os valores do professor que irá operá-lo. Isto inviabiliza a utilização de bases com decisões padronizadas, ainda que esta seja atualizada pelo agente. A ideia principal, nesse sentido, é que cada decisão seja como se fosse única.

Com isso, as contribuições da presente pesquisa vão além da proposição de um modelo de AMA que busca promover o engajamento comportamental de alunos em grupos de aprendizagem colaborativa. A aplicação das proposições aqui exposta pode ser ampliada para outros contextos, estratégias e técnicas de ensino e aprendizagem. Por exemplo, estão entre algumas das contribuições deste trabalho a definição e fundamentação da abordagem *top-down* como a mais adequada para o desenvolvimento de AMAs capazes de suportar processos de ensino e aprendizagem e a instituição de princípios deontológicos e critérios utilitaristas para tomadas de

decisão nesse contexto. Tais contribuições delineiam e estabelecem diretivas e orientações para futuros pesquisadores na área.

Além disso, este trabalho traz a questão da ética em IA na Educação para o âmbito da ética por design, de onde podem surgir soluções concretas para problemas de alinhamento de valores em IA. Nesse sentido, pode-se concluir afirmando que, embora se reconheça que a solução proposta e descrita no decorrer destas páginas não seja a definitiva, as contribuições aqui delineadas certamente impulsionam o progresso da aplicação da informática na educação. Especificamente, destacam-se avanços no âmbito do uso ético da inteligência artificial, da IA explicável e na promoção da confiança das pessoas no emprego dessas tecnologias para aprimorar os processos educacionais.

REFERÊNCIAS

ABDALLA, Reem; MISHRA, Alok. Agent-oriented software engineering methodologies: Analysis and future directions. **Complexity**, v. 2021, p. 1629419, 2021. Publisher: Hindawi.

ALIMAN, Nadisha-Marie; KESTER, Leon. Requisite variety in ethical utility functions for AI value alignment. **Arxiv**, 2019.

ALLEN, Colin; SMIT, Iva; WALLACH, Wendell. Artificial morality: Top-down, bottom-up, and hybrid approaches. **Ethics and Information Technology**, v. 7, n. 3, p. 149–155, 2005. ISSN 1572-8439.

ALMULLA, Mohammed Abdullatif. The effectiveness of the project-based learning (PBL) approach as a way to engage students in learning. **Sage Open**, v. 10, n. 3, p. 2158244020938702, 2020. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

ALVARES, Luis Otavio; SICHMAN, Jaime Simão. Introdução aos sistemas multiagentes. *In*: **JORNADA DE ATUALIZAÇÃO EM INFORMÁTICA**. *[S.l.]*: UnB, 1997.

ANDERSON, Michael; ANDERSON, Susan Leigh. Ethical healthcare agents. **Advanced Computational Intelligence Paradigms in Healthcare - 3**, p. 233–257, 2008. Publisher: Springer, Berlin, Heidelberg.

ANDERSON, Michael; ANDERSON, Susan Leigh. Robot be good. **Scientific American**, Scientific American, a division of Nature America, Inc., v. 303, n. 4, p. 72–77, 2010. ISSN 00368733, 19467087. Disponível em: http://www.jstor.org/stable/26002215.

ANDERSON, Michael; ANDERSON, Susan Leigh. GenEth: a general ethical dilemma analyzer. **Paladyn, Journal of Behavioral Robotics**, v. 9, n. 1, p. 337–357, 2018.

ANDERSON, Michael; ANDERSON, Susan Leigh; ARMEN, Chris. MedEthEx: Toward a medical ethics advisor. *In*: BICKMORE, Timothy W. (Ed.). **Caring Machines: AI in Eldercare, Papers from the 2005 AAAI Fall Symposium, Arlington, Virginia, USA, November 4-6, 2005**. *[S.l.]*: AAAI Press, 2005. (AAAI Technical Report, FS-05-02), p. 9–16.

ANDINO, C. El lugar de la ética entre los saberes técnicos. Un abordaje filosófico. **Revista Científica de la UCSA**, scielo, v. 2, p. 85 – 94, 12 2015. ISSN 2409-8752.

ANNAS, Julia. Virtue ethics. **The Oxford handbook of ethical theory**, p. 515–536, 2006. Publisher: Oxford University Press New York.

ANSARI, Jamal Abdul Nasir; KHAN, Nawab Ali. Exploring the role of social media in collaborative learning the new domain of learning. **Smart Learning Environments**, v. 7, n. 1, p. 1–16, 2020. Publisher: SpringerOpen.

ARNOLD, Thomas; KASENBERG, Daniel; SCHEUTZ, Matthias. Value alignment or misalignment - what will keep systems accountable? *In*: . *[s.n.]*, 2017. Disponível em: https://openreview.net/forum?id=rJZMs1-ubH.

AROSKAR, Mila A. Anatomy of an ethical dilemma: The theory. **AJN The American Journal of Nursing**, v. 80, n. 4, p. 658–660, 1980. Publisher: LWW.

AUGELLO, Agnese; GENTILE, Manuel; DIGNUM, Frank. Social agents for learning in virtual environments. *In*: BOTTINO, Rosa; JEURING, Johan; VELTKAMP, Remco C. (Ed.). **Games and Learning Alliance**. *[S.l.]*: Springer International Publishing, 2016. p. 133–143. ISBN 978-3-319-50182-6.

BALDONI, Matteo; BAROGLIO, Cristina; MICALIZIO, Roberto. Fragility and robustness in multiagent systems. *In*: BAROGLIO, Cristina; HUBNER, Jomi F.; WINIKOFF, Michael (Ed.). **Engineering Multi-Agent Systems**. *[S.l.]*: Springer International Publishing, 2020. p. 61–77. ISBN 978-3-030-66534-0.

BARBOZA, Luciana Caixeta; GIORDAN, M. Analisando o diálogo em um processo de tutoria. *In*: **Anais do XIV Encontro Nacional de Didática e Prática de Ensino**. *[S.l.]*: EdiPUCRS, 2008.

