

صوری سازی شدت برخی احساسات خلاف واقع در منطق موجبات

مجید علی زاده *

دانشیار، دانشکده ریاضی، آمار و علوم کامپیوتر، دانشکده علوم، دانشگاه تهران، تهران، ایران

پست الکترونیکی: majidalizadeh@ut.ac.ir

فاطمه مشهدی راویز

دانشجوی دکتری، دانشکده علوم انسانی، گروه فلسفه و حکمت و منطق، دانشگاه تربیت مدرس، تهران، ایران

پست الکترونیکی: fmkm201@gmail.com

لطف الله نبوی

استاد، دانشکده علوم انسانی، گروه فلسفه و حکمت و منطق، دانشگاه تربیت مدرس، تهران، ایران

پست الکترونیکی: nabavi_l@modares.ac.ir

چکیده

مانند گناه یا شرم را پس از انجام یک کنش شناسایی کند یا حتی بروز دهد، اعتماد کاربر به سیستم افزایش یافته و تعاملات طبیعی تر و سازنده تر شکل می گیرد. علاوه بر این، با صوری سازی احساسات و شدت آن ها می توان به عامل های هوشمند چارچوبی برای ارزیابی پیامدهای اعمالشان از دیدگاه هنجارهای اخلاقی و معیارهای اجتماعی آموخت. این پژوهش، پلی میان تحلیل های کیفی منطقی و نیاز به معیارهای کمی در کاربردهای عملی صوری سازی احساسات در سیستم های چندعاملی ایجاد می کند.

واژه های کلیدی: احساسات خلاف واقع، درجه ی گناه، درجه ی شرم، سیستم های چندعاملی، صوری سازی.

در این مقاله سعی بر این است تا با استفاده از فرمول های صوری برخی از احساسات خلاف واقع نظیر گناه و شرم، راهی برای سنجش شدت و درجه ی این احساسات در عامل ها بیابیم. برای سنجش شدت، مفاهیم «درجه اهمیت» هنجارها و معیارها، «درجه غیراخلاقی بودن»، «درجه ناکافی بودن» (از دیدگاه خود و دیگران) و «درجه اجتناب پذیری خلاف واقع» معرفی می شوند. پس از آن مفاهیم درجه ی گناه و درجه ی شرم با توجه به تعاریف صوری آن ها، با استفاده از ترکیب وزن دار مفاهیم ذکر شده بیان می شوند. این مؤلفه ها به عامل های هوشمند امکان می دهد تا شدت یک احساس را محاسبه کرده و بر اساس آن واکنش مناسبی نشان دهند، که منجر به تعاملات انسان-ماشین باورپذیرتر و همدلانه تر می شود. اگر یک هوش مصنوعی بتواند احساساتی

۱ مقدمه

در دنیای امروز، با گسترش روزافزون سیستم های هوشمند و عامل های مصنوعی که در تعامل مستقیم

- ایجاد معیارهای کمی برای تحلیل‌های کیفی: صوری‌سازی، پلی میان تحلیل‌های کیفی منطقی احساسات و نیاز به معیارهای کمی در کاربردهای عملی ایجاد می‌کند و به طراحان سیستم‌های هوشمند اجازه می‌دهد تا شدت یک احساس را در عامل محاسبه کرده و بر اساس آن، واکنش‌های مناسب و قابل پیش‌بینی را طراحی کنند.

- پیش‌بینی و تحلیل رفتار عامل‌ها: با داشتن مدلی کمی از شدت احساسات، می‌توان رفتار عامل‌ها را در موقعیت‌های مختلف با دقت بیشتری پیش‌بینی و تحلیل کرد. این موضوع در ارزیابی عملکرد سیستم‌های پیچیده و اطمینان از عملکرد مطلوب آن‌ها در موقعیت‌های بحرانی کاربرد دارد.

تلاش برای صوری‌سازی شدت احساسات، گامی مهم در جهت توسعه هوش مصنوعی است که نه تنها هوشمند، بلکه دارای سطحی از «هوش احساسی» مصنوعی نیز باشد تا بتواند به شکلی مؤثرتر و مسئولانه‌تر در جهان انسانی عمل کند.

در این مقاله سعی داریم تا با استفاده از فرمول‌های صوری که برای برخی احساسات خلاف‌واقع^۱، مانند گناه و شرم معرفی کرده‌ایم، راهی برای بررسی و اندازه‌گیری شدت این احساسات نیز بیابیم. احساسات خلاف‌واقع پاسخ‌های احساسی هستند که از تفکر خلاف‌واقع سرچشمه می‌گیرند. این احساسات هنگامی پدید می‌آیند که افراد پیامدهای واقعی را با گزینه‌های فرضی مقایسه می‌کنند و به احساساتی منجر می‌شوند که از ماهیت این مقایسه شکل می‌گیرند.

در ابتدا به طور خلاصه مفاهیم کلیدی و فرمول‌های صوری این احساسات را بیان کرده و سپس به بررسی شدت این احساسات می‌پردازیم. از آن‌جا که هنجارها^۲ و معیارها^۳ نقش مهمی در صوری‌سازی و بررسی شدت این احساسات دارند، نخست به طور خلاصه توصیفی

با انسان‌ها هستند، ضرورت درک و شبیه‌سازی احساسات انسانی بیش از پیش احساس می‌شود. صوری‌سازی احساسات، به ویژه احساسات پیچیده‌ای مانند گناه و شرم، و همچنین سنجش شدت آن‌ها، از چند جهت برای سیستم‌های هوشمند مصنوعی حیاتی است:

- تعاملات انسان-ماشین باورپذیرتر و همدلانه‌تر: اگر یک هوش مصنوعی بتواند احساساتی مانند پشیمانی یا شرم را پس از انجام یک کنش نادرست یا ناکافی شناسایی و حتی بروز دهد (مثلاً با عذرخواهی یا اصلاح رفتار)، اعتماد کاربر به سیستم افزایش یافته و تعاملات طبیعی‌تر و سازنده‌تری شکل می‌گیرد. این موضوع به ویژه در عامل‌های هوشمند اجتماعی، دستیارهای مجازی و سیستم‌های آموزشی اهمیت دارد.

- تصمیم‌گیری اخلاقی و مبتنی بر ارزش: با صوری‌سازی احساسات و شدت آن‌ها، می‌توان به عامل‌های هوشمند چارچوبی برای ارزیابی پیامدهای اعمالشان از دیدگاه هنجارهای اخلاقی و معیارهای اجتماعی آموخت. محاسبه «درجه گناه» یا «درجه شرم» می‌تواند به عنوان یک هشدار داخلی برای عامل عمل کرده و او را به سمت تصمیماتی هدایت کند که با ارزش‌های انسانی سازگارتر است.

- بهبود همکاری در سیستم‌های چندعاملی: در محیط‌هایی که چندین عامل هوشمند با یکدیگر همکاری می‌کنند، توانایی درک حالات احساسی (یا معادل‌های صوری آن‌ها) در سایر عامل‌ها می‌تواند به هماهنگی بهتر، حل تعارضات و تصمیم‌گیری جمعی مؤثرتر منجر شود. به عنوان مثال، درکی از اینکه یک عامل به دلیل نقض یک هنجار مشترک «احساس گناه» می‌کند، می‌تواند سایر عامل‌ها را به تطبیق رفتار خود با هنجارها ترغیب نماید.

