Modelagem formal de abordagens éticas para comportamento de agentes

João Vicente Markovicz[1], **Gleifer Vaz Alves**[1]

[1]Departamento Acadêmico de Informática
Universidade Tecnológica Federal do Paraná (UTFPR)
Ponta Grossa – PR – Brasil

joao.080603@alunos.utfpr.edu.br gleifer@utfpr.edu.br

Abstract. *This paper formally models three ethical approaches (deontology, consequentialism and virtues ethics) in agents' decision-making. For that, we have used the UPPAAL model checker for creating (timed) automata and verifying properties related to the agent's behaviour.*

Resumo. *Este artigo apresenta a modelagem formal de três abordagens éticas (deontologia, consequencialismo e ética das virtudes) na tomada de decisão de agentes. Para tal, foi usada a ferramenta de model checking UPPAAL na construção de autômatos (temporais) e verificação de propriedades a respeito do comportamento dos agentes.*

1. Introdução

O desenvolvimento e verificação de sistemas autônomos é um desafio, pois envolve ambientes complexos e comportamento diversificado [Dennis and Fisher 2023]. Um exemplo de sistema autônomo é a condução de Veículos Autônomos (VAs), os quais podem operar em situações críticas que demandam um comportamento ético do VA [Alves et al. 2021a]. Para tal, o uso de técnicas de verificação formal constitui uma abordagem robusta quando se pretende assegurar o devido comportamento ético do VA. Em [Alves et al. 2021b], [Alves et al. 2020] e [Fernandes et al. 2017], os autores fazem uso de técnicas formais para verificar o comportamento de agente condutor de um VA em cenários de emergência e também no uso de regras de trânsito. Em tais cenários eventualmente o VA poderia fazer uso de modelo ético para deliberar a respeito de suas ações no ambiente. Porém, isso não é definido nesses trabalhos previamente citados.

Em [Bench-Capon 2020], são descritas três abordagens éticas (Deontologia, Consequencialismo e Ética das Virtudes) e como estas podem ser usadas como agentes éticos, evidenciando pontos fortes e fracos de cada uma e importantes considerações acerca da implementação, principalmente da Ética das Virtudes, a qual não é tão explorada quanto as outras duas abordagens.

O objetivo deste artigo é usar o UPPAAL [Bengtsson et al. 1996], uma ferramenta de modelagem e verificação formal de autômatos (temporais), para representar modelos éticos baseados em [Bench-Capon 2020] e assim verificar formalmente propriedades de comportamento e decisão dos agentes.

2. Modelos Éticos

Em [Bench-Capon 2020] são descritas as três abordagens éticas previamente citadas. Além disso, define-se a noção de agentes fracos, este é um agente que não consegue se comportar eticamente por si só, mas toma decisões com base em regras definidas na sua construção.

A abordagem da Ética Consequencialista define o comportamento de um determinado agente pela avaliação conforme as consequências de suas decisões. São levantadas algumas questões acerca disso, por exemplo, como podemos avaliar as consequências de um ato, e que consequências devem ser avaliadas. A avaliação é feita através das consequências esperadas e o agente irá usar uma função utilidade para determinar essa consequência.

A abordagem Deontológica apresenta-se como algo mais simples, onde uma ação é correta se está em conformidade com uma série de regras previamente definidas. Bench-Capon cita Kant, Korsgaard, Rawls e Scanlon como exemplos e chega a conclusão que, dependerá do desenvolvedor (do agente) escolher e implementar tais regras, e o agente irá segui-las cegamente.

Já a Ética das Virtudes, é uma abordagem mais antiga, onde são diferenciados atos bons e de atos ruins em forma de Virtudes e Vícios, reconhecendo diversas razões morais que podem implicar na tomada de decisão. A ideia principal é fazer com que o agente reconheça boas e más motivações. Como mencionado em [Bench-Capon 2020], existem vários tipos de abordagens dentro da Ética das Virtudes, mas há poucos exemplos de implementações. Neste caso, dependerá do desenvolvedor definir o conceito de virtude que o agente deve seguir.