BAUMANE-VITOLINA, Ilona; CALS, Igo; SUMILO, Erika. Is ethics rational? teleological, deontological and virtue ethics theories reconciled in the context of traditional economic decision making. **Procedia Economics and Finance**, v. 39, p. 108–114, 2016. ISSN 2212-5671.

BEAUCHAMP, Tom L; CHILDRESS, James F. **Principles of biomedical ethics**. 5. ed. *[S.l.]*: Oxford University Press, 2001.

BENTHAM, Jeremy. **The principles of moral and legislation**. *[S.l.]*: Prometheus, 1988.

BERCHT, Magda. **Em Direção a Agentes Pedagógicos com Dimensões Afetivas**. 2001. Tese, 2001.

BERENDT, Bettina; LITTLEJOHN, Allison; BLAKEMORE, Mike. AI in education: learner choice and fundamental rights. **Learning, Media and Technology**, 2020. ISSN 1743-9884. Publisher: Routledge.

BLASS, Joseph A. Interactive learning and analogical chaining for moral and commonsense reasoning. *In*: **Thirtieth AAAI Conference on Artificial Intelligence**. *[S.l.: s.n.]*, 2016.

BOCCONI, Stefania; CHIOCCARIELLO, Augusto; KAMPYLIS, Panagiotis; DAGIENė, Valentina; WASTIAU, Patricia; ENGELHARDT, Katja; EARP, Jeffrey; HORVATH, Milena; JASUTė, Eglė; MALAGOLI, Chiara; others. Reviewing computational thinking in compulsory education: State of play and practices from computing education. 2022. Publisher: Publications Office of the European Union.

BONNEMAINS, Vincent; SAUREL, Claire; TESSIER, Catherine. Embedded ethics: some technical and ethical challenges. **Ethics and Information Technology**, v. 20, n. 1, p. 41–58, 2018. Publisher: Springer.

BORDINI, Rafael Heitor. **Contributions to an Anthropological Approach to the Cultural Adaptation of Migrant Agents**. 1999. Tese, 1999.

BORGES, Oto; JULIO, Josimeire M; COELHO, Geide R. Efeitos de um ambiente de aprendizagem sobre o engajamento comportamental, o engajamento cognitivo e sobre a aprendizagem. **Encontro de Pesquisa em Ensino de Ciências**, v. 5, 2005.

BOSTROM, Nick. Astronomical waste: The opportunity cost of delayed technological development. **Utilitas**, v. 15, n. 3, p. 308–314, 2003. Publisher: Cambridge University Press.

BOSTROM, Nick. **Superintelligence: Paths, Dangers, Strategies**. *[S.l.]*: Oxford University Press, 2014.

BOURQUE, Pierre; FARLEY, Richard E. **Guide to the software engineering body of knowledge (SWEBOK (R)): Version 3.0**. *[S.l.]*: IEEE Computer Society Press, 2014.

BRASIL. **Lei 12.965, de 23 de abril de 2014. Estabelece princípios, garantias, direitos e deveres para o uso da Internet no Brasil**. *[S.l.]*: Presidência da República, 2014.

BRASIL. **Lei 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD)**. *[S.l.]*: Presidência da República, 2018.

BRASIL, Ministério da Ciência, Tecnologia e Informação. **Estratégia Brasileira de Inteligência Artificial -EBIA**. 2021. Disponível em: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/arquivosinteligenciaartificial/ia_estrategia_documento_referencia_4-979_2021.pdf.

BRATMAN, Michael. **Intention, Plans, and Practical Reason**. *[S.l.]*: Harvard University Press, 1987.

BRAUBACH, Lars; POKAHR, Alexander; LAMERSDORF, Winfried. Jadex: A BDI-agent system combining middleware and reasoning. *In*: **Software agent-based applications, platforms and development kits**. *[S.l.]*: Springer, 2005. p. 143–168.

BRIGGS, Gordon Michael; SCHEUTZ, Matthias. “sorry, i can’t do that”: Developing mechanisms to appropriately reject directives in human-robot interactions. *In*: **2015 AAAI Fall Symposium Series**. *[S.l.: s.n.]*, 2015.

BRITO, Josilene Almeida. **Engajamento em atividades assíncronas na modalidade de ensino a distância: requisitos de interfaces colaborativas**. 2010. Tese, 2010.

BROOKS, Rodney A. Intelligence without reason, computers and thought lecture. **Proceedings of IJCAI-91, Sidney, Australia**, v. 32, p. 37–38, 1991.

BRUFFEE, Kenneth A. Sharing our toys: Cooperative learning versus collaborative learning on JSTOR. **Change: The Magazine of Higher Learning**, v. 27, n. 1, p. 12–18, 1995.

CAETANO, Luciana Maria. A epistemologia genética de jean piaget. **ComCiência**, n. 120, p. 0–0, 2010. ISSN 1519-7654.

CAMARGO, Liseane Silveira; BECKER, Maria Luíza Rheingantz. O percurso do conceito de cooperação na epistemologia genética. **Educação & Realidade**, v. 37, p. 527–549, 2012. ISSN 0100-3143, 2175-6236. Publisher: Universidade Federal do Rio Grande do Sul - Faculdade de Educação.

CARDOSO, Rafael C.; FERRANDO, Angelo. A review of agent-based programming for multi-agent systems. **Computers**, v. 10, n. 2, 2021. ISSN 2073-431X.

CARNEIRO, Leonardo de Andrdade; GARCIA, Leandro Guimarães; BARBOSA, Gentil Veloso. Uma revisÃo sobre aprendizagem colaborativa mediada por tecnologias. **DESAFIOS - Revista Interdisciplinar da Universidade Federal do Tocantins**, v. 7, n. 2, p. 52–62, 2020. ISSN 2359-3652.

CASAS-ROMA, Joan; ARNEDO-MORENO, Joan. From games to moral agents: Towards a model for moral actions. **Artificial Intelligence Research and Development**, p. 19–28, 2019. Publisher: IOS Press.

CASAS-ROMA, Joan; CONESA, Jordi; CABALLé, Santi. Education, ethical dilemmas and AI: From ethical design to artificial morality. *In*: **Adaptive Instructional Systems. Design and Evaluation**. *[S.l.]*: Springer, Cham, 2021. p. 167–182.