1- Counterfactual emotions

2- Norms

3- Evaluation criteria

از این دو مفهوم ارائه می دهیم.

هنجارها و معیارهای ارزیابی دو مفهوم بنیادین هستند که برای تحلیل رفتار و احساسات در سیستم های چندعاملی به کار می روند. این دو مفهوم، اگرچه از نظر ساختاری مشابه به نظر می رسند، اما نقش های متفاوتی در تحلیل و تفسیر موقعیت های اخلاقی و عملکردی ایفا می کنند.

هنجارها مجموعه ای متناهی از گزاره های اتمی هستند که قواعد رفتاری در یک جامعه یا سیستم چندعاملی را مشخص می کنند. این مجموعه را با نماد Norm نشان می دهیم. این هنجارها ممکن است اخلاقی (مانند لزوم وفای به عهد)، قانونی (مانند ممنوعیت رانندگی بدون گواهینامه)، اجتماعی (مانند رعایت نوبت در صف) یا حرفه ای (مانند رعایت اصول محرمانگی در شغل های درمانی یا حقوقی) باشند. هر عنصر α در مجموعه هنجارها، نماینده ی یک قاعده مشخص است که می توان آن را به صورت منطقی صورت بندی کرد. برای نمونه، هنجار «عدم افشای اطلاعات محرمانه» ممکن است با گزاره ای مانند «عامل هیچگاه اطلاعات خصوصی سایر عامل ها را منتشر نمی کند» بازنمایی شود. این گزاره، در صورت پایبندی به هنجار، صادق و در صورت نقض آن، کاذب خواهد بود.

معیارهای ارزیابی مجموعه ای متناهی از شاخص ها و استانداردهایی هستند که برای سنجش کیفیت عملکرد، ویژگی های فردی، یا میزان موفقیت عامل ها به کار می روند. این مجموعه را با نماد EC نشان می دهیم. هر عنصر β در مجموعه EC، به گزاره ای نظیر β_i مربوط می شود که معنای آن این است که «عامل i ویژگی β را داراست». معیارهای ارزیابی در تعیین جایگاه یا ارزش گذاری عامل ها در یک سیستم نقش مهمی دارند. این معیارها نشان دهنده حالات یا خصوصیات ارزشمندی مانند فضایل (برای مثال، شجاعت یا هوش) هستند که عامل ها ممکن است برای

دستیابی به آن ها تلاش کنند یا بر اساس آن ها مورد قضاوت قرار گیرند. برای مثال، معیار «توانایی کار تیمی» می تواند به صورت گزاره ای بیان شود که می گوید «عامل i ویژگی توانایی کار تیمی را داراست» یا «عامل i در محیط های مشارکتی عملکرد موفقی دارد». به طور خلاصه، هنجارها ناظر بر آن هستند که چه باید کرد، در حالی که معیارهای ارزیابی نشان می دهند که چه چیزهایی ارزشمند یا مطلوب تلقی می شوند.

به منظور بیان مفاهیمی که در فرمول های صورتی گناه و شرم وجود دارد، به ارائه ی تعاریف زیر می پردازیم. در این تعاریف منظور از Act مجموعه ی کنش ها است که متناهی می باشد. گزاره ی اتمی $P_{i,r}^a$ به معنای این است که «عامل i کنش a را در زمان r انجام داده است». همچنین عملگرهای موجبات با توصیفات زیر به کار می روند: فرمول $\Box \varphi$ به این معناست: « φ ضرورتاً برقرار است». فرمول $\nabla_{i,r}^{a:b} \varphi$ به این معناست: «ممکن بود (در گذشته) که عامل i به جای کنش a در زمان r ، کنش b را انتخاب کند و در نتیجه، φ صادق می بود». فرمول $K_i \varphi$ به این معناست: «عامل i می داند که φ صادق است». فرمول $Goal_i \varphi$ به این معناست: «عامل i می خواهد φ صادق باشد».

تعریف ۱. صورتی سازی غیر اخلاقی 1 بودن: منظور از این که کنش a توسط عامل i در زمان r انجام شده و منجر به وضعیتی گردیده است که با φ توصیف می شود و بر اساس هنجارهای موجود در مجموعه ی Norm غیر اخلاقی تلقی می گردد؛ این است که «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». لذا غیر اخلاقی بودن به شکل زیر صورت بندی می شود:

$$Imm_{i,r}^a \varphi := HP_{i,r}^a \varphi \wedge \Box \left(HP_{i,r}^a \varphi \rightarrow \bigvee_{\alpha \in Norm} \neg \alpha \right)$$

نکر این نکته حائز اهمیت است که ما نقض تمامی هنجارها را در شکل صورتی غیر اخلاقی بودن خلاصه

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 مؤلفه‌ی اساسی زیر استوار است [۲، ۱]:

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

- دانش از غیراخلاقی بودن (با نماد $K_i(Imm_{i,r}^a \varphi)$): عامل i می‌داند که پیامد φ با نقض برخی هنجارها همراه بوده است.

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

لذا احساس گناه به صورت زیر صورت‌بندی می‌شود.
تعریف ۴. صورتی‌سازی احساس گناه:

$$Guilt_i \varphi := K_i(HP_{i,r}^a \varphi \wedge Imm_{i,r}^a \varphi \wedge CHA_{i,r}^a \varphi).$$

به طور کلی، احساس شرم به عنوان پذیرش درونی قضاوت منفی دیگران در قبال کنش‌های خود تعریف می‌شود. در این چارچوب منطقی، احساس شرم هنگامی پدیدار می‌شود که یک عامل بداند کنش او پیامدی را به دنبال داشته که معیارهای ارزیابی مشخصی را برآورده نکرده است، بداند دیگران نیز کنش‌های او را ناکافی می‌دانند، بداند که می‌توانست با انتخاب کنشی متفاوت، از وقوع پیامد جلوگیری کند و آرزو داشته باشد بر این برداشت منفی چیره شود. این احساس بر اساس چهار مؤلفه زیر تعریف می‌گردد [۲]:

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

- آگاهی از دانش دیگران (با نماد $K_i(K_j(HP_{i,r}^a \varphi) \wedge K_j(Inq_{i,r}^a \varphi))$): عامل i می‌داند که عامل j از کنش او و ناکافی بودن آن آگاه است.

کرده‌ایم؛ یعنی نقض یک هنجار اجتماعی و یا قانونی را نیز با همین فرمول بیان می‌کنیم هرچند که می‌توان انواع هنجارها را از هم متمایز کرد، از این کار صرف‌نظر می‌کنیم و از همین فرمول برای تمامی هنجارها استفاده می‌کنیم و در ادامه‌ی کار از لفظ غیراخلاقی بودن برای نقض هر نوع هنجاری استفاده خواهیم کرد.

