ASPI | Afarand Scholarly Publishing Institute; Turkey
ISSN: 2873-1973; Philosophical Thought. 2025;5(2):207-221. 10.58209/jpt.5.2.207

Logical Formalization of Some Counterfactual Emotions

ARTICLE INFO

Article Type
Original Research

Authors
Mashhadi Raviz F.
Department of Philosophy and Logic, Faculty of Humanities, Tarbiat Modares University, Tehran, Iran
Nabavi L.*
Department of Philosophy and Logic, Faculty of Humanities, Tarbiat Modares University, Tehran, Iran
Alizadeh M.
Department of Computer Science, School of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, Tehran, Iran

How to cite this article
Mashhadi Raviz F, Nabavi L, Alizadeh M. Logical Formalization of Some Counterfactual Emotions. Philosophical Thought. 2025;5(2):207-221.

***Correspondence**
Address: Department of Philosophy and Logic, Faculty of Humanities, Tarbiat Modares University, Jalal-Al Ahmad Street, Tehran, Iran. Postal Code: 139-14115
Phone: +98 (21) 82884643
nabavi_l@modares.ac.ir

Article History
Received: April 23, 2025
Accepted: May 28, 2025
ePublished: May 31, 2025

ABSTRACT

Counterfactual emotions, such as guilt and shame, represent a significant category of factors that influence human behavior in social and moral contexts. These emotions enable agents to evaluate their choices concerning prior experiences. In this process, they consider how they might have acted differently and then contemplate how their choices might have led to different outcomes. They endeavor to modify their behavior. This paper proposes a logical framework designed to represent counterfactual emotions within the contexts of Multi-Agent Systems formally. The proposed formalism uses a specific syntax and semantics to express the relationship between an agent's actions, its knowledge, norms, and evaluative criteria. This approach provides a systematic explanation of the concepts of guilt and shame. This work provides a foundation for the development of intelligent systems that may be able to understand and learn about the complex ethical and social norms that are a part of societies.

Keywords Counterfactual emotions; Guilt; Shame; Multi-Agent Systems

CITATION LINKS

[Covert *et al.*, 2003] Shame-proneness, guilt-proneness, and interpersonal problem solving: A social cognitive analysis; [Guiraud *et al.*, 2011] The face of emotions: A logical formalization of expressive speech acts; [Lorini & Schwarzentruber, 2011] A logic for reasoning about counterfactual emotions; [Roese & Morrison, 2009] The psychology of counterfactual thinking; [Steunebrink *et al.*, 2012] A formal model of emotion triggers: An approach for BDI agents; [Tangney *et al.*, 2007] Moral emotions and moral behavior;

نوع مقاله: پژوهشی اصیل

صوری‌سازی منطقی برخی احساسات خلاف‌واقع

فاطمه مشهدی راویز
گروه فلسفه و حکمت و منطق، دانشکده علوم انسانی، دانشگاه تربیت مدرس، تهران، ایران
لطف‌اله نبوی*
گروه فلسفه و حکمت و منطق، دانشکده علوم انسانی، دانشگاه تربیت مدرس، تهران، ایران
مجید علی‌زاده
گروه علوم کامپیوتر، دانشکده ریاضی، آمار و علوم کامپیوتر، دانشکدگان علوم، دانشگاه تهران، تهران، ایران

چکیده

احساسات خلاف‌واقع، مانند گناه و شرم، رفتار انسان را در تلاقی دنیای اجتماعی و اخلاقی شکل می‌دهند. چنین احساساتی این امکان را برای افراد فراهم می‌سازند تا با اندیشیدن به آنچه می‌توانست رخ دهد در برابر آنچه رخ داده، انتخاب‌های خود را مورد سنجش قرار دهند. سپس، با تامل در اینکه چگونه گزینه‌های دیگر می‌توانستند به نتایجی متفاوت ختم شوند، در پی اصلاح رفتار خود برمی‌آیند. در این مقاله چارچوبی منطقی برای صوری‌سازی این احساسات در حوزه هوش مصنوعی و سیستم‌های چندعاملی ارایه شده است. برای این منظور، نحو و معناشناسی طراحی شده که رفتار یک عامل را به دانش، هنجارها و معیارهای او گره می‌زند و از این راه، احساساتی چون گناه و شرم را به‌خوبی بازسازی می‌کند. این پژوهش، زمینه‌ای برای طراحی سیستم‌های هوشمندی فراهم می‌آورد که بتوانند پیچیدگی‌های انسانی و اصول اخلاقی جامعه را درک کرده و با آنها همگام شوند.

کلیدواژگان: احساسات خلاف‌واقع، گناه، شرم، سیستم‌های چندعاملی

***نویسنده مسئول:** nabavi_l@modares.ac.ir
آدرس مکاتبه: تهران، خیابان جلال آل احمد، دانشگاه تربیت مدرس، دانشکده علوم انسانی
تلفن محل کار: ۸۲۸۸۴۶۴۳ (۰۲۱)

تاریخ دریافت: ۱۴۰۴/۰۲/۰۳
تاریخ پذیرش: ۱۴۰۴/۰۳/۰۷
تاریخ انتشار: ۱۴۰۴/۰۳/۱۰

مقدمه

یکی از بزرگ‌ترین چالش‌ها در هوش مصنوعی، خلق سیستم‌های تعاملی هوشمندی است که در رفتار، عملکرد و احساساتشان تا جای ممکن به انسان‌ها شبیه باشند. به طوری که بتوان آنها را موجوداتی انسان‌گونه تصور کرد. برای رسیدن به این هدف، چنین سیستمی باید بتواند درباره احساسات خود فکر کند، شخصیت و احساسش را نشان دهد، احساسات را به انسان‌ها نسبت دهد، پیش‌بینی کند که رفتارش چه تاثیری بر دیگران می‌گذارد، و براساس آن، واکنش‌هایش را تنظیم یا اصلاح کند.

از این جهت، بررسی پدیده‌های احساسی به یکی از حوزه‌های اصلی در هوش مصنوعی تبدیل شده است تا نسل جدیدی از سیستم‌های تعاملی احساسی پدید آید. در سال‌های اخیر، محاسبات احساسی پیشرفت‌های چشمگیری داشته‌اند. برخی پژوهشگران چارچوب‌های منطقی را برای تحلیل دقیق احساسات توسعه داده‌اند و بیشتر بر روش‌های منطقی تمرکز کرده‌اند تا سیستم‌ها بتوانند حالات احساسی خود را به‌خوبی ابراز کنند.

برای نمونه، در مطالعه لورینی و شوارزنتروبر [Lorini & Schwarzentruber, 2011]، مدلی منطقی برای احساسات خلاف‌واقع ارایه شده که بر ارتباط میان انتخاب‌ها، توانایی‌ها و تاثیرات کنش‌های عامل‌ها و گروه‌ها، دانش و خواسته‌هایشان تمرکز دارد.

در پژوهش استیونبرینک و همکاران [Steunebrink *et al.*, 2012]، بخشی از یک مدل روان‌شناختی معروف درباره احساسات به صورت منطقی بیان شده است. این کار شرایط برانگیخته‌شدن احساسات را به شکل منظم و سلسله‌مراتبی بررسی کرده و نتایج آن برای هدایت تحلیل محرک‌های احساسی به کار گرفته است.

همچنین، در مطالعه گیرود و همکاران [Guiraud *et al.*, 2011]، نظریه احساسات با نظریه کنش گفتاری و منطق در هم آمیخته شده است. با بهره‌گیری از مفاهیم باور، هدف، آرمان و مسئولیت در منطق موجهات، توضیح داده شده که یک سیستم در گفت‌وگو با سیستمی دیگر چه چیزی را منتقل می‌کند.