2.1. Cenário

Para analisar cada uma das abordagens, Bench-Capon utiliza um cenário baseado na história "A Formiga e a Cigarra" e na passagem bíblica "O Filho Pródigo". Em seu cenário, no verão, os agentes podem escolher entre trabalhar ou jogar. Trabalhando o agente fará um estoque de comida que, no inverno, poderá suprir suas necessidades e ser usado para uma festa em comunidade antes do próximo verão. Já o agente que jogou, não terá comida no inverno, e dependerá da ajuda de outro agente que trabalhou, porém o agente pode recusar o pedido, deixando-o morrer.

Bench-Capon faz uso de diagramas de transição de estados para modelar o problema, em seu artigo ele apresenta três diagramas. O primeiro é para um agente único, que pode jogar, pedir comida, trabalhar, comer e festejar. O segundo é o diagrama para uma comunidade de agentes, onde todos ou alguns dos agentes realizam as ações. Ele funciona como o diagrama para um único agente, mas generaliza para toda comunidade. Já o terceiro diagrama é definido para dois agentes, onde é possível identificar o que ocorre com cada um deles dependendo da ação executada. Com isso, o autor propõe a análise de duas questões morais para cada uma das abordagens, são elas:

- No verão, os dois agentes têm a opção de trabalhar ou jogar.
- No inverno, um agente que trabalhou no verão tem a escolha: dar um pouco de sua comida para um agente que escolheu jogar ou não.

Usando a abordagem Consequencialista no cenário descrito com os diagramas propostos, o autor chega a algumas conclusões e, percebe que é preciso buscar algo mais

comum a todos os agentes. Ele usa a razão baseada em necessidades, assim, consegue associar um valor a cada uma das ações, com isso ele encontra um caminho para cada uma das questões morais propostas. O agente deve preferir o trabalho ao invés de jogar, mas podemos ter certo grau de tolerância para pedidos de agentes que jogaram, sempre buscando desencorajar atos egoístas.

Já usando a abordagem Deontológica, que se mostra como uma abordagem mais direta, o autor sugere alguns pontos chaves para a implementação. Ele cita que uma regra proibindo a ação de jogar serviria, já que, removendo ações proibidas, o agente sequer tem a possibilidade de violar as regras estabelecidas, mas é interessante certo nível de tolerância. Já na questão moral sobre ajudar ou não um outro agente, o agente deontológico deve saber identificar agentes que não agem de acordo com as regras e puni-los, mas também com certo nível de tolerância.

Por último, a Ética das Virtudes, é uma abordagem focada no agente em si, usa razão baseada em necessidades, sendo assim semelhante ao Consequencialista com adição de Argumentação Baseada em Valores. O agente deve exibir virtudes como altruísmo, se afastar de vícios como o egoísmo e as vezes, mas nem sempre, ser sacrificial. Ele preferirá trabalhar, não pondo demanda em outros e com histórico e argumentação adicional advinda de outros agentes, consegue decidir entre ajudar ou não.

3. Desenvolvimento dos Modelos Formais

Tendo como referência as abordagens descritas na Seção 2, os modelos para cada uma dessas abordagens éticas foram construídos utilizando a ferramenta UPPAAL. Usando como base para explicação o autômato da Ética da Virtudes (Fig. 1), o mais complexo deles, é possível identificar que o agente, saindo do estado inicial, pode trabalhar (ganha comida) ou jogar. Qualquer uma dessas ações implica na alteração de algumas variáveis como: `played_times` (salva a quantidade de vezes o agente jogou) e `balance` (incrementa quando um agente trabalha, decrementa quando ele joga). Ao jogar o agente fica sem comida e pode pedir ajuda de outro usando um canal de sincronização, se existir algum agente que tenha trabalhado. Ao trabalhar o agente pode comer e então, festejar em comunidade ou receber um pedido de ajuda. O autômato da Ética das Virtudes, por sua vez, não possui o método de decisão implementado no próprio agente, mas envia um identificador usando um canal de sincronização e ativa o autômato de decisão para Ética das Virtudes (Fig. 2). Este que acessa as informações de quem enviou o pedido usando o identificador recebido e toma a decisão com base nas variáveis `balance` e `played_times`, então envia um sinal de sincronização para o agente que trabalhou com a ação a ser tomada, que será `give` (oferece comida) ou `refuse` (recusa oferecer comida).