CASTELFRANCHI, Cristiano. Founding agents’"autonomy"on dependence theory. *In*: **ECAI**. *[S.l.: s.n.]*, 2000. v. 1, p. 353–357.

CASTELFRANCHI, Cristiano; MICELI, Maria; CESTA, Amedeo. Dependence relations among autonomous agents. **decentralized AI**, v. 3, p. 215–227, 1992.

CASTRO, Alberto; MENEZES, Crediné. Aprendizagem colaborativa com suporte computacional. **Sistemas Colaborativos. Rio de Janeiro: Campus. ISBN**, p. 978–85, 2011.

CERVANTES, José-Antonio; LóPEZ, Sonia; RODRíGUEZ, Luis-Felipe; CERVANTES, Salvador; CERVANTES, Francisco; RAMOS, Félix. Artificial moral agents: A survey of the current status. **Science and Engineering Ethics**, v. 26, n. 2, p. 501–532, 2020. ISSN 1471-5546.

CHANEL, Guillaume; LALANNE, Denis; LAVOUé, Elise; LUND, Kristine; MOLINARI, Gaëlle; RINGEVAL, Fabien; WEINBERGER, Armin. Grand challenge problem 2: Adaptive awareness for social regulation of emotions in online collaborative learning environments. **Grand Challenge Problems in Technology-Enhanced Learning II: MOOCs and Beyond**, Springer, Cham, p. 13–16, 2016.

COELHO, Luana; PISONI, Silene. Vygotsky: sua teoria e a influência na educação. **Revista e-PED**, v. 2, n. 1, p. 144–152, 201.

COLABORAR. 7Graus, 2020. Disponível em: https://www.dicio.com.br/colaborar.

COLLAZOS, César A.; GUTIéRREZ, Francisco L.; GALLARDO, Jesús; ORTEGA, Manuel; FARDOUN, Habib M.; MOLINA, Ana Isabel. Descriptive theory of awareness for groupware development. **Journal of Ambient Intelligence and Humanized Computing**, v. 10, n. 12, p. 4789–4818, 2019. ISSN 1868-5145.

COOPERAR. 7Graus, 2020. Disponível em: https://www.dicio.com.br/cooperar.

CORDOVA, Paulo Roberto; VICARI, Rosa Maria. A CONCEPTUAL MODEL FOR ARTIFICIAL MORAL AGENTS (AMA) IN THE EDUCATIONAL CONTEXT. **International Journal of Development Research**, v. 11, p. 47869–47897, 2021. ISSN 2230-9926.

COSTA, Antônio Carlos da Rocha; COELHO, Helder Manuel Ferreira. Interactional moral systems: A model of social mechanisms for the moral regulation of exchange processes in agent societies. **IEEE Transactions on Computational Social Systems**, v. 6, n. 4, p. 778–796, 2019.

COSTAGUTA, Rosanna Nieves; MISSIO, Daniela Margarita; MANSILLA, Pablo Fernando Santana; LESCANO, Germán Ezequiel; MENINI, María de Los Ángeles. Caracterización de las interacciones colaborativas en ambientes de e-learning considerando conductas grupales y habilidades de colaboración. **Revista Internacional de Aprendizaje**, v. 2, n. 5, p. 139–159, 2019. ISSN 2575-5560.

CRISTANI, Matteo; BURATO, Elisa. Approximate solutions of moral dilemmas in multiple agent system. **Knowledge and Information Systems**, v. 18, n. 2, p. 157–181, 2009. ISSN 0219-3116.

CRUZ, Anderson; SANTOS, André V. dos; SANTIAGO, Regivan H. N.; BEDREGAL, Benjamin. A fuzzy semantic for BDI logic. **Fuzzy Information and Engineering**, v. 0, n. 0, p. 1–15, 2021.

CóRDOVA, Paulo Roberto; FILHO, Iderli Pereira De Souza; GUEDES, Gilleanes Thorwald Araujo; VICARI, Rosa Maria. ETHOSCHOOL: An artificial moral agent model for collaborative learning. *In*: FRASSON, Claude; MYLONAS, Phivos; TROUSSAS, Christos (Ed.). **Augmented Intelligence and Intelligent Tutoring Systems**. *[S.l.]*: Springer Nature Switzerland, 2023. v. 13891, p. 281–289. ISBN 978-3-031-32882-4 978-3-031-32883-1.

CóRDOVA, Paulo Roberto; VICARI, Rosa Maria. Practical ethical issues for artificial intelligence in education. *In*: REIS, Arsénio; BARROSO, João; MARTINS, Paulo; JIMOYIANNIS, Athanassios; HUANG, Ray Yueh-Min; HENRIQUES, Roberto (Ed.). **Technology and Innovation in Learning, Teaching and Education**. *[S.l.]*: Springer Nature Switzerland, 2022. v. 1720, p. 437–445. ISBN 978-3-031-22917-6 978-3-031-22918-3.

CóRDOVA, Paulo Roberto; VICARI, Rosa Maria; BRUSIUS, Carlos; COELHO, Helder. A proposal for artificial moral pedagogical agents. *In*: ROCHA, Álvaro; ADELI, Hojjat; DZEMYDA, Gintautas; MOREIRA, Fernando; CORREIA, Ana Maria Ramalho (Ed.). **Trends and Applications in Information Systems and Technologies**. *[S.l.]*: Springer International Publishing, 2021. p. 396–401. ISBN 978-3-030-72657-7.

DAVIDSON, Neil; MAJOR, Claire Howell. Boundary crossings: Cooperative learning, collaborative learning, and problem-based learning. **Journal on excellence in college teaching**, v. 25, 2014.

DAVIDSON, Neil; WORSHAM, Toni. **Enhancing Thinking through Cooperative Learning.** *[S.l.]*: Teachers College Press, 1234 Amsterdam Avenue, New York, NY 10027 ($21.95). Tel: 212-678-3963; Tel: 800-575-6560 (Toll Free)., 1992. ISBN 978-0-8077-3157-4.