در این مقاله، سعی بر ارایه چارچوبی منطقی برای احساسات خلاف‌واقع در سیستم‌های چندعاملی است که عامل با توجه به هنجارها و معیارهای موجود در سیستم (جامعه)، کنش‌ها و پیرو آن رویدادهای رخ داده را بررسی می‌کند. در این چارچوب به احساسات اخلاقی مهم مانند گناه و شرم پرداخته شده و ویژگی‌ها و نقش هر یک را در استدلال اخلاقی و اجتماعی نشان داده است. گناه به عنوان آگاهی سیستم از نقض یک هنجار اخلاقی تعریف می‌شود، در حالی که شرم از درک قضاوت منفی دیگران ریشه می‌گیرد.

معرفی احساسات خلاف‌واقع

در این مقاله یک چارچوب منطقی به نام احساسات خلاف‌واقع (CFE; Counterfactual Emotions) ارایه می‌دهیم که در آن می‌توان احساسات خلاف‌واقعی مانند گناه یا شرم را بیان کرد این چارچوب به گونه‌ای متفاوت از منطق معرفتی پویا و ترکیب آن با زمان نشات گرفته است. در ادامه به بررسی نحو و معناشناسی CFE می‌پردازیم.

نحوشناسی

فرض کنید $Agt = \{1, \dots, n\}$ یک مجموعه متناهی از عامل‌ها باشد، $Act = \{a_1, \dots, a_m\}$ یک مجموعه متناهی غیرتهی از کنش‌های اتمی باشد و $T = \{1, 2, \dots\}$ مجموعه‌ای برای زمان باشد. همچنین، فرض کنید $Atom$ یک مجموعه قابل شمارش از گزاره‌های اتمی باشد و $Norm$ یک زیرمجموعه متناهی از $Atom$ باشد که شامل گزاره‌های اتمی هنجاری مانند هنجارهای اخلاقی، قانونی و اجتماعی (و یا هنجارهای دیگر) است. ما از نماد α برای عناصر $Norm$ استفاده می‌کنیم؛ هر $\alpha \in Norm$ یک گزاره‌ی اتمی است که نشان‌دهنده یک هنجار خاص است. برای مثال، هنجار «به دیگران آسیب نرسانید» در $Norm$ به صورت گزاره «هیچ‌کس از کنش‌های عامل آسیب نمی‌بیند» یا «کنش‌های عامل به هیچ‌کس آسیب نمی‌رسانند» مدل‌سازی می‌شود. این گزاره وقتی هنجار رعایت شود، صادق است و در صورت نقض آن، کاذب خواهد بود.

علاوه بر این، EC یک مجموعه متناهی از معیارهای ارزیابی (Evaluation criteria) برای عامل‌ها است. ما از نماد β برای عناصر EC استفاده می‌کنیم. برای هر $i \in Agt$، فرمول β_i به این معنا است: « عامل i ویژگی β را دارا است». بنابراین، برای هر i، مجموعه $EC_i = \{\beta_i \mid \beta \in EC\}$ زیرمجموعه‌ای از $Atom$ است. به عنوان مثال، اگر «شجاعت» برابر β باشد، آنگاه β_i به این معنا است که «عامل i شجاع است» یا « عامل i ویژگی شجاعت را دارا است». β_i در صورتی صادق است که عامل i معیار شجاعت را برآورده کند، و در غیر این صورت، کاذب است.

ما مجموعه $Atom$ را با افزودن اتم‌های $P_{i,r}^{a}$ برای هر عامل $i \in Agt$، کنش $a \in Act$ و زمان $r \in T$ گسترش می‌دهیم، که در آن $P_{i,r}^{a}$ به این معنا است: «عامل i کنش a را در زمان r انجام داده است». مجموعه همه چنین اتم‌هایی را P می‌نامیم.

زبان L با استفاده از دستور زبان زیر تعریف می‌شود:

$$\varphi :: \bot \mid p \mid P_{i,r}^{a} \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\,\varphi \mid \nabla_{i,r}^{a:b}\varphi \mid K_i\varphi \mid Goal_i\varphi$$

که در آن $p \in Atom$، $i \in Agt$، $a \in Act$ و $r \in T$ است.

فرمول $\Box\,\varphi$ به این معنا است که «φ ضرورتاً برقرار است» و ($\Diamond\varphi = \neg\,\Box\,\neg\varphi$). فرمول $\nabla_{i,r}^{a:b}\varphi$ به این معنا است که «ممکن بود (در گذشته) که عامل i به جای کنش a در زمان r، کنش b را انتخاب کند و در نتیجه، φ صادق می‌بود». فرمول $K_i\varphi$ به این معنا است که «عامل i می‌داند که φ صادق است». فرمول $Goal_i\varphi$ به این معنا است که «عامل i می‌خواهد φ صادق باشد».

هنجارها و معیارهای ارزیابی

هنجارها $(Norm)$ مجموعه‌ای متناهی از گزاره‌های اتمی هستند که به طور دقیق هنجارهای رفتاری را مشخص می‌کنند. این هنجارها، که عامل‌ها در یک سیستم چندعاملی موظف به رعایت آنها هستند، شامل هنجارهای اخلاقی، قانونی و اجتماعی می‌شوند. هر عنصر α در مجموعه $Norm$ یک گزاره اتمی است که نشان‌دهنده یک هنجار خاص است. برای مثال، هنجار «به دیگران آسیب نرسانید» به صورت گزاره‌ای منطقی بیان می‌شود: «کنش‌های عامل به هیچ‌کس آسیب نمی‌رسانند». این گزاره در صورت رعایت هنجار صادق و در صورت نقض آن کاذب است. هنجارها تجویزی (Prescriptive) هستند؛ به این معنی که به عامل‌ها دستور می‌دهند که چه باید بکنند یا از چه باید اجتناب کنند. این هنجارها به عنوان محدودیت‌های خارجی بر رفتار عمل می‌کنند و نقض آنها به طور رسمی با مفهوم غیراخلاقی‌بودن گره خورده است. چنین تخلف‌هایی احساساتی مانند گناه را در عامل‌ها برمی‌انگیزند و از این رو، نقش مهمی در تنظیم تعاملات اجتماعی دارند.

در مقابل، معیارهای ارزیابی (EC) مجموعه‌ای متناهی از معیارها هستند که برای سنجش دستاوردها یا ویژگی های مطلوب عامل‌ها در سیستم چندعاملی طراحی شده‌اند. هر عنصر β در EC با گزاره‌ای به نام βi برای یک عامل خاص i مرتبط است، به این معنا که «عامل i ویژگی β را دارا است» یا «عامل i معیار β را برآورده کرده است». این معیارها نشان‌دهنده حالات یا خصوصیات ارزشمندی مانند فضایل (برای مثال، شجاعت یا هوش) هستند که عامل‌ها ممکن است برای دستیابی به آنها تلاش کنند یا براساس آنها مورد قضاوت قرار گیرند. معیارهای ارزیابی توصیفی (Descriptive) هستند؛ به این معنی که این معیارها رفتار را دیکته نمی‌کنند، بلکه صرفاً وضعیت یا عملکرد عامل‌ها را توصیف و ارزیابی می‌کنند. ناکامی در برآورده‌کردن این معیارها به عدم کفایت می‌انجامد که معمولاً با احساس شرم همراه است.

معرفی چند مفهوم جدید

تعریف. در این بخش، مفاهیم جدیدی را بر پایه نمادها و مفاهیم موجود معرفی می‌کنیم.

۱) انجام کنش:

$$HP_{i,r}^{a}\varphi \;:=\; P_{i,r}^{a} \wedge \Box\left(P_{i,r}^{a} \rightarrow \varphi\right)$$

این فرمول بیان می‌کند که «عامل i کنش a را در زمان r انجام داده و ضرورتاً پس از آن، φ برقرار است».

لازم به ذکر است که HP مخفف عبارت Has Performed است.

۲) غیراخلاقی (Immorality) بودن:

$$Imm_{P^a_{i,r}}\varphi \coloneqq HP^a_{i,r}\varphi \wedge \ \Box\left(HP^a_{i,r}\varphi \rightarrow \bigvee_{\{\alpha \in Norm\}} \neg\,\alpha\right)$$

این فرمول نشان‌دهنده این است که کنش a توسط عامل i در زمان r انجام شده و منجر به وضعیتی شده است که با φ توصیف می‌شود و براساس هنجارهای موجود در $Norm$، غیراخلاقی تلقی می‌گردد؛ یعنی «عامل i در زمان r، کنش a را انجام داده که ضرورتاً φ را در پی داشته است و ضرورتاً حداقل یکی از هنجارها را نقض کرده است».