تعریف ۲. صورتی‌سازی ناکافی^۱ بودن:

منظور از این‌که کنش a توسط عامل i در زمان r انجام شده و به وضعیتی منجر شده که با φ توصیف می‌شود و بر اساس معیارهای ارزیابی در EC ناکافی است؛ به این معنا «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». لذا ناکافی‌بودن به شکل زیر صورت‌بندی می‌شود:

$$Inq_{i,r}^a \varphi := HP_{i,r}^a \varphi \wedge \square \left(HP_{i,r}^a \varphi \rightarrow \bigvee_{\beta \in EC} \neg \beta \right)$$

تعریف ۳. صورتی‌سازی اجتناب‌پذیری خلاف‌واقع^۲: رابطه‌ی میان انتخاب و مسئولیت یعنی «عامل i کنش a را در زمان r انجام داده که ضرورتاً φ را در پی داشته است، اما می‌توانست با انتخاب کنش متفاوت b ، از وقوع φ جلوگیری کند و $\neg \varphi$ صادق باشد». و لذا اجتناب‌پذیری خلاف‌واقع به شکل زیر صورت‌بندی می‌شود:^۳

$$CHA_{i,r}^a \varphi := HP_{i,r}^a \varphi \wedge \left(\bigvee_{b \in Act, a \neq b} \nabla_{i,r}^{a:b} \neg \varphi \right)$$

با معرفی عناصر اصلی در تعریف گناه و شرم، اکنون تعریف این دو احساس را به شکل صورتی ارائه می‌دهیم. به‌طور کلی، احساس گناه به‌عنوان پذیرش درونی مسئولیت در قبال نقض هنجارها یا وارد آوردن آسیب به دیگران تعریف می‌شود. در این چارچوب منطقی، احساس گناه هنگامی پدیدار می‌شود که یک عامل تشخیص دهد کنش او پیامدی را به دنبال داشته که

1- Inadequacy

2- Counterfactual avoidability

۳- CHA مخفف عبارت Could Have Avoided است.

برای استدلال در مورد انتخابها، تواناییها و تأثیرات کنشهای عاملها و گروهها فراهم می آورد. عملگر اصلی آن $\varphi[J]$ به این معناست که «گروه J تضمین می کند که φ صادق باشد، صرف نظر از اینکه سایر عاملها در $J \setminus \text{Agt}$ چه کاری انجام دهند». که در آن Agt مجموعه ای عاملهاست. با این حال، نسخه کامل STIT با عاملها و گروهها با مشکل جدی عدم تصمیم پذیری و عدم امکان اصل بندی متناهی مواجه است که کاربرد عملی آن را محدود می کند.

برای غلبه بر این محدودیت، نویسندگان یک قطعه نحوی خاص از STIT به نام dfSTIT را معرفی نمودند. ویژگی مهم dfSTIT محدودیت در تودرتویی عملگرهای STIT است. این محدودیت، ضمن حفظ قدرت بیانی لازم برای صورتی سازی احساسات خلاف واقع، منجر به تصمیم پذیری و قابلیت اصل بندی متناهی می شود. سپس dfSTIT را با عملگرهای معرفتی ترکیب می کنند تا منطق dfKSTIT را بسازند. این امر برای صورتی سازی بُعد ذهنی و مبتنی بر دانش احساسات ضروری است. صورتی سازی احساسات بر دو مفهوم کلیدی استوار است که در dfKSTIT تعریف می شوند:

- توانایی جلوگیری: $\text{CHP}_J\varphi := \varphi \wedge \neg[\text{Agt} \setminus J]\varphi$ این فرمول بیان می کند که φ در حال حاضر صادق است، اما گروه $J \setminus \text{Agt}$ (دیگران) آن را تضمین نکرده اند، به این معنی که انتخاب جایگزینی برای J وجود داشته که می توانست منجر به $\neg\varphi$ شود.
- مطلوبیت^۲: $\text{DES}_i\varphi := K_i\text{GOOD}_i\varphi$ یک وضعیت φ برای عامل i به صورت معرفتی تعریف می شود: عامل i می داند که φ در تمام جهان هایی که برای او خوب و مطلوب هستند (که با good_i مشخص می شوند) برقرار است.

با استفاده از این مؤلفه ها، چهار احساس خلاف واقع پیشیمانی، آسودگی، ناامیدی و وجد صورتی سازی

- آگاهی از امکان اجتناب (با نماد $(K_i(\text{CHA}_{i,r}^a\varphi))$: عامل i می داند که می تواند با انتخاب کنشی دیگر از φ جلوگیری کند.

- خواست قضاوت مثبت (با نماد $(\text{Goal}_i(K_j(\neg\text{Inq}_{i,r}^a\varphi)))$: عامل i می خواهد که بر قضاوت منفی دیگران غلبه سازد.

بنابراین شرم را می توان به صورت زیر صورتی سازی کرد.

تعریف ۵. صورتی سازی شرم:

$$\text{Shame}_i^j\varphi := K_i(\text{HP}_{i,r}^a\varphi \wedge \text{Inq}_{i,r}^a\varphi \wedge K_j(\text{HP}_{i,r}^a\varphi) \wedge K_j(\text{Inq}_{i,r}^a\varphi) \wedge \text{CHA}_{i,r}^a\varphi) \wedge \text{Goal}_i(K_j(\neg\text{Inq}_{i,r}^a\varphi)).$$

۲ مرور کارهای پیشین

در مقاله [۳] یک چارچوب منطقی برای مدل سازی، صورتی سازی و استدلال در مورد دسته ای مهم از احساسات شناختی، یعنی احساسات خلاف واقع ارائه شده است. نویسندگان با تأکید بر اهمیت روزافزون هوش مصنوعی احساسی و نیاز به ساخت عاملهای مصنوعی باورپذیر و قادر به تعاملات اجتماعی و احساسی پیچیده با انسانها، انگیزه پژوهش خود را بیان می کنند. چالش اصلی در این مسیر، توسعه یک زبان منطقی است که هم قدرت بیانی کافی برای به تصویر کشیدن دقیق مؤلفه های شناختی و ارزیابی های زیربنایی این احساسات را داشته باشد و هم از ویژگی های محاسباتی مطلوب مانند تصمیم پذیری و پیچیدگی محاسباتی قابل قبول برخوردار باشد تا بتوان از آن در سیستم های هوشمند استفاده کرد.

نویسندگان برای مواجهه با این چالش، منطق STIT^۱ را به عنوان چارچوب پایه انتخاب کردند. STIT، که ریشه در فلسفه تحلیلی دارد و به حوزه علوم کامپیوتر و سیستم های چندعاملی وارد شده است، ابزاری قدرتمند

2- Could Have Prevented (CHP)

3- Desirability (DES)

1- See-To-It-That

می‌شوند.

این مقاله یک همکاری مهم در تلاقی منطق، هوش مصنوعی و علوم شناختی محسوب می‌شود. نویسندگان موفق شده‌اند یک رویکرد منطقی دقیق و از نظر محاسباتی کنترل‌شده برای مدل‌سازی دسته‌ای پیچیده از احساسات انسانی ارائه دهند.

نوآوری اصلی در شناسایی و بهره‌برداری از قطعه $dfSTIT/dfKSTIT$ است که تعادل مناسبی بین قدرت بیانی و خواص محاسباتی برقرار می‌کند و راهکاری عملی برای غلبه بر محدودیت‌های محاسباتی منطق STIT استاندارد فراهم می‌آورد. نتایج مربوط به تصمیم‌پذیری و اصل‌بندی ارزش نظری قابل توجهی دارند و فرمول‌بندی صریح احساسات می‌تواند به عنوان یک پایه محکم برای تحقیقات در زمینه‌های هوش مصنوعی احساسی، مدل‌سازی عامل‌های باورپذیر، تعامل انسان و ربات، و حتی تحلیل‌های فلسفی و نظری در مورد احساسات و انتخاب عقلانی به کار رود. این مقاله به خوبی قدرت ابزارهای منطق صوری را در مدل‌سازی مفاهیم پیچیده روان‌شناختی و شناختی نشان می‌دهد.