DEHGHANI, Morteza; TOMAI, Emmett; FORBUS, Kenneth D; KLENK, Matthew. An integrated reasoning approach to moral decision-making. *In*: **AAAI**. *[S.l.: s.n.]*, 2008. p. 1280–1286.

DELJOO, Ameneh; ENGERS, Tom M van; DOESBURG, Robert van; GOMMANS, Leon; LAAT, Cees de; others. A normative agent-based model for sharing data in secure trustworthy digital market places. *In*: **ICAART (1)**. *[S.l.: s.n.]*, 2018. p. 290–296.

DENNETT, Daniel Clement. **The intentional stance**. *[S.l.]*: MIT press, 1989.

DENNIS, Louise; FISHER, Michael; SLAVKOVIK, Marija; WEBSTER, Matt. Formal verification of ethical choices in autonomous systems. **Robotics and Autonomous Systems**, v. 77, p. 1–14, 2016. Publisher: Elsevier.

DIGNUM, Virginia. Ethics in artificial intelligence: introduction to the special issue. **Ethics and Information Technology**, v. 20, n. 1, p. 1–3, 2018. ISSN 1572-8439.

DIGNUM, Virginia; BALDONI, Matteo; BAROGLIO, Cristina; CAON, Maurizio; CHATILA, Raja; DENNIS, Louise; GéNOVA, Gonzalo; HAIM, Galit; KLIEß, Malte S; LOPEZ-SANCHEZ, Maite; others. Ethics by design: Necessity or curse? *In*: **Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society**. *[S.l.: s.n.]*, 2018. p. 60–66.

DILLENBOURG, Pierre. What do you mean by collaborative learning? *In*: **Collaborativelearning: Cognitive and Computational Approaches.** *[S.l.]*: Elsevier, 1999. p. 1 – 19.

DORRI, Ali; KANHERE, Salil S.; JURDAK, Raja. Multi-agent systems: A survey. **IEEE Access**, v. 6, p. 28573–28593, 2018.

DOURISH, Paul; BELLOTTI, Victoria. Awareness and coordination in shared workspaces. *In*: **Proceedings of the 1992 ACM conference on Computer-supported cooperative work**. *[S.l.: s.n.]*, 1992. p. 107–114.

ENGELS, R; REHBEIN, M; SCHMIEDCHEN, F; STAPF-FINé, H; SüLZEN, A. Policy paper on the asilomar principles on artificial intelligence. **Federation of German Scientist (VDW): Berlin, Germany**, 2018.

EPSTEIN, Joshua M. **Agent_Zero: Toward Neurocognitive Foundations for Generative Social Science**. *[S.l.]*: Princeton University Press, 2014. ISBN 978-0-691-15888-4.

FAVARETTO, Maddalena; CLERCQ, Eva De; ELGER, Bernice Simone. Big data and discrimination: perils, promises and solutions. a systematic review. **Journal of Big Data**, Springer Open, v. 6, n. 1, p. 1–27, 2019. ISSN 2196-1115.

FILHO, Osterne Nonato Maia; CHAVES, Hamilton Viana; SEIXAS, Pablo de Sousa. Por uma educação para a autonomia de sujeitos situados no mundo. **Psicologia da Educação**, n. 46, 2018. ISSN 2175-3520. Number: 46.

FINN, Jeremy D; ROCK, Donald A. Academic success among students at risk for school failure. **Journal of applied psychology**, v. 82, n. 2, p. 221, 1997. Publisher: American Psychological Association.

FISHER, Michael; GHIDINI, Chiara; HIRSCH, Benjamin. Organising logic-based agents. *In*: **Formal Approaches to Agent-Based Systems**. *[S.l.]*: Springer, Berlin, Heidelberg, 2002. p. 15–27.

FLI, Future of Life Institute. **AI Principles - Future of Life Institute**. 2021. Disponível em: https://futureoflife.org/ai-principles/?cn-reloaded=1.

FORMOSA, Paul; RYAN, Malcolm. Making moral machines: why we need artificial moral agents. **AI & SOCIETY**, Springer, London, p. 1–13, 2020. ISSN 1435-5655.

FREDRICKS, Jennifer; McColskey, Wendy; MELI, Jane; MORDICA, Joy; MONTROSSE, Bianca; MOONEY, Kathleen. Measuring student engagement in upper elementary through high school: A description of 21 instruments. issues & answers. REL 2011-no. 098. **Regional Educational Laboratory Southeast**, 2011. Publisher: ERIC.

FUKS, Hugo; RAPOSO, Alberto; GEROSA, Marco A.; PIMENTAL, Mariano; LUCENA, Carlos J. P. chapter, **The 3C Collaboration Model**. *[S.l.]*: IGI Global, 2008.

FUKS, Hugo; RAPOSO, Alberto Barbosa; GEROSA, Marco Aurélio; PIMENTEL, Mariano; FILIPPO, Denise; LUCENA, CJP de. Teorias e modelos de colaboração. **Sistemas colaborativos**, Elsevier, p. 16–33, 2011.

GARCíA-PEñALVO, Francisco José; LLORENS-LARGO, Faraón; VIDAL, Javier. La nueva realidad de la educación ante los avances de la inteligencia artificial generativa. **Revista Iberoamericana de Educación a Distancia**, v. 27, n. 1, 2023. ISSN 1390-3306.

GEORGEFF, Michael; PELL, Barney; POLLACK, Martha; TAMBE, Milind; WOOLDRIDGE, Michael. The belief-desire-intention model of agency. *In*: **International workshop on agent theories, architectures, and languages**. *[S.l.]*: Springer, 1998. p. 1–10.

GIPS, James. **Toward the ethical robot**. 1995.

GIRAFFA, Lucia Maria Martins. **Uma Arquitetura de tutor utilizando estados mentais**. 1999. Tese, 1999.

GIRARDI, Rosario. Engenharia de software baseada em agentes. *In*: **Procedimentos do IV Congresso Brasileiro de Ciência da Computação CBComp**. *[S.l.]*: Editora UNIVALI, 2004. p. 913–937.