۳) ناکافی (Inadequacy) بودن:

$$Inq_{P^a_{i,r}}\varphi \coloneqq HP^a_{i,r}\varphi \wedge \ \Box\left(HP^a_{i,r}\varphi \rightarrow \bigvee_{\{\beta \in EC\}} \neg\,\beta\right)$$

این فرمول بیانگر آن است که کنش a توسط عامل i در زمان r انجام شده و به وضعیتی منجر شده که با φ توصیف می‌شود و براساس معیارهای ارزیابی در EC، ناکافی است؛ به این معنا که «عامل i در زمان r، کنش a را انجام داده که ضرورتاً φ را در پی داشته است و ضرورتاً حداقل یکی از معیارهای ارزیابی را برآورده نکرده است».

۴) اجتناب‌پذیری خلاف‌واقع (Counterfactual avoidability):

$$CHA_{P^a_{i,r}}\varphi \coloneqq HP^a_{i,r}\varphi \wedge \left(\bigvee_{\{b \in Act, a \neq b\}} \nabla^{a:b}_{i,r} \neg\varphi\right)$$

این فرمول رابطه میان انتخاب و مسئولیت را نشان می‌دهد. یعنی «عامل i کنش a را در زمان r انجام داده که ضرورتاً به φ منجر شده، اما می‌توانست با انتخاب کنش متفاوت b، از وقوع φ جلوگیری کند و $\neg\varphi$ صادق باشد» (در اینجا نیز CHA مخفف عبارت Could Have Avoided است).

اکنون زبان جدید L_{CFE} را با دستور زبان زیر معرفی می‌کنیم:

$\sigma ::= P^a_{i,r} \mid \neg\sigma \mid \sigma \wedge \sigma$ (فرمول‌های کنش اتمی)

$\chi ::= \bot \mid p \mid \neg\chi \mid \chi \wedge \chi$ (فرمول‌های گزاره‌ای)

$\theta ::= \nabla^{a:b}_{i,r}\chi \mid \neg\theta \mid \theta \wedge \theta$ (فرمول‌های خلاف‌واقع)

$\psi ::= HP^a_{i,r}\chi \mid Imm_{P^a_{i,r}}\chi \mid Inq_{P^a_{i,r}}\chi \mid CHA_{P^a_{i,r}}\chi$ (فرمول‌های مفاهیم جدید)

$\varphi ::= \sigma \mid \chi \mid \theta \mid \psi \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\ \varphi \mid K_i\varphi \mid Goal_i\varphi$ (فرمول‌های زبان L_{CFE})

این زبان امکان بیان دقیق کنش‌ها، گزاره‌ها، مفاهیم جدید، دانش و اهداف عامل‌ها را فراهم می‌آورد.

معناشناسی

هر جهان ممکن، دنباله‌ای متناهی از ماتریس‌های کنش‌های عامل‌ها (یا همان ماتریس‌های تصمیم) است که تاریخچه کنش‌های انجام‌شده توسط آنها را ثبت می‌کند. این ساختار منظم به ما امکان می‌دهد تا تکامل کنش‌ها و تاثیرشان بر استدلال‌های خلاف‌واقع را به دقت بررسی کنیم.

تعریف. مجموعه Λ شامل ماتریس‌هایی است که کنش‌های ممکن عامل‌ها در سیستم را نشان می‌دهند.

هر $\lambda_{\ell_t} \in \Lambda$ یک ماتریس کنش‌های عامل‌ها است و به شکل زیر تعریف می‌شود:

۱. λ_{ℓ_t} یک ماتریس $n \times m$ است و $t \in \mathbb{T}$،

۲. هر سطر به یک عامل $i \in \mathrm{Agt}$ اختصاص دارد،

۳. هر ستون به یک کنش $a \in \mathrm{Act}$ مربوط است،

۴. در هر سطر، دقیقاً یک آرایه مقدار 1 دارد که نشان‌دهنده کنش انتخاب‌شده توسط عامل است و بقیه آرایه‌ها 0 هستند.

$$\lambda_{\ell_t} = \begin{pmatrix} k_{\lambda^1_{\ell_t} a_1} & k_{\lambda^1_{\ell_t} a_2} & \cdots & k_{\lambda^1_{\ell_t} a_m} \\ k_{\lambda^2_{\ell_t} a_1} & k_{\lambda^2_{\ell_t} a_2} & \cdots & k_{\lambda^2_{\ell_t} a_m} \\ \vdots & \vdots & \ddots & \vdots \\ k_{\lambda^n_{\ell_t} a_1} & k_{\lambda^n_{\ell_t} a_2} & \cdots & k_{\lambda^n_{\ell_t} a_m} \end{pmatrix}$$

در این ماتریس:

۱. مقدار $k_{\lambda^i_{\ell_t} a_j} \in \{0, 1\}$ است،

۲. برای هر عامل i (یعنی هر سطر)، فقط یک j وجود دارد که $k_{\lambda^i_{\ell_t} a_j} = 1$ (یعنی عامل i کنش a_j را انجام داده است)،

۳. سایر آرایه‌های سطر 0 هستند، به این معنا که عامل هیچ کنش دیگری انجام نداده است،

۴. $\lambda^i_{\ell_t}$ نشان‌دهنده سطر i در ماتریس λ_{ℓ_t} است.

مجموعه ماتریس‌های معتبر کنش‌های عامل‌ها این‌گونه تعریف می‌شود:

$$\Lambda = \{\lambda_{\ell_t} = [k_{\lambda^i_{\ell_t} a_j}]_{n \times m} \quad | k_{\lambda^i_{\ell_t} a_j} \in \{0,1\}, \quad \sum_{j=1}^{m} k_{\lambda^i_{\ell_t} a_j} = 1, \quad i \in Agt \text{ برای هر}\}$$

تعریف. جهان‌های ممکن را با ساختن دنباله‌هایی از ماتریس‌های کنش‌های عامل‌ها تعریف می‌کنیم:

۱. $C_1 = \Lambda$ (مجموعه کنش‌های آغازین عامل‌ها)،

۲. C_2 (دنباله‌هایی با طول ۲): دو ماتریس از C_1 را انتخاب می‌کنیم و دنباله‌ای با طول ۲ می سازیم، این دنباله، کنش‌های عامل‌ها را در دو گام متوالی نشان می‌دهد،

۳. C_i (دنباله‌هایی با طول i): i ماتریس از C_1 را انتخاب کرده و دنباله‌ای با طول i می‌سازیم، یا یک ماتریس از C_1 را به دنباله‌ای از C_{i-1} اضافه می‌کنیم، این دنباله، سیر کنش‌های عامل‌ها را در i گام متوالی نشان می‌دهد،

۴. مجموعه جهان‌های ممکن W:

- W شامل همه دنباله‌های متناهی از ماتریس‌های C_1 است،
- به شکل صوری:

$$W = \{w_t = \lambda_{\ell_1} \lambda_{\ell_2} \cdots \lambda_{\ell_t} \mid \lambda_{\ell_r} \in C_1, \quad 1 \le r \le t, \quad t \in \mathbb{T}\}$$

- یعنی هر جهان در W، دنباله‌ای از ماتریس‌ها است که نشان می‌دهد کنش‌های عامل‌ها چگونه پیش می‌روند.

تعریف. اگر یک جهان w_t به صورت $w_t = \lambda_{\ell_1} \lambda_{\ell_2} \cdots \lambda_{\ell_t}$ باشد، طول آن را این‌گونه تعریف می‌کنیم:

$$\mathrm{Len}(w_t) = t$$

یا اگر $w \in C_j$ باشد، آنگاه $\mathrm{Len}(w) = j$.