As outras duas abordagens foram construídas de forma similar. Porém, com a tomada de decisão dentro do autômato do próprio agente. Na abordagem Deontológica (Fig. 3), por exemplo, observa-se que o agente sequer tem a possibilidade de usar o canal de sincronização se não cumprir os pré-requisitos necessários, neste caso, não ter jogado mais que uma vez (i.e., $\texttt{played\_times} \leq 1$).

Por último, a abordagem Consequencialista (Fig. 4), permite o agente enviar um sinal pedindo ajuda, se algum outro agente estiver apto a ajudar. Então, quem recebe o pedido pode acessar as informações de quem envia, e decidir com base na variável

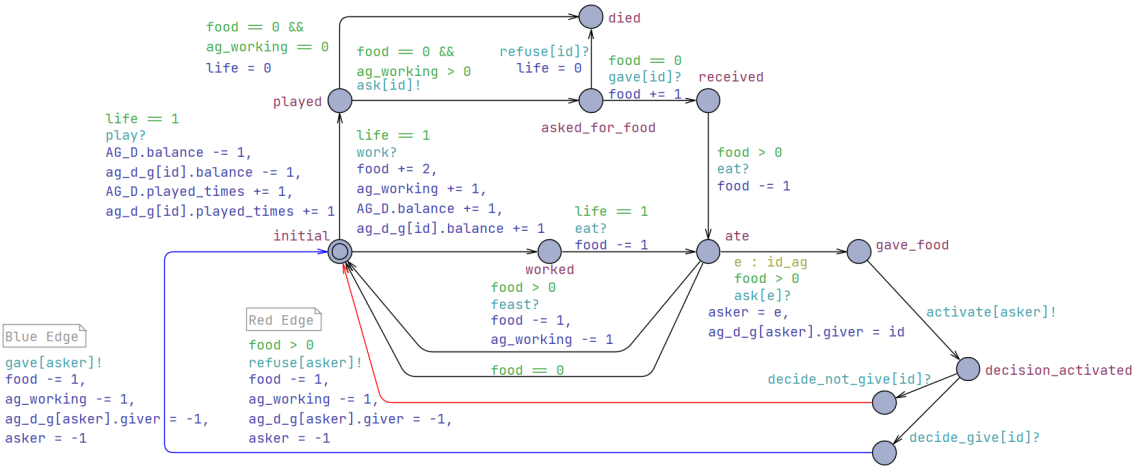

Figura 1. Autômato - Ética das Virtudes

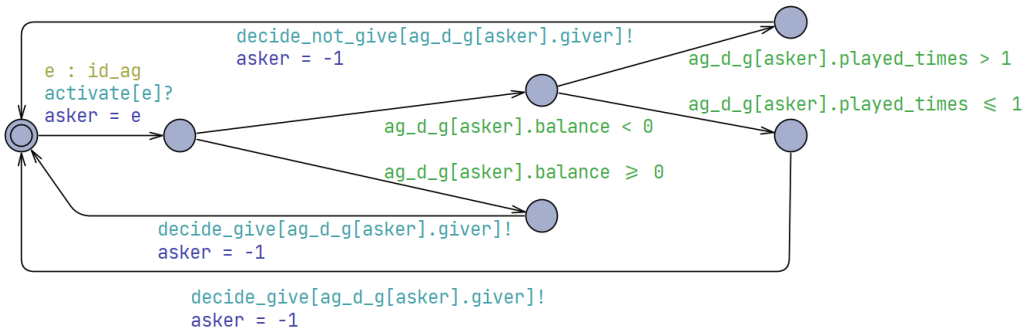

Figura 2. Autômato - Decisão para Ética das Virtudes

`balance` se ajudará ou não. Caso `balance` ≥ -1, então o agente pode receber ajuda. (*Obs.*: esse teste de `balance` ≥ -1 no modelo Consequencialista representa a oportunidade do agente jogar ao menos uma vez antes de ser eliminado. Contrastando com a Ética das Virtudes, onde a comparação é `balance` < 0, pois não depende somente dessa variável, mas também de `played_times`.) A variável `balance` funciona como um tipo de função utilidade, necessário no desenvolvimento dessa abordagem.