GUEDES, Gilleanes Thorwald Araujo. MASRML-a domain-specific modeling language for multi-agent systems requirements. **International Journal of Software Engineering & Applications (IJSEA)**, v. 11, n. 5, 2020.

GUEDES, Gilleanes Thorwald Araujo; VICARI, Rosa Maria. Specific UML-derived languages for modeling multi-agent systems. *In*: **Handbook on Artificial Intelligence-Empowered Applied Software Engineering: VOL. 1: Novel Methodologies to Engineering Smart Software Systems**. *[S.l.]*: Springer, 2022. p. 159–223.

GUERINI, Marco; PIANESI, Fabio; STOCK, Oliviero. Is it morally acceptable for a system to lie to persuade me? *In*: **Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence**. *[S.l.: s.n.]*, 2015.

GUNNING, David; AHA, David. DARPA's explainable artificial intelligence (XAI) program. **AI Magazine**, v. 40, n. 2, p. 44–58, 2019. ISSN 2371-9621.

HAAS, Julia. Moral gridworlds: A theoretical proposal for modeling artificial moral cognition. **Minds and Machines**, v. 30, n. 2, p. 219–246, 2020. ISSN 1572-8641.

HAN, Shengnan; KELLY, Eugene; NIKOU, Shahrokh; SVEE, Eric-Oluf. Aligning artificial intelligence with human values: reflections from a phenomenological perspective. **AI & SOCIETY**, Springer, London, p. 1–13, 2021. ISSN 1435-5655.

HAN, The Anh; PEREIRA, Luís Moniz. Evolutionary machine ethics. *In*: BENDEL, Oliver (Ed.). **Handbuch Maschinenethik**. *[S.l.]*: Springer Fachmedien Wiesbaden, 2019. p. 229–253. ISBN 978-3-658-17483-5.

HAYASHI, Yugo. Gaze awareness and metacognitive suggestions by a pedagogical conversational agent: an experimental investigation on interventions to support collaborative learning process and performance. **International Journal of Computer-Supported Collaborative Learning**, Springer US, v. 15, n. 4, p. 469–498, 2020. ISSN 1556-1615.

HOLTGRAVES, T. M.; ROSS, S. J.; WEYWADT, C. R.; HAN, T. L. Perceiving artificial social agents. **Computers in Human Behavior**, v. 23, n. 5, p. 2163–2174, 2007. ISSN 0747-5632.

HONARVAR, Ali Reza; GHASEM-AGHAEE, Nasser. Casuist BDI-agent: a new extended BDI architecture with the capability of ethical reasoning. *In*: **International conference on artificial intelligence and computational intelligence**. *[S.l.]*: Springer, 2009. p. 86–95.

IEEE, Institute of Electrical Electronics Engineers. **The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition.** 2019. Disponível em: https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.htm.

ISOHäTäLä, Jaana; NäYKKI, Piia; JäRVELä, Sanna; BAKER, Michael J.; LUND, Kristine. Social sensitivity: a manifesto for CSCL research. **International Journal of Computer-Supported Collaborative Learning**, v. 16, n. 2, p. 289–299, 2021. ISSN 1556-1615.

ISO/IEC/IEEE, 29119-1. **Software and systems engineering - Software testing - Part 1: Concepts and definitions.** *[S.l.]*: International Organization for Standardization, 2022.

JACOBS, George M. Collaborative learning or cooperative learning? the name is not important; flexibility is. **Online Submission**, v. 3, n. 1, p. 32–52, 2015.

JAZAYERI, Ali; BASS, Ellen J. Agent-oriented methodologies evaluation frameworks: a review. **International Journal of Software Engineering and Knowledge Engineering**, v. 30, n. 9, p. 1337–1370, 2020. Publisher: World Scientific.

JONG, Ton de; GILLET, Denis; RODRíGUEZ-TRIANA, María Jesús; HOVARDAS, Tasos; DIKKE, Diana; DORAN, Rosa; DZIABENKO, Olga; KOSLOWSKY, Jens; KORVENTAUSTA, Miikka; LAW, Effie; PEDASTE, Margus; TASIOPOULOU, Evita; VIDAL, Gérard; ZACHARIA, Zacharias C. Understanding teacher design practices for digital inquiry–based science learning: the case of go-lab. **Educational Technology Research and Development**, Springer US, v. 69, n. 2, p. 417–444, 2021. ISSN 1556-6501.

JäRVELä, Sanna; ROSé, Carolyn. Forms of collaboration matters: CSCL across the contexts. **International Journal of Computer-Supported Collaborative Learning**, v. 16, n. 2, p. 145–149, 2021. ISSN 1556-1615.

KANKANHALLI, A; CHARALABIDIS, Yannis; MELLOULI, Sehl. IoT and AI for smart government: A research agenda. **Government Information Quarterly**, v. 36, n. 2, p. 304–309, 2019. ISSN 0740-624X. Publisher: JAI.

KANT, Immanuel. **Metafísica dos Costumes**. *[S.l.]*: Editora Vozes; Editora São Francisco, 2013.

KIM, Tae Wan; DONALDSON, Thomas; HOOKER, John. Grounding value alignment with ethical principles. **arXiv preprint arXiv:1907.05447**, 2019.

KIM, Tae Wan; HOOKER, John; DONALDSON, Thomas. Taking principles seriously: A hybrid approach to value alignment in artificial intelligence. **Journal of Artificial Intelligence Research**, v. 70, p. 871–890, 2021. ISSN 1076-9757.

LIBANEO, José Carlos. Tendências pedagógicas na prática escolar. **Revista da Associação Nacional de Educação–ANDE**, v. 3, p. 11–19, 1983.

LIN, Lijia; GINNS, Paul; WANG, Tianhui; ZHANG, Peilin. Using a pedagogical agent to deliver conversational style instruction: What benefits can you obtain? **Computers & Education**, v. 143, p. 103658, 2020. ISSN 0360-1315.

LUCK, Michael; D'INVERNO, Mark. A formal framework for agency and autonomy. *In*: **Icmas**. *[S.l.: s.n.]*, 1995. v. 95, p. 254–260.