تعریف. CFE-قاب‌ها سه‌تایی‌هایی به صورت $F = \langle W, \mathcal{E}, \mathcal{G} \rangle$ هستند که:

۱. $W = \{w_t = \lambda_{\ell_1}\lambda_{\ell_2}\cdots\lambda_{\ell_t} \mid \lambda_{\ell_r} \in C_1, \quad 1 \le r \le t, \quad t \in \mathbb{T}\}$

۲. $\mathcal{E}: \mathrm{Agt} \to 2^{W\times W}$ یک رابطه هم‌ارزی معرفتی برای هر عامل $i \in \mathrm{Agt}$ است،

۳. $\mathcal{G}: \mathrm{Agt} \to 2^{W\times W}$ یک رابطه سریالی برای هر عامل $i \in \mathrm{Agt}$ است.

این قاب دارای دو محدودیت زیر است:

- C1: برای هر $w = \lambda_{\ell_1}\cdots\lambda_{\ell_t}$ و $w' = \lambda'_{\ell_1}\cdots\lambda'_{\ell_t}$ در W،و برای هر عامل i:

  اگر $w\mathcal{E}_i w'$، آنگاه $\lambda_t^i = \lambda_t'^i$ و برای همه $s < t$، داریم $\lambda_s = \lambda_s'$،

- C2: برای هر $w, w' \in W$ و هر عامل i:

  اگر $w\mathcal{E}_i w'$، آنگاه $\mathcal{G}_i(w) = \mathcal{G}_i(w')$.

رابطه $\mathcal{E}_i$ مجموعه جهان‌های ممکن را به دسته‌های هم‌ارزی تقسیم می‌کند. هر دسته شامل جهان‌هایی است که عامل i، با توجه به دانش خود، آنها را ممکن می‌بیند. برای مثال، اگر دو جهان تا پیش از گام t تاریخچه یکسانی از کنش‌ها داشته باشند و عامل i در این گام کنش یکسانی را در هر دو جهان انجام داده باشد، این دو جهان در یک دسته قرار می‌گیرند. این نشان‌دهنده یک موقعیت واقعی است: عوامل می‌دانند چه اتفاقی افتاده و خودشان چه کرده‌اند، اما شاید از کارهای فعلی دیگران خبر نداشته باشند. براساس شرط C1، عامل i گذشته را کاملاً به یاد دارد و از کنش فعلی خودش کاملاً آگاه است و آن را می‌داند، ولی ممکن است نداند دیگران الان چه کنشی را انتخاب می‌کنند، مثل یک بازی تیمی که شما همه لحظه‌های بازی را به خاطر دارید و حرکت خودتان را می‌دانید، اما از تصمیم‌های همزمان هم‌تیمی‌های خود مطمئن نیستید.

رابطه $\mathcal{G}_i$ اهداف یا دستاوردهای دلخواه عامل i را نشان می‌دهد و برای هر جهان اولیه w، مجموعه‌ای از «جهان‌های هدف» را مشخص می‌کند که خواسته‌ها یا هدف‌های عامل را بازتاب می‌دهند. ویژگی سریالی این اطمینان را می‌دهد که عوامل هیچ‌وقت بدون هدف نمی‌مانند. شرط C2 می‌گوید اهداف عامل باید فقط به دانسته‌های او بستگی داشته باشد، نه به چیزهای پنهانی که نمی‌تواند بفهمد. پس اگر دو جهان از نگاه عامل i (براساس $\mathcal{E}_i$) یکسان به نظر بیایند، هدف‌های او در این دو جهان باید یکی باشد.

تعریف. CFE-مدل‌ها دوتایی‌هایی به صورت $M = \langle F, V \rangle$ هستند که $V: \mathrm{Atom} \to 2^W$ تابع ارزش‌گذاری است و به هر گزاره اتمی $p \in \mathrm{Atom}$، مجموعه‌ای از جهان‌ها را نسبت می‌دهد که p در آنها صادق است و F یک CFE-قاب است.

اکنون، با درنظرگرفتن مدل $M = \langle F, V \rangle$، برای هر فرمول φ و هر جهان w در این مدل، عبارتی مانند $M, w \vDash \varphi$ به این معنا است که «فرمول φ در جهان w از مدل M صادق است». به جای عبارت «$M, w \vDash \varphi$ برقرار نیست»، می‌نویسیم: $M, w \nvDash \varphi$.

شرایط صدق برای فرمول‌ها به صورت زیر تعریف می‌شوند (برای هر $w = \lambda_{\ell_1}\cdots\lambda_{\ell_r}\cdots\lambda_{\ell_t} \in W$ و $a, b \in$ Act و $a \neq b$):

- $M, w \vDash p$ اگر و تنها اگر $w \in V(p)$،
- $M, w \vDash P_{i,r}^a$ اگر و تنها اگر $k_{\lambda_{\ell_r}^i a} = 1$،

- $M,w \vDash \neg\varphi$ اگر و تنها اگر $M,w \nvDash \varphi$،
- $M,w \vDash \varphi_1 \wedge \varphi_2$ اگر و تنها اگر $M,w \vDash \varphi_1$ و $M,w \vDash \varphi_2$،
- $M,w \vDash \Box\varphi$ اگر و تنها اگر برای هر $w' = \lambda_{\ell_1} \cdots \lambda_{\ell_{t-1}} \lambda'_{\ell_t}$ که $\lambda_{\ell_t}{}' \in C_1$، داشته باشیم $M,w' \vDash \varphi$،
- $M,w \vDash \nabla_{i,r}^{a:b}\varphi$ اگر و تنها اگر برای برخی $w' = \lambda'_{\ell_1} \cdots \lambda'_{\ell_r} \cdots \lambda'_{\ell_t}$ در W، برای $r < t$، داشته باشیم:

۱. $k_{\lambda^i_{\ell_r} a} = 1$،

۲. λ'_{ℓ_r} مشابه λ_{ℓ_r} است جز اینکه $k_{\lambda'^i_{\ell_r} b} = 1$ و $k_{\lambda'^i_{\ell_r} a} = 1$،

۳. $\lambda_{\ell_j}{}' = \lambda_{\ell_j}$ برای $j \neq r$،

و $M,w' \vDash \varphi$،

- $M,w \vDash K_i\varphi$ اگر و تنها اگر برای هر $w' \in W$ که $w\mathcal{E}_i w'$، داشته باشیم $M,w' \vDash \varphi$.
- $M,w \vDash Goal_i\varphi$ اگر و تنها اگر برای هر $w' \in W$ که $w\mathcal{G}_i w'$، داشته باشیم $M,w' \vDash \varphi$.

فرمول φ در یک CFE-مدل به نام M صادق است اگر و تنها اگر برای هر جهان w در M، رابطه $M,w \vDash \varphi$ برقرار باشد. همچنین، فرمول φ، CFE-معتبر ($\vDash \varphi$) است اگر و تنها اگر φ در تمام CFE-مدل‌ها صادق باشد.

مثال. تصور کنید در یک موقعیت جنگی قرار داریم که در آن، عامل i، یک سرباز در خط مقدم جنگ و عامل j، سرباز دیگری (یا فرمانده) است. هنجار α_1 را «کنش‌های عامل به هیچ‌کس آسیب نمی‌رسانند» و معیار ارزیابی β_1 را «شجاعت» در نظر می‌گیریم. همچنین ما دو کنش a_1 «شلیک‌کردن با اسلحه» و a_2 «فرارکردن از میدان نبرد» داریم.