مقاله [۴] تلاشی قابل توجه برای پر کردن یک شکاف مهم در تلاقی نظریه کنش‌گفتاری^۱، نظریه احساسات و منطق صوری ارائه می‌دهد. این مقاله به طور خاص بر روی کنش‌های گفتاری بیانی^۲ تمرکز دارد، دسته‌ای از کنش‌های منظوری^۳ که توسط سرل^۴ طبقه‌بندی شده‌اند اما در مقایسه با دسته‌های دیگر (مانند اظهارهای یا دستوری‌ها) کمتر مورد توجه صوری‌سازی منطقی قرار گرفته‌اند.

هدف اصلی نویسندگان، ارائه یک چارچوب منطقی یکپارچه به نام MLC^۵ است که قادر به توصیف، صوری‌سازی و استدلال در مورد این کنش‌های بیانی

باشد. هسته اصلی مقاله، معرفی منطق MLC است که یک منطق موجهات چندوجهی بر پایه چارچوب‌های BDI^۶ (باور-میل-قصد) است. این منطق با ترکیب عملگرهای موجهات برای حالات ذهنی و یک عملگر جدید برای ارتباطات، یک زبان غنی برای تحلیل تعاملات احساسی فراهم می‌کند.

مقاله با پیروی از نظریه‌های ارزیابی در روانشناسی، بر ساختار شناختی احساسات تمرکز می‌کند، یعنی ترکیب خاصی از حالات ذهنی (باورها، اهداف، آرمان‌ها، ارزیابی مسئولیت) که منجر به پیدایش یک احساس خاص می‌شوند، و سپس به صوری‌سازی دو دسته از احساسات، احساسات پایه و احساسات پیچیده، می‌پردازد.

ایده محوری مقاله این است که یک کنش گفتاری بیانی، اساساً بیان عمومی یک حالت روان‌شناختی احساس است. نویسندگان با استفاده از عملگر $Exp_{i,j,H\varphi}$ ، کنش‌های بیانی مختلف مانند ابراز خوشحالی کردن یا عذرخواهی کردن را صوری‌سازی می‌کنند. این رویکرد امکان تحلیل دقیق ساختار معنایی، پیش‌فرض‌ها و شرایط صدق هر کنش بیانی را فراهم می‌کند و نشان می‌دهد چگونه کنش‌های پیچیده‌تر از ترکیب بیان احساسات ساده‌تر و باورهای مرتبط ساخته می‌شوند.

این مقاله به طور موفقیت‌آمیزی سه حوزه متمایز منطق موجهات، نظریه کنش‌گفتاری و نظریه احساسات را در یک چارچوب صوری واحد و منسجم ترکیب می‌کند. این منطق عملگرهای جدید به‌ویژه Exp را برای مدل‌سازی ارتباطات و استفاده از عملگر Cd را برای مسئولیت ارائه می‌دهد که قدرت بیانی لازم برای مدل‌سازی جنبه‌های پیچیده تعاملات احساسی و ارتباطی، از جمله امکان عدم صداقت، را دارد. چارچوب MLC می‌تواند به عنوان مبنایی برای توسعه مشخصات عامل‌های هوشمند، تحلیل شیوه‌های

1- Speech act

2- Expressive speech act

3- Illocutionary Act

4- John Rogers Searle

5- Modal Logic of Communication

6- Belief-Desire-Intention

می‌شوند. برای نمونه، ممکن است فردی در موقعیتی خاص پابندی به قانونی اجتماعی را در اولویت قرار دهد، اما در شرایطی دیگر، همان قانون از دیدگاه او اهمیت کمتری یافته و معیار متفاوتی جایگزین آن شود. این پویایی ارزش‌گذاری، بازتابی از ماهیت انعطاف‌پذیر اولویت‌های انسانی است؛ ویژگی که سازگاری مدل با پیچیدگی‌های واقعیت اجتماعی را برآورده کرده و امکانی فراهم می‌آورد تا میزان تعهد عامل به هنجارهای اجتماعی یا معیارهای شخصی به‌دقت ارزیابی شود.

بر این اساس، دو شاخص مهم تعریف می‌شوند: «درجه غیراخلاقی بودن» و «درجه ناکافی بودن». شاخص نخست، میزان نقض هنجارها توسط پیامد یک کنش را محاسبه می‌کند. شاخص دوم فاصله پیامد از معیارهای ارزیابی عامل، مانند شایستگی یا جایگاه اجتماعی، را اندازه‌گیری می‌کند. هر دو شاخص، بینشی درباره‌ی ناسازگاری‌های درونی که در پس تجربه‌ی احساساتی چون گناه و شرم نهفته‌اند، ارائه می‌دهند و ابزارهایی برای کمی‌سازی شدت خطاهای اخلاقی یا ناکامی‌های ارزشی درک‌شده فراهم می‌سازند.

یکی دیگر از عناصر محوری در این بحث، «درجه اجتناب‌پذیری خلاف‌واقع» است؛ معیاری که نسبت کنش‌های جایگزینی را می‌سنجد که می‌توانستند از وقوع پیامد نامطلوب جلوگیری کنند. این شاخص با تأکید بر نقش عامل و مسئولیت‌پذیری، درک ما از تجربه‌های احساسی را عمیق‌تر می‌سازد. زمانی که عامل بداند گزینه‌های دیگری در دسترس داشته، گزینه‌هایی که در صورت انتخاب می‌توانستند پیامد متفاوتی رقم بزنند، بار احساسی کنش انجام‌شده افزایش می‌یابد. این معیار به‌روشنی نشان می‌دهد که درک امکان اجتناب چگونه می‌تواند بر شدت تجربه‌هایی چون گناه، پشیمانی یا شرم اثرگذار باشد، و از این طریق پیوند میان اختیار، مسئولیت و پاسخ‌های احساسی را برجسته می‌سازد.

درجه اهمیت هر هنجار و معیار ارزیابی به ترتیب با استفاده از نمادهای D_{Norm} و D_{EC} بیان می‌شود، که

ارتباطی-احساسی، و درک عمیق‌تر از نحوه عملکرد و بیان احساسات در زمینه‌های اجتماعی و ارتباطی به کار گرفته شود و مسیری برای پژوهش در این حوزه‌های میان‌رشته‌ای باز می‌کند.

۳ روش پیشنهادی

در این پژوهش روشی کمی برای اندازه‌گیری شدت احساسات معرفی می‌شود که بر اساس درجه اهمیت هنجارها و معیارهای ارزیابی محاسبه می‌گردد. این روش عددی، امکان پیش‌بینی و تحلیل دقیق‌تر رفتار عامل‌ها را در سیستم‌های چندعاملی فراهم می‌کند، چیزی که در آثار پیشین مورد توجه قرار نگرفته است.