LäMSä, Joni; HäMäLäINEN, Raija; KOSKINEN, Pekka; VIIRI, Jouni; LAMPI, Emilia. What do we do when we analyse the temporal aspects of computer-supported collaborative learning? a systematic literature review. **Educational Research Review**, v. 33, p. 100387, 2021. ISSN 1747-938X. Publisher: Elsevier.

MABASO, Bongani Andy. Artificial moral agents within an ethos of AI4sg. **Philosophy & Technology**, 2020. ISSN 2210-5441.

MADL, Tamas; FRANKLIN, Stan. Constrained incrementalist moral decision making for a biologically inspired cognitive architecture. *In*: TRAPPL, Robert (Ed.). **A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations**. *[S.l.]*: Springer International Publishing, 2015. p. 137–153.

MALLE, Bertram F. Integrating robot ethics and machine morality: the study and design of moral competence in robots. **Ethics and Information Technology**, v. 18, n. 4, p. 243–256, 2016. ISSN 1572-8439. Disponível em: https://doi.org/10.1007/s10676-015-9367-8.

MARTHA, Ati Suci Dian; SANTOSO, Harry B. The design and impact of the pedagogical agent: A systematic literature review. **Journal of Educators Online**, Journal of Educators Online. 500 University Drive, Dothan, v. 16, n. 1, 2019. ISSN EISSN-1547-500X.

MATTHEWS, Roberta S. Collaborative learning: Creating knowledge with students. **Teaching on solid ground: Using scholarship to improve practice**, v. 1996, p. 101–124, 1996.

MATTHEWS, Roberta S.; COOPER, James L.; DAVIDSON, Neil; HAWKES, Peter. Building bridges between cooperative and collaborative learning. **Change: The Magazine of Higher Learning**, v. 27, n. 4, p. 35–40, 1995.

McCarthy, John. What is artificial intelligence? 2007.

McIntyre, Alison. Doctrine of double effect. *In*: ZALTA, Edward N. (Ed.). **The Stanford Encyclopedia of Philosophy**. Spring 2019. *[S.l.]*: Metaphysics Research Lab, Stanford University, 2019.

MISSELHORN, Catrin. Artificial systems with moral capacities? a research design and its implementation in a geriatric care system. **Artificial Intelligence**, v. 278, p. 103179, 2020. ISSN 0004-3702.

MOOR, J.H. The nature, importance, and difficulty of machine ethics. **IEEE Intelligent Systems**, v. 21, n. 4, p. 18–21, 2006.

MORRIS, R.; HADWIN, A. F.; GRESS, C. L. Z.; MILLER, M.; FIOR, M.; CHURCH, H.; WINNE, P. H. Designing roles, scripts, and prompts to support CSCL in gStudy. **Computers in Human Behavior**, v. 26, n. 5, p. 815–824, 2010. ISSN 0747-5632.

NASCIMENTO, Ernandes Rodrigues do; PADILHA, Maria Auxiliadora Soares. Aprendizagem por meio do ensino híbrido na educação superior: narrando o engajamento dos estudantes. **Revista Diálogo Educacional**, v. 20, p. 252 – 271, 2020. ISSN 1981-416x. Publisher: scielo.

NATH, Rajakishore; SAHU, Vineet. The problem of machine ethics in artificial intelligence. **AI & SOCIETY**, v. 35, n. 1, p. 103–111, 2020. ISSN 1435-5655.

NG, Peggy M. L.; CHAN, Jason K. Y.; LIT, Kam Kong. Student learning performance in online collaborative learning. **Education and Information Technologies**, v. 27, n. 6, p. 8129–8145, 2022. ISSN 1573-7608.

NOLFI, Stefano. Power and the limits of reactive agents. **Neurocomputing**, Elsevier, v. 42, n. 1, p. 119–145, 2002.

NUNES, Clarice. Historiografia comparada da escola nova: algumas questões. **Revista da Faculdade de Educação**, v. 24, p. 105–125, 1998. ISSN 0102-2555.

OECD. **Recommendation of the council on artificial intelligence**. Paris: OECD, 2021a. (OECD Legal Instruments, n. 0449), 2021. Disponível em: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

OLIVEIRA, Leonam Cordeiro de; GUERINO, Guilherme Corredato; OLIVEIRA, Leander Cordeiro de; PIMEN, Andrey Ricardo. Information and communication technologies in education 4.0 paradigm: a systematic mapping study. **Informatics in Education**, v. 22, n. 1, p. 71–98, 2023. ISSN 1648-5831. Publisher: Vilnius University Institute of Data Science and Digital Technologies.

PFORDTEN, Dietmar von der. Five elements of normative ethics - a general theory of normative individualism. **Ethical Theory and Moral Practice**, v. 15, n. 4, p. 449–471, 2012. ISSN 1572-8447.

POKAHR, Alexander; BRAUBACH, Lars; LAMERSDORF, Winfried. Jadex: A BDI reasoning engine. *In*: **Multi-agent programming**. *[S.l.]*: Springer, 2005. p. 149–174.

QUINN, Warren S. Actions, intentions, and consequences: The doctrine of double effect. **Philosophy & Public Affairs**, p. 334–351, 1989. Publisher: JSTOR.

REDDY, Sandeep; ALLAN, Sonia; COGHLAN, Simon; COOPER, Paul. A governance model for the application of AI in health care. **Journal of the American Medical Informatics Association**, v. 27, n. 3, p. 491–497, 2020. Publisher: Oxford Academic.

RIEDER, Travis N; HUTLER, Brian; MATHEWS, Debra JH. Artificial intelligence in service of human needs: Pragmatic first steps toward an ethics for semi-autonomous agents. **AJOB neuroscience**, v. 11, n. 2, p. 120–127, 2020.

RIOS, Terezinha Azeredo. **Ética na docência universitária: a caminho de uma universidade pedagógica**. *[S.l.]*: USP, 2009.

RODRIGUES, Karin Débora; BARROS, Irany Gomes; FRAGUAS, Andreia Dutra. TENDÊNCIAS PEDAGÓGICAS ATUAIS. *In*: **Educação como (re)Existência: mudanças, conscientização e conhecimentos**. *[S.l.]*: Editora Realize, 2020.