در اوج درگیری، سرباز i باید تصمیم بگیرد: آیا با دشمن بجنگد و شلیک کند (a_1) یا عقب‌نشینی کند و فرار کند (a_2)؟ شلیک‌کردن ممکن است نشان‌دهنده شجاعت باشد (β_1)، اما خطر زیر پا گذاشتن هنجار «کنش‌های عامل به هیچ‌کس آسیب نمی‌رسانند» (α_1) را دارد؛ مخصوصاً اگر باعث آسیب غیرضروری یا تلفات ناخواسته شود. از طرف دیگر، فرارکردن جلوی آسیب به دیگران را می‌گیرد، ولی ممکن است به عنوان ترس تعبیر شود و شجاعت سرباز را زیر سوال ببرد. فرض کنیم φ نتیجه تصمیم سرباز i باشد. در این صورت اگر سرباز i شلیک کند و در پی آن دشمن کشته یا زخمی شود، داریم $HP_{i,r}^{a_1}\varphi$، یعنی «سرباز i در زمان r شلیک کرد و ضرورتاً نتیجه‌ای مثل کشته یا زخمی‌شدن دشمن (φ) رخ داد.» این نتیجه (φ) هنجار «کنش‌های عامل به هیچ‌کس آسیب نمی‌رسانند» (α_1) را نقض می‌کند، به عبارتی دیگر $Imm_{P_{i,r}^{a_1}}\varphi$. سرباز i می‌توانست با فرارکردن (a_2) از این نتیجه جلوگیری کند، یعنی با شلیک‌نکردن امکان داشت به کسی آسیب نرسد ($CHA_{P_{i,r}^{a_1}}\varphi$).

اگر سرباز i فرار کند و در پی آن شجاعت او زیر سوال برود، داریم $HP_{i,r}^{a_2}\varphi$، یعنی «سرباز i در زمان r فرار کرد و ضرورتاً برچسب ترسوبودن به آن خورده (φ)، که شجاعت او را زیر سوال برده است.» این نتیجه شجاعت را زیر سوال می‌برد و با معیار β_1 سازگار نیست، یعنی $\mathrm{Inq}_{P_{i,r}^{a_2}}\varphi$. سرباز i می‌توانست با شلیک‌کردن (a_1) از این نتیجه جلوگیری کند، یعنی با جنگیدن، می‌توانست از ترسو نامیده‌شدن جلوگیری کند ($CHA_{P_{i,r}^{a_2}}\varphi$).

اصول احساسات خلاف‌واقع

اصول و قواعد CFE به صورت زیر هستند: (برای همه $i \in Agt$، $a, b \in Act, (a \neq b)$ و برای هر نقطه زمانی r تا تاریخچه فعلی t).

- همه اصول و قواعد منطق گزاره‌ها
- همه اصول و قواعد منطق S5 برای $\Box$
- همه اصول و قواعد منطق S5 برای K_i
- همه اصول و قواعد منطق KD برای Goal_i
- $\bigvee_{a \in Act} P^a_{i,r}$ **(Act1)**
- $P^a_{i,r} \wedge P^b_{i,r} \rightarrow \bot$ **(Act2)**
- $P^a_{i,r} \rightarrow K_i P^a_{i,r}$ **(Act3)**
- $\nabla^{a:b}_{i,r} \varphi \rightarrow P^a_{i,r}$ **(CF1)**
- $\nabla^{a:b}_{i,r} \varphi \vee \nabla^{a:b}_{i,r} \psi \leftrightarrow \nabla^{a:b}_{i,r} (\varphi \vee \psi)$ **(CF2)**
- $\nabla^{a:b}_{i,r} (\varphi \wedge \psi) \rightarrow \nabla^{a:b}_{i,r} \varphi \wedge \nabla^{a:b}_{i,r} \psi$ **(CF3)**
- $\nabla^{a:b}_{i,r} \varphi \wedge \nabla^{a:b}_{i,r} \neg\varphi \rightarrow \bot$ **(CF4)**
- $\nabla^{a:b}_{i,r} \varphi \rightarrow K_i (\nabla^{a:b}_{i,r} \varphi)$ **(CF5)**
- $\text{Goal}_i \varphi \rightarrow K_i \text{Goal}_i \varphi$
- $\neg\text{Goal}_i \varphi \rightarrow K_i \neg\text{Goal}_i \varphi$

تعریف. یک برهان برای فرمول φ در سیستم اصل موضوعی CFE، دنباله‌ای متناهی از فرمول‌ها به شکل $\varphi_1, \dots, \varphi_n$ است که:

۱. $\varphi_n = \varphi$،

۲. هر فرمول φ_i در این دنباله یا

- نمونه‌ای از اصول موضوع CFE است یا
- از به‌کارگیری قواعد استنتاج بر روی فرمول‌های پیشین $\varphi_1, \dots, \varphi_{i-1}$ به‌دست آمده است.

اگر برای φ در سیستم CFE برهانی وجود داشته باشد، می‌نویسیم $\vdash_{\text{CFE}} \varphi$ یا به طور خلاصه $\vdash \varphi$ و می‌گوییم φ یک قضیه در CFE است، یعنی φ در CFE برهان‌پذیر است.

قضیه درستی (Soundness theorem): هر قضیه از CFE، CFE-معتبر است ($\vdash \varphi \Rightarrow \vDash \varphi$).

برهان. یک روش رایج برای اثبات قضیه درستی این است که نشان دهیم:

- هر اصل از CFE در هر CFE-مدل معتبر است.
- قواعد استنتاج، اعتبار را حفظ می‌کنند.

سپس، با استفاده از استقراء روی طول برهان‌ها، می‌توان ثابت کرد که هر قضیه در هر CFE-مدل معتبر است. برای این کار معتبربودن اصول را بررسی می‌کنیم. برای طولانی‌نشدن برهان، معتبربودن دو اصل Act3 و CF5 را نشان می‌دهیم و به بقیه موارد را به خواننده واگذار می‌کنیم.

(Act3): فرض کنید $w = \lambda_{\ell_1} \cdots \lambda_{\ell_r} \cdots \lambda_{\ell_t}$ و $M, w \vDash P^a_{i,r}$، این یعنی در جهان w در ماتریس کنش λ_{ℓ_r} عامل i کنش a را انجام داده است. فرض کنید w' جهانی است که با w رابطه معرفتی دارد یعنی $w\mathcal{E}_i w'$ و $w' = \lambda'_{\ell_1} \cdots \lambda'_{\ell_r} \cdots \lambda'_{\ell_t}$، پس با توجه به محدودیت C1 داریم:

$\lambda_s = \lambda_s'$، $s < t$ و برای همه $\lambda^i_t = \lambda_t'^i$

بنابراین، در جهان w'، کنش انجام‌شده توسط عامل i در گام r همچنان a است، پس $M, w' \vDash P^a_{i,r}$.

از آنجا که انتخاب w' دلخواه بود و این برای همه جهان‌های قابل دسترس از w از طریق $\mathcal{E}_i$ برقرار است، پس برای هر w' که $w\mathcal{E}_i w'$ داریم $M, w' \vDash P^a_{i,r}$. بنابراین $M, w \vDash \mathrm{K}_i P^a_{i,r}$. و در آخر چون M و w دلخواه هستند پس $\vDash P^a_{i,r} \to K_i P^a_{i,r}$.