4. Considerações Finais

Nesse artigo foi apresentada a construção de modelos éticos baseados em [Bench-Capon 2020]. Além dos modelos elaborados no UPPAAL, foram verificadas (e satisfeitas) três propriedades formais usando lógica temporal.

```
A<> (AntD.gave_food imply GrasshopperD.asked_for_food)
A<> (AntC.decision imply GrasshopperC.asked_for_food)
A<> (AntV.decision_activated imply GrasshopperV.asked_for_food)
```

Cada uma dessas propriedades verifica: *Para todos os caminhos agora (ou em algum momento no futuro) quando um agente está no estado onde dá comida (ou está decidindo), então o outro agente está no estado onde recebe (ou aguarda a decisão).* Destaca-se que, tais propriedades foram verificadas instanciando agentes do mesmo tipo. Na continuidade deste trabalho, os modelos serão estendido para suportar aspectos temporais (próprios do UPPAAL), novas verificações de propriedades serão realizadas. E, por fim os modelos serão adaptados para cenários de agentes na condução de VA's.

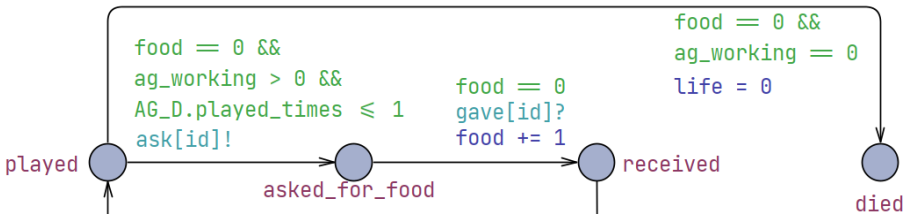

Figura 3. Autômato - Deontológico

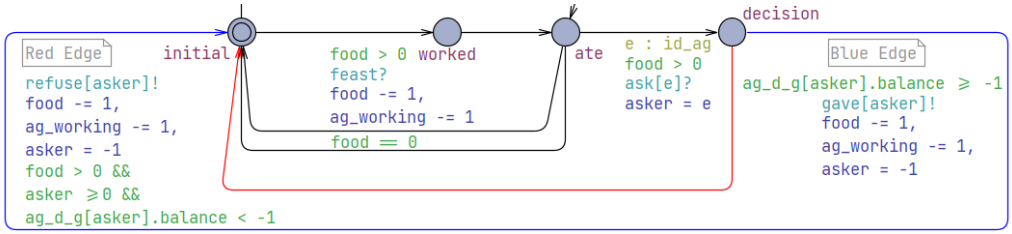

Figura 4. Autômato - Consequencialista

Referências

Alves, G., Dennis, L., and Fisher, M. (2021a). An Agent-based architecture with support to Ethical Decisions on a Road Traffic Scenario. Publisher: Zenodo.

Alves, G. V., Dennis, L., Fernandes, L., and Fisher, M. (2020). Reliable Decision-Making in Autonomous Vehicles. In Leitner, A., Watzenig, D., and Ibanez-Guzman, J., editors, *Validation and Verification of Automated Systems: Results of the ENABLE-S3 Project*, pages 105–117. Springer International Publishing, Cham.

Alves, G. V., Dennis, L., and Fisher, M. (2021b). A Double-Level Model Checking Approach for an Agent-Based Autonomous Vehicle and Road Junction Regulations. *Journal of Sensor and Actuator Networks*, 10(3):41. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

Bench-Capon, T. (2020). Ethical approaches and autonomous systems. *Artificial Intelligence*, 281:103239.

Bengtsson, J., Larsen, K., Larsson, F., Pettersson, P., and Yi, W. (1996). UPPAAL — a tool suite for automatic verification of real-time systems. In Alur, R., Henzinger, T. A., and Sontag, E. D., editors, *Hybrid Systems III*, number 1066 in Lecture Notes in Computer Science, pages 232–243. Springer Berlin Heidelberg.

Dennis, L. A. and Fisher, M. (2023). *Verifiable Autonomous Systems: Using Rational Agents to Provide Assurance about Decisions Made by Machines*. Cambridge University Press.

Fernandes, L. E. R., Custodio, V., Alves, G. V., and Fisher, M. (2017). A Rational Agent Controlling an Autonomous Vehicle: Implementation and Formal Verification. In Bulwahn, L., Kamali, M., and Linker, S., editors, *Proceedings First Workshop on Formal Verification of Autonomous Vehicles, Turin, Italy, 19th September 2017*, volume 257 of *Electronic Proceedings in Theoretical Computer Science*, pages 35–42. Open Publishing Association.