ROSENBLUETH, Arturo; WIENER, Norbert. Purposeful and non-purposeful behavior. **Philosophy of Science**, The University of Chicago Press, Philosophy of Science Association, v. 17, n. 4, p. 318–326, 1950. ISSN 00318248, 1539767X.

ROSS, William David. **The right and the good**. *[S.l.]*: Clarendon Press, 1930.

RUSSEL, Stuart; NORVIG, Peter. **INTELIGÊNCIA ARTIFICIAL**. 3. ed. *[S.l.]*: Elsevier, 2013.

RUSSELL, Stuart; DEWEY, Daniel; TEGMARK, Max. Research priorities for robust and beneficial artificial intelligence. **Ai Magazine**, v. 36, n. 4, p. 105–114, 2015.

RUZZI, Patricio E.; PITT, Jeremy; BUSQUETS, Dídac. Electronic social capital for self-organising multi-agent systems. **ACM Transactions on Autonomous and Adaptive Systems (TAAS)**, 2017. Publisher: ACM PUB27 New York, NY, USA.

RăILEANU, Silviu; ANTON, Florin Daniel; BORANGIU, Theodor; ANTON, Silvia. Design of high availability manufacturing resource agents using JADE framework and cloud replication. *In*: BORANGIU, Theodor; TRENTESAUX, Damien; THOMAS, André; CARDIN, Olivier (Ed.). **Service Orientation in Holonic and Multi-Agent Manufacturing: Proceedings of SOHOMA 2017**. *[S.l.]*: Springer International Publishing, 2018. p. 201–215.

SALLES, Arleen; EVERS, Kathinka; FARISCO, Michele. Anthropomorphism in AI. **AJOB Neuroscience**, v. 11, n. 2, p. 88–95, 2020.

SANTOS, Francisco Alan de Oliveira; JUNIOR, Erivaldo A. S.; OLIVEIRA, Lydia Bruna A.; DUARTE, Salvino. Mapeamento sistemático sobre aprendizagem colaborativa com suporte

computacional no brasil / systematic mapping on collaborative learning with computational support in brazil. **Brazilian Journal of Development**, v. 6, n. 1, p. 91–102, 2020. ISSN 2525-8761.

SAPUTRA, DhanarIntan Surya; MANONGGA, Danny. The concept and future of intelligent personal assistant based on artificial autonomy for gamers in indonesia. **Design Engineering**, p. 1940–1956, 2021.

SARKER, Iqbal H.; HOQUE, Mohammed Moshiul; UDDIN, Md Kafil; ALSANOOSY, Tawfeeq. Mobile data science and intelligent apps: Concepts, AI-based modeling and research directions. **Mobile Networks and Applications**, v. 26, n. 1, p. 285–303, 2021. ISSN 1572-8153.

SCHROEDER, Mark. Normative ethics and metaethics. *In*: **Routledge handbook of metaethics**. *[S.l.]*: Routledge, 2017. p. 674–686.

SEARLE, John R. Is the brain's mind a computer program? **Scientific American**, Scientific American, a division of Nature America, Inc., v. 262, n. 1, p. 25–31, 1990. ISSN 00368733, 19467087.

SEARLE, John R; SEARLE, John Rogers. **Speech acts: An essay in the philosophy of language**. *[S.l.]*: Cambridge university press, 1969. v. 626.

SHOHAM, Yoav. Agent-oriented programming. **Artificial intelligence**, Elsevier, v. 60, n. 1, p. 51–92, 1993.

SILVA, Viviane Torres Da; LUCENA, Carlos JP de. MAS-ML: a multi-agent system modeling language. *In*: **Companion of the 18th annual ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications**. *[S.l.: s.n.]*, 2003. p. 126–127.

SOARES, Nate; FALLENSTEIN, Benja. Aligning superintelligence with human interests: A technical research agenda. **Machine Intelligence Research Institute (MIRI) technical report**, v. 8, 2014. Publisher: Citeseer.

STAHL, Gerry; KOSCHMANN, Timothy D; SUTHERS, Daniel D. **Computer-supported collaborative learning: An historical perspective**. *[S.l.]*: Cambridge handbook of the learning sciences, 2006.

SUBAGDJA, Budhitama; TAN, Ah-Hwee. Beyond autonomy: The self and life of social agents. *In*: **Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems**. *[S.l.: s.n.]*, 2019. p. 1654–1658.

SVENMARCK, Peter; LUOTSINEN, Linus; NILSSON, Mattias; SCHUBERT, Johan. Possibilities and challenges for artificial intelligence in military applications. *In*: **STO-MP-IST-160**. *[S.l.: s.n.]*, 2018. ISBN 978-92-837-2181-9.

THORNTON, Sarah M.; PAN, Selina; ERLIEN, Stephen M.; GERDES, J. Christian. Incorporating ethical considerations into automated vehicle control. **IEEE Transactions on Intelligent Transportation Systems**, v. 18, n. 6, p. 1429–1439, 2017.

TONI, Francesca; BENTAHAR, Jamal. Computational logic-based agents. **Autonomous Agents and Multi-Agent Systems**, Boston, MA: Kluwer Academic Publishers, 1998, v. 16, n. 3, p. 211–213, 2008.

TORRES, Patrícia Lupion; Irala, Esrom Adriano Freitas. Aprendizagem colaborativa: teoria e prática. **Complexidade: redes e conexões na produção do conhecimento. Curitiba: Senar**, p. 61–93, 2014.

UNESCO. **FIRST DRAFT OF THE RECOMMENDATION ON THE ETHICS OF ARTIFICIAL INTELLIGENCE**. 2020.

VAMPLEW, Peter; DAZELEY, Richard; FOALE, Cameron; FIRMIN, Sally; MUMMERY, Jane. Human-aligned artificial intelligence is a multiobjective problem. **Ethics and Information Technology**, v. 20, n. 1, p. 27–40, 2018. ISSN 1572-8439.