(CF5): فرض کنید $w = \lambda_{\ell_1} \cdots \lambda_{\ell_r} \cdots \lambda_{\ell_t}$ و $M, w \vDash \nabla^{a:b}_{i,r} \varphi$. پس وجود دارد $w' = \lambda'_{\ell_1} \cdots \lambda'_{\ell_r} \cdots \lambda'_{\ell_t}$ در W، به طوری که برای $r < t$، داریم:

۱. $k_{\lambda^i_{\ell_r} a} = 1$،

۲. λ'_{ℓ_r} مشابه λ_{ℓ_r} است جز اینکه $k_{\lambda'^i_{\ell_r} b} = 1$ و $k_{\lambda'^i_{\ell_r} a} = 0$،

۳. $\lambda'_{\ell_j} = \lambda_{\ell_j}$ برای $j \neq r$،

و $M, w' \vDash \varphi$،

جهانی دلخواه مانند $w'' = \lambda''_{\ell_1} \cdots \lambda''_{\ell_r} \cdots \lambda''_{\ell_t}$ را در نظر می‌گیریم به طوری که $w\mathcal{E}_i w''$. با توجه به محدودیت C1 داریم: $\lambda^i_t = \lambda_t''^i$ و برای همه $s < t$، $\lambda_s = \lambda_s''$. حال این تساوی‌ها را در موارد ۱، ۲ و ۳ قرار می‌دهیم:

۱. $k_{\lambda^i_{\ell_r} a} = k_{\lambda''^i_{\ell_r} a} = 1$،

۲. λ'_{ℓ_r} مشابه $\lambda_{\ell_r} = \lambda''_{\ell_r}$ است جز اینکه $k_{\lambda'^i_{\ell_r} b} = 1$ و $k_{\lambda'^i_{\ell_r} a} = 0$،

۳. $\lambda'_{\ell_j} = \lambda_{\ell_j} = \lambda''_{\ell_j}$ برای $j \neq r$،

و $M, w' \vDash \varphi$. پس $M, w'' \vDash \nabla^{a:b}_{i,r}$. انتخاب w'' دلخواه بود و این برای همه جهان‌های قابل دسترس از w از طریق $\mathcal{E}_i$ برقرار است پس $M, w \vDash \mathrm{K}_i \nabla^{a:b}_{i,r} \varphi$. و در آخر $\vDash \nabla^{a:b}_{i,r} \varphi \to K_i(\nabla^{a:b}_{i,r} \varphi)$.

تعریف. مدل کانونی $M_c = \langle W_c, \mathcal{E}_c, \mathcal{G}_c, V_c \rangle$ به صورت زیر تعریف می‌شود:

- **جهان‌های کانونی (W_c):**

$W_c = \{ w_c = (w, \Gamma) \mid w = \lambda_{\ell_1} \cdots \lambda_{\ell_t}$ ، $\lambda_{\ell_r} \in C_1$ ، $1 \leq r \leq t$ ، Γ یک مجموعه سازگار ماکسیمال است که با w هماهنگ است $\}$

که در آن:

– هماهنگی Γ با $w = \lambda_{\ell_1} \cdots \lambda_{\ell_t}$ دو شرط زیر تعریف می‌شود:

- اگر $P^a_{i,r} \in \Gamma$، $(1 \leq r \leq t)$ $k_{\lambda^i_{\ell_r} a} = 1$، $\boldsymbol{a} \in Act$ ،$\boldsymbol{i} \in Agt$
- اگر $\neg P^a_{i,r} \in \Gamma$، $(1 \leq r \leq t)$ $k_{\lambda^i_{\ell_r} a} = 0$، $\boldsymbol{a} \in Act$ ،$\boldsymbol{i} \in Agt$

- مجموعه $\Gamma \subseteq \mathcal{L}_{CF}$ سازگار ماکسیمال است:
  - سازگاری: اگر $\varphi \in \Gamma$ آنگاه $\neg\varphi \notin \Gamma$،
  - ماکسیمال‌بودن: برای هر $\varphi \in \mathcal{L}_{CF}$، یا $\varphi \in \Gamma$ یا $\neg\varphi \in \Gamma$.
- برای هر جهان $w = \lambda_{\ell_1} \cdots \lambda_{\ell_t}$ تعریف می‌کنیم:

$$\Gamma_w = \{P^a_{i,r} \mid k_{\lambda^i_{\ell_r} a} = 1, 1 \le r \le t, \boldsymbol{i} \in Agt, \boldsymbol{a} \in Act\} \cup \{\neg P^a_{i,r} \mid k_{\lambda^i_{\ell_r} a} = 0, 1 \le r \le t, \boldsymbol{i} \in Agt, \boldsymbol{a} \in Act\}$$

- از آنجا که λ_{ℓ_r} دقیقاً یک کنش به هر عامل نسبت می‌دهد، Γ_w سازگار است و می‌تواند به یک مجموعه سازگار ماکسیمال Γ گسترش یابد به طوری که $\Gamma_w \subseteq \Gamma$.

- **روابط کانونی:**
  - رابطه معرفتی $(\mathcal{E}_c)$: برای هر $\boldsymbol{i} \in Agt$ و $w = \lambda_{\ell_1} \cdots \lambda_{\ell_t}$ و $w' = \lambda'_{\ell_1} \cdots \lambda'_{\ell_t}$، $(w, \Gamma)\mathcal{E}_{c,i}(w', \Gamma')$ اگر:
    - $\mathrm{Len}(w) = \mathrm{Len}(w')$،
    - $\lambda_{\ell_s} = \lambda_{\ell_s}{}'$ $(1 \le s < t)$،
    - $\lambda^i_{\ell_t} = \lambda_{\ell_t}{}'^i$،
    - برای هر φ، اگر $K_i\varphi \in \Gamma$ آنگاه $\varphi \in \Gamma'$.
  - رابطه هدف $(\mathcal{G}_c)$: برای هر $\boldsymbol{i} \in Agt$ و $w = \lambda_{\ell_1} \cdots \lambda_{\ell_t}$ و $w' = \lambda'_{\ell_1} \cdots \lambda'_{\ell_t}$، $(w, \Gamma)\mathcal{E}_{c,i}(w', \Gamma')$ اگر:
    - برای هر فرمول φ، اگر $Goal_i\varphi \in \Gamma$ آنگاه $\varphi \in \Gamma'$.
    - برای هر (w, Γ) حداقل یک (w', Γ') وجود دارد به طوری که $(w, \Gamma)\mathcal{G}_{c,i}(w', \Gamma')$.
    - اگر $(w, \Gamma)\mathcal{E}_{c,i}(w', \Gamma')$ آنگاه $\mathcal{G}_{c,i}(w, \Gamma) = \mathcal{G}_{c,i}(w', \Gamma')$.

- **ارزش‌گذاری (V_c):**
  - برای $p \in Atom \setminus \{P\}$، $V_c(p) = \{(w, \Gamma) \in W_c \mid p \in \Gamma\}$،
  - برای $P^a_{i,r}$، هماهنگ‌بودن Γ و w نتیجه می‌دهد: $P^a_{i,r} \in \Gamma$ اگر و تنها اگر $k_{\lambda^i_{\ell_r} a} = 1$.

لم صدق (Truth Lemma): برای هر فرمول φ و هر $(w, \Gamma) \in W_c$ داریم:

$$M_c, (w, \Gamma) \vDash \varphi \quad \text{اگر و تنها اگر} \quad \varphi \in \Gamma$$

برهان. این لم را با استفاده از استقراء روی پیچیدگی فرمول φ اثبات می‌کنیم: (برای طولانی‌نشدن برهان، این لم را برای فرمول $\varphi = \nabla^{a:b}_{i,r}\psi$ و برای قسمت «تنها اگر» اثبات می‌کنیم و به بقیه موارد را به خواننده واگذار می‌کنیم).

بنا بر شرایط صدق $M_c, (w, \Gamma) \vDash \nabla_{i,r}^{a:b}\psi$ اگر و تنها اگر وجود داشته باشد $(w', \Gamma') \in W_c$ به طوری که:

$w' = \lambda'_{\ell_1} \cdots \lambda'_{\ell_r} \cdots \lambda'_{\ell_t}$ در W، به طوری که برای $r < t$، داریم:

۱. $(w = \lambda_{\ell_1} \cdots \lambda_{\ell_r} \cdots \lambda_{\ell_t})$، $k_{\lambda^i_{\ell_r} a} = 1$

۲. λ'_{ℓ_r} مشابه λ_{ℓ_r} است جز اینکه $k_{\lambda'^i_{\ell_r} b} = 1$ و $k_{\lambda'^i_{\ell_r} a} = 0$،

۳. $j \neq r$ برای $\lambda'_{\ell_j} = \lambda_{\ell_j}$،

و $M_c, (w', \Gamma') \vDash \varphi$.

فرض کنید $\nabla_{i,r}^{a:b}\psi \in \Gamma$، که در آن Γ یک مجموعه سازگار ماکسیمال و هماهنگ با $w = \lambda_{\ell_1} \cdots \lambda_{\ell_r} \cdots \lambda_{\ell_t}$ است. هدف ما این است که نشان دهیم یک جهان $(w', \Gamma') \in W_c$ وجود دارد که شرایط فوق را برآورده می‌کند به طوری که $M_c, (w', \Gamma') \vDash \psi$.