۱.۳ شدت احساس گناه و شرم

گناه و شرم از جمله احساساتی هستند که اکثر افراد در برهه‌ای از زندگی خود با آن‌ها روبه‌رو می‌شوند، اما شدت این احساسات در میان افراد و موقعیت‌ها به طور گسترده‌ای متفاوت است. این تفاوت‌ها از تعامل پیچیده‌ای میان هنجارها، معیارها، انتظارات اجتماعی و شرایط خاصی که کنش‌ها در آن رخ می‌دهند، نشأت می‌گیرد. در این بخش این احساسات را تجزیه و تحلیل کرده و کمی‌سازی و شدت آن‌ها را مورد بررسی قرار می‌دهیم. با استفاده از تعاریف صورتی که برای این احساسات معرفی شد، هدف ما روشن ساختن درجه و شدت برای مؤلفه‌های موجود در این تعاریف صورتی است که این پاسخ‌های احساسی را شکل می‌دهند.

مفهوم «درجه اهمیت» در هسته این بحث قرار دارد که با اختصاص مقادیر عددی بین ۰ تا ۱، نشان می‌دهد یک عامل در لحظه‌ای خاص به هر هنجار یا معیار ارزیابی چه مقدار اهمیت می‌دهد. این مقادیر ثابت نیستند، بلکه در گذر زمان و در پاسخ به عوامل گوناگون همچون تجربیات زیسته‌ی عامل، زمینه‌های فرهنگی، وضعیت عاطفی، و الزامات موقعیتی، دستخوش تغییر

ممکن است مقدار اهمیت هنجار کاهش پیدا کند برای مثال $D_{\text{Norm}}(i, w', \alpha) = 0.3$.

با در نظر گرفتن این تغییرات می توان پیش بینی کرد که افراد در موقعیت های مختلف چگونه عمل می کنند. برای مثال، اگر یک هنجار با درجه اهمیت بالا نقض شود، ممکن است فرد احساس گناه بیشتری کند تا زمانی که هنجاری با اهمیت کمتر نقض شود.

درجه اهمیتی که یک عامل برای نظر یا دیدگاه عامل دیگر در یک جهان قائل است، با نماد $D_{\text{Social}}(i, j, w)$ بیان می شود که:

$$D_{\text{Social}}(i, j, w) : \text{Agt} \times \text{Agt} \times W \rightarrow (0, 1)$$

مقدار $D_{\text{Social}}(i, j, w)$ یک عدد بین ۰ و ۱ است که نشان دهنده میزان اهمیتی است که عامل i برای نظر عامل j قائل است. هر چه $D_{\text{Social}}(i, j, w)$ به صفر نزدیک تر باشد، این به این معناست که نظر عامل j برای عامل i اهمیت کمتری دارد. از سوی دیگر، هر چه $D_{\text{Social}}(i, j, w)$ به یک نزدیک تر باشد، یعنی نظر عامل j برای عامل i بسیار مهم و تعیین کننده است.

مقدار $D_{\text{Social}}(i, j, w)$ به زمینه و شرایط بستگی دارد و تحت تأثیر عوامل مختلفی تغییر می کند. برای مثال، نوع رابطه بین دو عامل، مانند دوستی، ریاست یا رقابت، می تواند تأثیر زیادی بر اهمیت نظر عامل j برای عامل i داشته باشد. در جوامعی که احترام به سلسله مراتب یا مقامات اهمیت دارد، نظر افراد ارشد ارزش بیشتری پیدا می کند. در موقعیت های خاص، مانند شرایط بحرانی یا اضطراری، نظر افراد متخصص یا رهبران اهمیت بیشتری می یابد. برای مثال، در یک موقعیت جنگی، اگر عامل j فرمانده باشد، مقدار $D_{\text{Social}}(i, j, w)$ برای یک سرباز i احتمالاً بسیار بالا خواهد بود، زیرا دستورات فرمانده در چنین شرایطی حیاتی است. اما اگر عامل j فقط یک سرباز هم رده باشد، این مقدار ممکن است کمتر باشد.

مفهوم $D_{\text{Social}}(i, j, w)$ ابزاری ریاضی است که در

هر هنجار یا معیار را به یک عدد برای هر عامل در یک جهان معین نسبت می دهد:

$$D_{\text{Norm}} : \text{Agt} \times W \times \text{Norm} \rightarrow (0, 1)$$

$$D_{\text{EC}} : \text{Agt} \times W \times \text{EC} \rightarrow (0, 1)$$

درجه اهمیت یک مفهوم اساسی در این مدل است که نشان می دهد عامل i تا چه حد یک هنجار α یا یک معیار ارزیابی β را در یک جهان w ارزشمند می داند. این مقدار با عددی بین ۰ (بی اهمیت) تا ۱ (اهمیت بسیار زیاد) بیان می شود. نکته قابل توجه این است که درجه اهمیت ثابت نیست و می تواند در طول زمان تغییر کند. به همین دلیل ارزش هر هنجار یا معیار با توجه به عامل و جهان متغیر است.

درجه اهمیت به این دلیل متغیر است که انسان ها در موقعیت های مختلف با اولویت های متفاوتی روبه رو می شوند. برای مثال، ارزشی که یک فرد برای یک هنجار قائل است، ممکن است تحت تأثیر شرایط محیطی، نیازهای فوری، یا حتی تحول در نگرش او تغییر کند. تجربه یک شکست یا موفقیت می تواند دیدگاه فرد را نسبت به معیارها و هنجارها تغییر دهد. مثلاً کسی که به دلیل عدم رعایت قوانین جریمه شده، ممکن است در آینده به آن ها اهمیت بیشتری بدهد. احساساتی مثل ترس، شادی یا غم می توانند تصمیم گیری و اولویت بندی را تحت تأثیر قرار دهند. در شرایط بحرانی، نیازهای فوری مثل بقا می توانند بر سایر معیارها و هنجارها غالب شوند.

برای مثال فرض کنید فردی در یک جامعه قانون مدار زندگی می کند و به رعایت قوانین اهمیت زیادی می دهد، مثلاً درجه اهمیت او برای هنجارهای قانونی $D_{\text{Norm}}(i, w, \alpha) = 0.9$ است. اما اگر این فرد در یک موقعیت اضطراری مانند زلزله قرار گیرد و مجبور شود برای نجات جان خود از یک منطقه ممنوعه عبور کند، که عبور نکردن از آن مکان از هنجارهای قانونی جامعه بوده است، اولویت او تغییر می کند. در این حالت،

یک عامل در یک موقعیت مشخص تعریف کرد. یکی از مسائل مهم در تحلیل رفتارها و احساسات انسانی، مانند شرم، این است که بتوانیم به شکلی دقیق و قابل‌فهم بررسی کنیم که یک کنش و پیامد آن تا چه حد با معیارهای ارزیابی شخصی یا اجتماعی یک فرد هم‌راستا یا در تعارض است. در این راستا، مفهوم «درجه ناکافی بودن» به عنوان ابزاری ریاضیاتی و منطقی معرفی می‌شود که به ما امکان می‌دهد میزان ناکامی یک عامل در برآورده کردن انتظارات یا معیارهای مطلوب را به صورت عددی و مشخص اندازه‌گیری کنیم. این مفهوم، برخلاف «درجه غیراخلاقی بودن» که بر نقض هنجارها تمرکز دارد، بیشتر به ناتوانی در دستیابی به ویژگی‌های مثبت یا معیارهای ارزیابی می‌پردازد که برای عامل اهمیت دارند. به زبان ساده، درجه ناکافی بودن نشان می‌دهد که وقتی یک تصمیم یا کنش به پیامدی خاص منجر می‌شود، این پیامد تا چه حد با معیارهایی که عامل برای خودش ارزشمند می‌داند، سازگار یا ناسازگار است.