VEIGA, Feliciano Henriques. Envolvimento dos alunos na escola: Elaboração de uma nova escala de avaliação. **International Journal of Developmental and Educational Psychology**, v. 1, p. 441–450, 2013.

VICARI, Rosa Maria. Influências das tecnologias da inteligência artificial no ensino. **Estudos Avançados**, v. 35, n. 101, p. 73–84, 2021.

VYGOTSKY, Lev. **A formação social da mente: o desenvolvimento dos processos psicológicos superiores**. *[S.l.]*: Martins Fontes, 1998. (6. ed.).

WALLACH, Wendell; ALLEN, Colin. **Moral machines: Teaching robots right from wrong**. *[S.l.]*: Oxford University Press, 2009.

WALLACH, Wendell; FRANKLIN, Stan; ALLEN, Colin. A conceptual and computational model of moral decision making in human and artificial agents. **Topics in cognitive science**, v. 2, n. 3, p. 454–485, 2010. Publisher: Wiley Online Library.

WATSON, David. The rhetoric and reality of anthropomorphism in artificial intelligence. **Minds and Machines**, v. 29, n. 3, p. 417–440, 2019. ISSN 1572-8641.

WEISS, Gerhard. **Multiagent systems: a modern approach to distributed artificial intelligence**. *[S.l.]*: MIT press, 1999.

WOOLDRIDGE, Michael. Intelligent agents: The key concepts. *In*: **ECCAI Advanced Course on Artificial Intelligence**. *[S.l.]*: Springer, 2001. p. 3–43.

WOOLF, Beverly. Introduction to IJAIED special issue, FATE in AIED. **International Journal of Artificial Intelligence in Education**, v. 32, n. 3, p. 501–503, 2022. Publisher: Springer.

WYNSBERGHE, Aimee van; ROBBINS, Scott. Critiquing the reasons for making artificial moral agents. **Science and Engineering Ethics**, v. 25, n. 3, p. 719–735, 2019. ISSN 1471-5546.

Zagzebski, Linda. EXEMPLARIST VIRTUE THEORY. **Metaphilosophy**, v. 41, n. 1, p. 41–57, 2010.

ZOSHAK, John; DEW, Kristin. Beyond kant and bentham: How ethical theories are being used in artificial moral agents. *In*: **Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems**. *[S.l.: s.n.]*, 2021. p. 1–15.

ANEXO

ANEXO A – ANEXO A - LISTA COMPLETA DOS PARÂMETROS PARA PESOS E CONTRA-PESOS USADOS PELO ALGORITMO HAU

Tipo	Cenário	Escopo	Tempo Decorrido	Prazer	Probabilidade	Duração
GNI	true	G	0	0	0.5	0
GNI	true	G	0.25	-1	0.5	1
GNI	true	G	0.5	1	0.75	1
GNI	true	G	0.75	2	0.8	2
GNI	true	G	0.9	1	0.95	1
GNI	false	G	0	0	0.95	0
GNI	false	G	0.25	0	0.7	0
GNI	false	G	0.5	0	0.5	0
GNI	false	G	0.75	0	0.8	0
GNI	false	G	0.9	-1	0.95	2
ANT	true	G	0	0	0.5	0
ANT	true	G	0.25	-1	0.5	1
ANT	true	G	0.5	0	0.7	0
ANT	true	G	0.75	1	0.8	2
ANT	true	G	0.9	-1	0.95	1
ANT	false	G	0	0	0.5	0
ANT	false	G	0.25	0	0.6	0
ANT	false	G	0.5	-1	0.6	1
ANT	false	G	0.75	-1	0.6	1
ANT	false	G	0.9	-2	0.5	2
ANT	true	A	0	0	0.7	0
ANT	true	A	0.25	-1	0.5	1
ANT	true	A	0.5	1	0.75	1
ANT	true	A	0.75	1	0.5	1
ANT	true	A	0.9	1	0.7	1
ANT	false	A	0	0	0.5	0
ANT	false	A	0.25	1	0.6	1
ANT	false	A	0.5	0	0.6	1
ANT	false	A	0.75	-1	0.6	1
ANT	false	A	0.9	-1	0.5	2
ANI	true	G	0	1	0.5	1
ANI	true	G	0.25	1	0.6	1
ANI	true	G	0.5	2	0.7	2
ANI	true	G	0.75	1	0.7	2
ANI	true	G	0.9	1	0.9	2
ANI	false	G	0	0	0.5	0
ANI	false	G	0.25	0	0.6	0
ANI	false	G	0.5	-1	0.7	1
ANI	false	G	0.75	-2	0.7	2
ANI	false	G	0.9	-2	0.9	2
ANI	true	A	0	-1	0.5	2
ANI	true	A	0.25	-1	0.6	2
ANI	true	A	0.5	0	0.7	2
ANI	true	A	0.75	1	0.8	2
ANI	true	A	0.9	1	0.7	2
ANI	false	A	0	0	0.5	0
ANI	false	A	0.25	0	0.5	0
ANI	false	A	0.5	1	0.6	2
ANI	false	A	0.75	-1	0.6	2

ANI	false	A	0.9	-1	0.9	2
ANE	true	G	0	0	0.5	0
ANE	true	G	0.25	1	0.5	1
ANE	true	G	0.5	1	0.7	2
ANE	true	G	0.75	2	0.8	2
ANE	true	G	0.9	2	0.95	2
ANE	false	G	0	0	0.5	0
ANE	false	G	0.25	0	0.5	0
ANE	false	G	0.5	0	0.7	0
ANE	false	G	0.75	-1	0.8	1
ANE	false	G	0.9	-1	0.9	2
ANE	true	A	0	-2	0.6	2
ANE	true	A	0.25	-1	0.6	2
ANE	true	A	0.5	1	0.5	2
ANE	true	A	0.75	1	0.6	2
ANE	true	A	0.9	-1	0.7	2
ANE	false	A	0	0	0.5	0
ANE	false	A	0.25	0	0.5	0
ANE	false	A	0.5	0	0.5	0
ANE	false	A	0.75	-1	0.6	1
ANE	false	A	0.9	-1	0.9	1