از آنجا که $\nabla_{i,r}^{a:b}\psi \in \Gamma$ و Γ با w هماهنگ است و براساس CF1، داریم: $k_{\lambda^i_{\ell_r} a} = 1$.

اکنون یک $w' = \lambda_{\ell_1}' \cdots \lambda_{\ell_r}' \cdots \lambda_{\ell_t}'$ معرفی می‌کنیم که λ'_{ℓ_r} مشابه λ_{ℓ_r} است جز اینکه کنش عامل i تغییر می‌کند. به عبارتی $k_{\lambda_{\ell_r}, i_a} = 0$ و $k_{\lambda_{\ell_r}, i_b} = 1$. و همچنین برای هر $j \neq r$ داریم $\lambda_{\ell_j}' = \lambda_{\ell_j}$. در واقع w' جهانی را نشان می‌دهد که در آن، در گام r عامل i به جای کنش a کنش b را انتخاب کرده است و سایر موارد بدون تغییر باقی مانده‌اند.

ما به یک مجموعه‌ی سازگار ماکسیمال مانند Γ' که با w' هماهنگ باشد و $\psi \in \Gamma'$ نیاز داریم. مجموعه $S = \{\psi\} \cup \Gamma_{w'}$ را در نظر می‌گیریم. S سازگار است. حال S را به مجموعه‌ی سازگار ماکسیمالی مانند Γ' که با w' هماهنگ است گسترش می‌دهیم. چون $\psi \in \Gamma'$ با توجه به فرض استقرا داریم $M_c, (w', \Gamma') \vDash \psi$.

اکنون (w', Γ') همه شرایط معنایی را برآورده می‌کند: در گام r عامل i به جای کنش a می‌توانست کنش b را انتخاب کند و در جهان حاصل ψ صادق است، بنابراین $M_c, (w, \Gamma) \vDash \nabla_{i,r}^{a:b}\psi$.

قضیه تمامیت (Completeness theorem): هر فرمول CFE-معتبر، در CFE برهان‌پذیر است ($\vDash \varphi \Rightarrow \vdash \varphi$).

برهان. فرض کنید φ معتبر است ($\vDash \varphi$) اما برهان‌پذیر نیست ($\vdash \varphi$). آنگاه مجموعه $\{\neg\varphi\}$ سازگار است و می‌تواند به یک مجموعه سازگار ماکسیمال Γ هماهنگ با w گسترش یابد. اما بنا بر لم صدق $M_c, (w, \Gamma) \vDash \neg\varphi$، در نتیجه $M_c, (w, \Gamma) \nvDash \varphi$ و این با معتبربودن φ در تناقض است. پس φ برهان‌پذیر است.

احساسات خلاف‌واقع

احساسات خلاف‌واقع، نظیر گناه و شرم، از جمله پدیده‌های روان‌شناختی برجسته‌ای هستند که فرد را به تامل در این موضوع وا می‌دارند که اگر در گذشته انتخاب‌هایی متفاوت گرفته شده بود، شاید سیر رویدادها به گونه‌ای دیگر رقم می‌خورد. این احساسات جایگاه مهمی در زندگی ما دارند و بر جنبه‌های گوناگونی، از رفتارهای اجتماعی و تصمیم‌گیری‌های شخصی گرفته تا عملکرد عامل‌های هوشمند در سیستم‌های چندعاملی، تاثیر می‌گذارند. این شیوه تفکر به ما این امکان را می‌بخشد که از تجربیات خویش، حتی اگر دشوار یا ناخوشایند باشند، آموخته‌هایی کسب کنیم و با برانگیختن واکنش‌های عاطفی، رفتار خود را اصلاح و ارتقا دهیم. در حقیقت، توانایی بازسازی ذهنی سناریوهای «چه می‌شد اگر»، بستری برای ظهور احساساتی چون گناه و شرم

فراهم می‌سازد؛ احساساتی که برای رشد اخلاقی و تقویت توانایی استدلال درباره مسایل اخلاقی نقش حیاتی ایفا می‌کنند [Roese & Morrison, 2009].

در چارچوب سیستم‌های چندعاملی، این احساسات به شکلی متمایز جلوه‌گر می‌شوند. به عنوان مثال، در یک بازی رقابتی، یک عامل مصنوعی باید بتواند پشیمانی ناشی از یک حرکت اشتباه یا آسودگی ناشی از اجتناب از یک خطا را درک کند تا استراتژی‌های خود را بهبود دهد. در تعاملات انسانی نیز، فهم این احساسات به عامل کمک می‌کند تا واکنش‌های احساسی کاربران را دقیق‌تر پیش‌بینی کرده و پاسخ‌هایی همدلانه‌تر ارایه دهد. به عنوان نمونه، یک سیستم مشاوره مصنوعی باید بتواند پشیمانی بیمار از تصمیم‌های گذشته را تشخیص دهد و با ارایه راهکارهای مناسب، به او کمک کند تا با این احساس کنار بیاید. این احساسات ما را قادر می‌سازند تا سیستم‌هایی طراحی کنیم که رفتار عامل‌های آنها با ارزش‌ها و اصول اخلاقی انسانی همسو باشد.

گناه

احساس گناه در گروه احساسات خودآگاه قرار می‌گیرد و بر نقش کلیدی خودآگاهی در پدیدآمدن آن تاکید شده است. برخلاف احساسات ساده که به طور مستقیم از محرک‌ها سرچشمه می‌گیرند، احساسات خودآگاه به این نیاز دارند که فرد یا عامل خود را به عنوان موضوعی برای ارزیابی بشناسد. این خودنگری به عامل‌ها اجازه می‌دهد تا به استدلال اخلاقی بپردازند و در فضاهای اجتماعی که رفتارها پیامدهای مهمی به دنبال دارند، راه خود را پیدا کنند. توانایی نگاه انتقادی به خود، بر درک عامل‌ها از جایگاهشان در سیستم و ارتباطشان با دیگران اثر می‌گذارد. در سیستم‌های چندعاملی، عامل‌هایی که این نوع احساسات را دارند، می‌توانند رفتارشان را با هنجارها و انتظارات گروه هماهنگ کنند و به این ترتیب، هم عملکرد اخلاقی‌شان را بهبود ببخشند و هم به پایداری سیستم کمک کنند [Tangney *et al.*, 2007].

احساس گناه یک حس شخصی و درونی است که وقتی عامل‌ها متوجه تناقض بین رفتار خود و هنجارهای اخلاقی‌شان می‌شوند، پدیدار می‌شود. عامل‌ها معمولاً وقتی احساس گناه می‌کنند که ببینند رفتارشان به کسی آسیب زده یا از هنجارهای پذیرفته‌شده دور شده است. این احساس مثل یک راهنمای درونی عمل می‌کند که آنها را به تامل در خود و جبران اشتباهاتشان تشویق می‌کند. در سیستم‌های چندعاملی، این حس اهمیت زیادی دارد، چون می‌تواند عامل‌ها را به سمت رفتارهایی هدایت کند که همکاری و انسجام اجتماعی را تقویت می‌کند [Tangney *et al.*, 2007].

گنجاندن احساس گناه در طراحی سیستم‌های چندعاملی، درک ژرف‌تری از انگیزه‌ها و تعاملات عامل‌ها به دست می‌دهد. شبیه‌سازی احساس گناه به عامل‌ها کمک می‌کند تا حس مسئولیت‌پذیری بیشتری پیدا کنند و به همکاری بیشتری روی بیاورند. این توانایی احساسی، نه‌تنها عملکرد هر عامل را بهتر می‌کند، بلکه پایه‌های هنجاری حاکم بر سیستم چندعاملی را نیز استوارتر می‌سازد.