در موقعیت‌های واقعی یا شرایط پیچیده‌تری، هر تصمیمی که می‌گیریم می‌تواند ما را به معیارهای مورد انتظارمان نزدیک‌تر یا از آن‌ها دورتر سازد. برای مثال، اگر یک سرباز در میدان جنگ تصمیم به فرار بگیرد، ممکن است این انتخاب با معیار شجاعت که برای او یا جامعه‌اش مهم است، در تضاد باشد. درجه ناکافی بودن به ما کمک می‌کند تا با نگاه به اهمیت این معیارها و میزان ناسازگاری یک پیامد با آن‌ها، شدت این ناکامی را محاسبه کنیم. این ابزار نه تنها در فهم بهتر احساساتی مانند شرم به ما یاری می‌رساند، بلکه در طراحی سیستم‌های هوشمندی که باید با پیچیدگی‌های انسانی و انتظارات اجتماعی هماهنگ شوند، کاربرد دارد.

تعریف ۷. درجه ناکافی بودن برای یک گزاره φ که پیامد کنش a توسط عامل i در زمان r است، به صورت

مدل‌سازی تعاملات اجتماعی به کار می‌رود و به ما کمک می‌کند تا درک بهتری از چگونگی تأثیر نظرات یک عامل بر تصمیم‌گیری‌ها و رفتارهای عامل دیگر داشته باشیم. یکی از چالش‌های مهم در بررسی شدت و درجه احساسی مانند گناه این است که بتوانیم به شکلی دقیق و قابل‌فهم بررسی کنیم که یک کنش و پیامد آن تا چه حد با هنجارهای یک فرد سازگار یا در تضاد است. مفهوم «درجه غیراخلاقی بودن» ابزاری صورتی و ریاضیاتی است که در این راستا به ما کمک می‌کند تا این موضوع را به صورت عددی و مشخص اندازه‌گیری کنیم. این ایده به زبان ساده، میزان غیراخلاقی بودن یک کنش و پیامد آن را با توجه به هنجارهایی که نقض می‌کند و اهمیت آن هنجارها برای فرد محاسبه می‌کند. برای مثال فرض کنید می‌خواهیم بدانیم دروغ گفتن برای یک عامل در یک موقعیت خاص چقدر غیراخلاقی است؛ این ابزار کمک می‌کند با نگاهی به هنجارهایی مثل راست‌گویی و اهمیت آن‌ها برای عامل، پاسخی روشن پیدا کنیم.

برای درک بهتر، می‌توان تصور کرد برخی از کنش‌هایی که انجام می‌دهیم، ممکن است با برخی از هنجارهای ما در تضاد باشند. درجه غیراخلاقی بودن به ما می‌گوید که وقتی کنشی خاص به پیامدی مشخص منجر می‌شود، کدام هنجارها زیر پا گذاشته می‌شوند و این هنجارها برای ما چقدر مهم هستند. این روش نه تنها به ما کمک می‌کند تا تصمیم‌هایمان را بهتر ارزیابی کنیم، بلکه در بحث‌های نظری یا موقعیت‌های پیچیده هم کاربردی است، زیرا به جای قضاوت‌های مبهم، یک معیار قابل‌مقایسه ارائه می‌دهد.

تعریف ۶. درجه غیراخلاقی بودن برای یک گزاره φ که پیامد کنش a توسط عامل i در زمان r است، به صورت زیر تعریف می‌شود:

$$D_{Imm_{P_{i,r}^a} \varphi} = \frac{\sum_{\alpha \in \text{Norm}, (\text{HP}_{i,r}^a \varphi \wedge \Box(\text{HP}_{i,r}^a \varphi \rightarrow \neg \alpha))} D_{\text{Norm}}(i, w, \alpha)}{\sum_{\alpha \in \text{Norm}} D_{\text{Norm}}(i, w, \alpha)}.$$

همان‌گونه که درجه غیراخلاقی بودن را تعریف کردیم، می‌توان درجه‌ی ناکافی بودن (یا عدم کفایت) را نیز برای

زیر تعریف می‌شود:

$$D_{\text{Inq}_{P_{i,r}^a}, \varphi} = \frac{\sum_{\beta \in \text{EC}, (\text{HP}_{i,r}^a \varphi \wedge \square(\text{HP}_{i,r}^a \varphi \rightarrow \neg \beta))} D_{\text{EC}}(i, w, \beta)}{\sum_{\beta \in \text{EC}} D_{\text{EC}}(i, w, \beta)}.$$

تعریف ۸. درجه ناکافی بودن برای یک گزاره φ که پیامد کنش a توسط عامل i در زمان r است، از دیدگاه عامل j نیز به صورت زیر تعریف می‌شود:

$$D_{j, \text{Inq}_{P_{i,r}^a}, \varphi} = \frac{\sum_{\beta \in \text{EC}, (\text{HP}_{i,r}^a \varphi \wedge \square(\text{HP}_{i,r}^a \varphi \rightarrow \neg \beta))} D_{\text{EC}}(j, w_t, \beta)}{\sum_{\beta \in \text{EC}} D_{\text{EC}}(j, w_t, \beta)}.$$

در تعریف ۸، مفهوم «درجه ناکافی بودن» از دیدگاه عامل دیگری، یعنی عامل j ، مورد توجه قرار گرفته است. در بسیاری از موقعیت‌های اجتماعی و اخلاقی، احساساتی چون شرم، سرزنش یا عدم پذیرش، نه تنها به ارزیابی درونی عامل از خود، بلکه به برداشت و داوری دیگران نسبت به کنش و پیامد آن وابسته‌اند. از این‌رو، سنجش ناکافی بودن از منظر سایر عامل‌ها امری ضروری در تحلیل اینگونه پدیده‌هاست.

در زندگی روزمره و حتی در تحلیل‌های پیچیده‌تر، اغلب با این پرسش روبه‌رو می‌شویم که یک فرد یا عامل تا چه حد می‌توانست با انتخاب‌های متفاوت، از رخ دادن یک نتیجه ناخواسته جلوگیری کند. مفهوم «درجه اجتناب‌پذیری خلاف واقع» ابزاری ریاضیاتی است که به ما کمک می‌کند این موضوع را به صورت عددی و روشن بررسی کنیم. به زبان ساده، این مفهوم نشان می‌دهد که اگر عامل به جای یک تصمیم خاص، راه دیگری را انتخاب می‌کرد، چقدر احتمال داشت که نتیجه‌ای نامطلوب رخ ندهد.

این مفهوم به‌ویژه در مواقعی که احساساتی مانند پشیمانی یا گناه در میان است، اهمیت پیدا می‌کند. درجه اجتناب‌پذیری خلاف واقع با محاسبه تعداد انتخاب‌های جایگزینی که می‌توانستند جلوی یک پیامد را بگیرند و مقایسه آن با کل گزینه‌های ممکن، عددی به ما می‌دهد که میزان اجتناب‌پذیر بودن آن پیامد را نشان می‌دهد. این روش نه تنها به ما در فهم بهتر تصمیم‌های گذشته

کمک می‌کند، بلکه در طراحی سیستم‌هایی که باید از تجربه‌ها یاد بگیرند و انتخاب‌های بهتری داشته باشند نیز کاربردی است.

تعریف ۹. درجه اجتناب‌پذیری خلاف واقع برای یک گزاره φ که نتیجه کنش a توسط عامل i در زمان r است، به صورت زیر تعریف می‌شود:

$$D_{\text{CHA}_{P_{i,r}^a}, \varphi} = \frac{\sum_{b \in \text{Act}, b \neq a} \text{sign}^+(\nabla_{i,r}^{a:b} \neg \varphi)}{|\text{Act}|}$$

که در آن:

$$\text{sign}^+(\psi) = \begin{cases} 1 & \text{اگر } \psi \text{ در جهان مورد نظر صادق باشد} \\ 0 & \text{در غیر این صورت} \end{cases}$$

اجتناب‌پذیری خلاف واقع مفهومی چندلایه و پیچیده است که عوامل بسیاری در شکل‌گیری آن نقش دارند؛ از جمله میزان آگاهی عامل نسبت به پیامدهای کنش‌ها، امکان‌پذیری گزینه‌های جایگزین، و توانایی عامل در انتخاب و اجرای آن گزینه‌ها. در تعریف ۹، ساده‌ترین صورت‌بندی این مفهوم ارائه شده است، تا زمینه‌ای ریاضی‌وار و قابل‌محاسبه برای تحلیل‌های پیچیده‌تر فراهم گردد. در این صورت‌بندی، درجه اجتناب‌پذیری تنها بر اساس تعداد کنش‌های جایگزینی محاسبه می‌شود که اگر به جای کنش واقعی a انجام می‌گرفتند، ممکن بود پیامد φ رخ ندهد. به بیان دیگر، این تعریف فقط بررسی می‌کند که آیا گزاره‌ی خلاف واقعی $\nabla_{i,r}^{a:b} \neg \varphi$ در مدل صادق است یا نه، و تعداد چنین مواردی را نسبت به کل کنش‌های ممکن می‌سنجد. در نتیجه، هرچه تعداد این موارد بیشتر باشد، درجه اجتناب‌پذیری بالاتر خواهد بود و نشان می‌دهد که عامل می‌توانست با انتخابی متفاوت از وقوع φ جلوگیری کند.

شایان ذکر است که این تعریف جنبه‌های دقیق‌تری چون احتمال موفقیت گزینه‌های جایگزین، هزینه‌ی تغییر کنش، یا درجه‌ی تفاوت پیامدهای حاصل از آن‌ها را در نظر نمی‌گیرد، بلکه تنها به امکان صرف جلوگیری از پیامد اشاره دارد. برای رسیدن به سنجشی دقیق‌تر،

مشابه، سرقت در صورتی که فرد آن را خطایی بزرگ بداند، درجه غیراخلاقی بودن بالایی خواهد داشت.

درجه اجتناب‌پذیری خلاف‌واقع مؤلفه دیگری است که به وجود گزینه‌های جایگزین و انتخاب نکردن آن‌ها وابسته است. برای مثال، اگر فردی با رانندگی شتاب‌زده دچار تصادف شود، در حالی که می‌دانسته کاهش سرعت می‌توانسته از حادثه جلوگیری کند، هر چه تعداد جایگزین‌های موجود بیشتر باشد، درجه اجتناب‌پذیری بالا خواهد بود.

این دو مؤلفه با توجه به وزن‌دهی شخصی فرد، که ممکن است بر اساس اولویت دادن به اخلاقی بودن کنش یا امکان پیشگیری از پیامد تعیین شود، ترکیب می‌شوند تا درجه گناه نهایی را مشخص کنند. اگر فرد به جنبه اخلاقی کنش اهمیت بیشتری دهد، وزن درجه غیراخلاقی بودن افزایش می‌یابد؛ در مقابل، اگر نتیجه را قابل اجتناب بداند، وزن درجه اجتناب‌پذیری بیشتر می‌شود.

فرض کنید یک عامل هوش مصنوعی به کاربر خود اطمینان می‌دهد که در بازه زمانی مشخصی به وظیفه‌ای پاسخ دهد، اما به دلایلی نامشخص، به‌جای انجام آن وظیفه، داده‌های ورودی را به گونه‌ای تغییر می‌دهد که نشان دهد مشغول است و این باعث نارضایتی سیستم‌های مرتبط می‌شود. اگر صداقت در گزارش‌دادن وضعیت و پایبندی به وظایف برای این ربات ارزش بالایی داشته باشد، درجه غیراخلاقی بودن رفتار آن افزایش می‌یابد. همچنین، اگر عامل آگاه باشد که با گزارش دقیق‌تر و مدیریت بهتر منابع می‌توانست از ایجاد این نارضایتی جلوگیری کند، درجه اجتناب‌پذیری پیامد منفی نیز بالا خواهد بود. در نتیجه، این وضعیت می‌تواند منجر به ثبت یک احساس گناه شود.

در مقابل، فرض کنید یک عامل هوش مصنوعی در یک سیستم امنیتی، برای محافظت از اطلاعات حساس، عمداً گزارشی ناقص ارائه می‌دهد یا داده‌ای را مخفی می‌کند تا از نفوذ یا دسترسی غیرمجاز جلوگیری کند.

می‌توان تعریف حاضر را گسترش داد و به‌جای تابع ساده‌ی $sign^+$ ، از توابعی بهره برد که جنبه‌های دیگر را نیز لحاظ می‌کنند. چنین تعمیم‌هایی، بنیانی غنی‌تر برای تحلیل مفاهیمی چون گناه، مسئولیت، و پشیمانی در سیستم‌های چندعاملی فراهم می‌سازند، به‌ویژه در شرایطی که ارزیابی کیفیت گزینه‌های جایگزین نیز به اندازه‌ی کمیت آن‌ها اهمیت دارد.

با معرفی و تعریف مفاهیم «درجه‌ی غیراخلاقی بودن»، «درجه‌ی ناکافی بودن» و «درجه‌ی اجتناب‌پذیری خلاف‌واقع» اکنون می‌توانیم به بررسی و تعریف درجه‌ی گناه و درجه‌ی شرم بپردازیم.

۲.۳ درجه‌ی گناه

احساس گناه یکی از پیچیده‌ترین و بنیادی‌ترین احساسات انسانی است که نقش محوری در شکل‌دهی به زندگی افراد ایفا می‌کند. این احساس به انسان امکان می‌دهد تا رفتارها و تصمیم‌های خود را در چارچوب هنجارهای اخلاقی و اجتماعی مورد ارزیابی قرار دهد. هنگامی که کنشی انجام می‌دهیم که از دیدگاه شخصی یا جامعه نادرست تلقی می‌شود، احساس گناه به مثابه هشدار عمل می‌کند که ما را به بازنگری در خطاها و بهبود عملکرد در آینده ترغیب می‌نماید. با این حال، برای تحلیل دقیق این احساس، نیاز به معیاری مشخص وجود دارد که شدت آن را به شکلی قابل اندازه‌گیری و قابل مقایسه نشان دهد. در این راستا، مفهوم «درجه گناه» به عنوان ابزاری صوری معرفی می‌شود که احساس گناه را بر اساس دو مؤلفه اصلی مورد سنجش قرار می‌دهد.