تعریف. به طور کلی، احساس گناه به عنوان پذیرش درونی مسئولیت در قبال نقض هنجارها یا واردآوردن آسیب به دیگران تعریف می‌شود. در این چارچوب منطقی، احساس گناه هنگامی بروز می‌کند که یک عامل تشخیص دهد کنش او پیامدی را به دنبال داشته که هنجارهای مشخصی را زیر پا گذاشته است و بداند که می‌توانست با انتخاب کنشی متفاوت، از وقوع آن پیامد جلوگیری کند. این احساس بر پایه سه مولفه اساسی زیر استوار است:

۱. انجام کنش: عامل i کنش a را در زمان r انجام داده که به پیامد φ منجر شده است. $HP_{i,r}^{a}\varphi$

۲. آگاهی از غیراخلاقی‌بودن: عامل i می‌داند که پیامد φ با نقض برخی هنجارها همراه بوده است.

$$K_i\left(Imm_{P_{i,r}^a}\varphi\right)$$

۳. آگاهی از امکان اجتناب: عامل i می‌داند که می‌توانست با انتخاب کنشی دیگر، از بروز φ پیشگیری کند. $K_i\left(CHA_{P_{i,r}^a}\varphi\right)$

بر این اساس، احساس گناه به صورت زیر تعریف می‌شود:

$$Guilt_i\varphi \coloneqq K_i\left(HP_{i,r}^a\varphi \wedge Imm_{P_{i,r}^a}\varphi \wedge CHA_{P_{i,r}^a}\varphi\right)$$

شرم

شرم نیز، به عنوان یک احساس خودآگاه مهم دیگر، نقش بسزایی در رفتار عامل‌ها در سیستم‌های چندعاملی ایفا می‌کند. برخلاف گناه که بر اعمال و پیامدهای آنها متمرکز است، شرم (که اغلب احساسی «عمومی» تلقی می‌شود) از آگاهی فرد نسبت به نگاه و قضاوت دیگران درباره رفتارش سرچشمه می‌گیرد. این احساس، بازتاب‌دهنده درک عامل از قضاوت‌های اجتماعی است و گاه به احساس بی‌کفایتی یا بی‌ارزشی منجر می‌شود. در سیستم‌های چندعاملی، شرم بر تصمیم‌گیری و رفتار عامل‌ها تاثیر می‌گذارد و اغلب آنها را به رعایت هنجارهای گروهی سوق می‌دهد تا از سرزنش یا طردشدن در امان بمانند.

عامل‌ها ممکن است به دلیل شرم، از تعاملات اجتماعی فاصله بگیرند یا خود را پنهان کنند، زیرا می‌خواهند از نگاه منفی دیگران بگریزند. این بُعد اجتماعی شرم در سیستم‌های چندعاملی اهمیتی ویژه دارد، چرا که عامل‌ها باید در محیط‌های پیچیده اجتماعی فعالیت کنند و روابطشان را با دیگران حفظ نمایند. وقتی عامل‌ها شرم را تجربه می‌کنند، ممکن است کارهایی را در اولویت قرار دهند که پذیرش اجتماعی‌شان را بیشتر کند، حتی اگر این به معنای چشم‌پوشی از خواسته‌ها یا اولویت‌های شخصی‌شان باشد [Covert *et al.*, 2003].

تعریف. احساس شرم معمولاً هنگامی پدیدار می‌شود که یک عامل بداند که دیگران کنش‌های او را ناکافی می دانند، اما در عین حال آرزو داشته باشد بر این برداشت منفی چیره شود. این احساس براساس چهار مولفه زیر تعریف می‌گردد:

۱. انجام کنش: عامل i کنش a را در زمان r انجام داده که به پیامد φ منجر شده است. $HP_{i,r}^a\varphi$

۲. آگاهی از دانش دیگران: عامل i می‌داند که عامل j از کنش او و ناکافی‌بودن آن آگاه است.

$$K_i\left(K_j\left(HP_{i,r}^a\varphi\right) \wedge \left(K_j\left(Inq_{P_{i,r}^a}\varphi\right)\right)\right.$$

۳. آگاهی از امکان اجتناب: عامل i می‌داند که می‌توانست با انتخاب کنشی دیگر از φ جلوگیری کند.

$$K_i\left(CHA_{P_{i,r}^a}\varphi\right)$$

٤. خواست قضاوت مثبت: عامل i می‌خواهد که بر قضاوت منفی دیگران غلبه سازد.

$$Goal_i\left(K_j\left(\neg Inq_{P_{i,r}^a}\varphi\right)\right)$$

بر این اساس، احساس شرم به صورت زیر تعریف می‌شود:

$$Shame_i^j\varphi \coloneqq K_i\left[HP_{i,r}^a\varphi \wedge K_j\left(HP_{i,r}^a\varphi\right) \wedge K_j\left(Inq_{P_{i,r}^a}\varphi\right) \wedge CHA_{P_{i,r}^a}\varphi\right]$$
$$\wedge Goal_i\left(K_j\left(\neg Inq_{P_{i,r}^a}\varphi\right)\right)$$

نتیجه‌گیری

در این مقاله چارچوبی منطقی برای صوری‌سازی و استدلال پیرامون احساسات خلاف‌واقع معرفی شده که با تکیه بر مجموعه‌های متناهی گزاره‌های هنجاری (Norm) و معیارهای ارزیابی (EC)، و افزودن اتم‌های جدیدی چون $P_{i,r}^{a}$، $HP_{i,r}^{a}\varphi$، $Imm_{P_{i,r}^{a}}\varphi$، $Inq_{P_{i,r}^{a}}\varphi$ و $CHA_{P_{i,r}^{a}}\varphi$ زمینه تعریف دقیق مفاهیمی مانند «غیراخلاقی»، «ناکافی» و «اجتناب‌پذیری خلاف‌واقع» را فراهم آورده است. معناشناسی CFE براساس ساختار جهان‌های ممکن برآمده از ماتریس‌های کنش چندعاملی است؛ این ساختار امکان مدل‌سازی تاریخچه کنش‌های عامل‌ها و بررسی نتایج انتخاب‌های جایگزین را میسر می‌سازد. CFE با اصول و قواعد ویژه خود، امکان استدلال صوری و استنتاج قضایای مرتبط با احساسات خلاف‌واقع را فراهم می‌کند. در پایان، دو احساس «گناه» و «شرم» به عنوان نمونه‌های برجسته احساسات خلاف‌واقع صوری‌سازی شدند.

تشکر و قدردانی: موردی برای گزارش وجود ندارد.
تاییدیه اخلاقی: موردی برای گزارش وجود ندارد.
تعارض منافع: موردی برای گزارش وجود ندارد.
سهم نویسندگان: فاطمه مشهدی راویز (نویسنده اول)، نویسنده اصلی(۴۰٪)؛ لطف‌اله نبوی (نویسنده دوم) پژوهشگر کمکی (۳۰٪)؛ مجید علی‌زاده (نویسنده سوم)، پژوهشگر کمکی(۳۰٪)
منابع مالی: موردی برای گزارش وجود ندارد.

منابع

- Covert MV, Tangney JP, Maddux JE, Heleno NM (2003). Shame-proneness, guilt-proneness, and interpersonal problem solving: A social cognitive analysis. Journal of Social and Clinical Psychology. 22(1):1-12.
- Guiraud N, Longin D, Lorini E, Pesty S, Rivière J (2011). The face of emotions: A logical formalization of expressive speech acts. Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems. Taipei: AAMAS. p. 1031-1038.
- Lorini E, Schwarzentruber F (2011). A logic for reasoning about counterfactual emotions. Artificial Intelligence. 175(3-4):814-847.
- Roese NJ, Morrison M (2009). The psychology of counterfactual thinking. Historical Social Research. 34(2):16-26.
- Steunebrink BR, Dastani M, Meyer JC (2012). A formal model of emotion triggers: An approach for BDI agents. Synthese. 185(Suppl 1):83-129.
- Tangney JP, Stuewig J, Mashek DJ (2007). Moral emotions and moral behavior. Annual Review of Psychology. 58:345-372.