مؤلفه‌ی نخست درجه غیراخلاقی بودن است. این مؤلفه بر اساس اهمیت هنجارهای اخلاقی برای فرد و میزان تخطی از آن‌ها سنجیده می‌شود. برای نمونه، اگر صداقت برای فردی ارزشی والا باشد و او مرتکب دروغ‌گویی شود، این کنش به عنوان نقض جدی‌تری تلقی شده و درجه غیراخلاقی بودن آن افزایش می‌یابد. به طور

تحلیلی معرفی می‌شود که شدت این احساس را با توجه به سه مؤلفه کلیدی مورد سنجش قرار می‌دهد: درجه ناکافی بودن از دیدگاه عامل، درجه ناکافی بودن از دیدگاه دیگران، و درجه اجتناب‌پذیری خلاف واقع.

نخستین مؤلفه، یعنی درجه ناکافی بودن از دیدگاه عامل، به میزان تطابق یا عدم تطابق پیامد یک کنش با معیارهای ارزیابی اشاره دارد. این معیارها ممکن است شامل شایستگی، جایگاه اجتماعی، یا هر ویژگی دیگری باشد که فرد برای خود ارزشمند می‌داند. برای مثال، اگر فردی در یک ارائه عمومی به دلیل عدم آمادگی ضعیف عمل کند و این عملکرد را با معیارهای موجود ناسازگار ببیند، درجه ناکافی بودن بالایی را تجربه خواهد کرد. این مؤلفه نشان‌دهنده خودانگاره فرد و میزان اهمیتی است که او برای برآورده ساختن انتظارات خود قائل است. در مقابل، مؤلفه دوم، یعنی درجه ناکافی بودن از دیدگاه دیگران، به تأثیر قضاوت‌ها و انتظارات اجتماعی بر احساس شرم می‌پردازد. این مؤلفه زمانی شدت می‌یابد که فرد بداند دیگران او را به دلیل کنشش ناکافی یا نامناسب ارزیابی می‌کنند، به ویژه اگر نظرات آن افراد برای او اهمیت داشته باشد. برای نمونه، اگر فردی در جمعی مرتکب خطایی شود و احساس کند که این خطا جایگاهش را در نظر دیگران خدشه‌دار کرده، شرم بیشتری را تجربه خواهد کرد.

مؤلفه سوم، درجه اجتناب‌پذیری خلاف واقع، به نقش عامل و مسئولیت فرد در پیدایش پیامدهای نامطلوب توجه دارد. این مؤلفه بررسی می‌کند که آیا فرد می‌توانست با انتخاب‌های متفاوت از نتیجه ناخواسته جلوگیری کند یا خیر. اگر فرد معتقد باشد که با تصمیم‌گیری بهتر یا تلاش بیشتر می‌توانست از موقعیت شرم‌آور اجتناب کند، شدت شرم او افزایش می‌یابد. به عنوان مثال، اگر فردی به دلیل بی‌توجهی در کاری شکست بخورد و بداند که با دقت بیشتر می‌توانست موفق شود، این حس اجتناب‌پذیری، وزن احساسی تجربه او را سنگین‌تر می‌کند.

اگر این رفتار به عنوان یک اقدام اخلاقی موجه در نظر گرفته شود، درجه غیراخلاقی بودن پایین خواهد بود. همچنین، اگر جایگزین مطمئنی برای این اقدام وجود نداشته باشد، درجه اجتناب‌پذیری پیامدهای منفی نیز کاهش می‌یابد. در این حالت، احتمال احساس گناه در عامل بسیار کم خواهد بود.

این نمونه‌ها نشان می‌دهند که درجه گناه به شدت به زمینه کنش و دیدگاه‌های عامل وابسته است و یک رفتار مشابه، مانند دروغ‌گویی، می‌تواند بر اساس شرایط، شدت احساس متفاوتی را برانگیزد.

درجه گناه با ترکیب دو مؤلفه‌ی درجه غیراخلاقی بودن و درجه اجتناب‌پذیری خلاف واقع که هر یک دارای وزن و ضریب خاصی هستند به صورت زیر تعریف می‌شود. تعریف ۱۰. درجه احساس گناه:

$$D_{\text{Guilt}_{i,\varphi}} := \mu_1 \times D_{\text{Imm}_{i,r}^a, \varphi} + \mu_2 \times D_{\text{CHA}_{i,r}^a, \varphi}.$$

که در آن $(\mu_1 + \mu_2 = 1, \quad 0 < \mu_1, \mu_2 < 1)$.

۳.۳ درجه‌ی شرم

احساس شرم نیز یکی از احساسات پیچیده‌ی انسانی است که در تعاملات اجتماعی و فرآیند خودارزیابی افراد نقش محوری ایفا می‌کند. این احساس اغلب زمانی پدیدار می‌شود که فرد خود را در برابر معیارهای شخصی یا انتظارات اجتماعی ناکافی یا ناسازگار می‌بیند و می‌تواند تأثیری دوگانه داشته باشد: از یک سو، به عنوان محرکی برای رشد و بهبود عمل کند و از سوی دیگر، در صورت شدت یافتن بیش از حد، به تجربه‌ای مخرب و ناسالم تبدیل شود. شرم، برخلاف گناه که بیشتر بر نقض هنجارهای اخلاقی متمرکز است، به درک فرد از ناکامی‌ها یا نقص‌های خود در دستیابی به معیارهای مطلوب می‌پردازد. برای بررسی این احساس و درک دقیق شدت آن در موقعیت‌های مختلف، نیاز به چارچوبی ساختاریافته و قابل اندازه‌گیری وجود دارد. در این راستا، مفهوم «درجه شرم» به عنوان ابزاری

از دیدگاه دیگران نیز افزایش می‌یابد. حال، اگر عامل به منابع کافی مانند پایگاه داده، تاریخچه گفت‌وگوها، یا امکان پرس‌وجوی داخلی دسترسی داشته ولی از آن‌ها استفاده نکرده باشد، مثلاً به خاطر انتخاب ضعیف بین راهبردهای پاسخ‌دهی یا نادیده گرفتن نشانه‌های واضح مشکل، در آن صورت، درجه اجتناب‌پذیری نیز بالا خواهد بود.

این تحلیل نشان می‌دهد که شدت شرم به شدت به زمینه کنش، دیدگاه‌های عامل، و پویایی‌های اجتماعی وابسته است. یک رفتار یکسان، مانند اشتباه در یک وظیفه، می‌تواند بسته به اهمیت معیارها و شرایط، احساس شرم متفاوتی را برانگیزد. درجه شرم با ارائه چارچوبی ریاضیاتی و منطقی، نه تنها درک ما از این احساس پیچیده را عمیق‌تر می‌کند، بلکه ابزاری کاربردی برای مطالعه رفتارهای انسانی و طراحی سیستم‌های هوشمندی که با احساسات و انتظارات افراد هماهنگ باشند، فراهم می‌آورد. این مدل با ترکیب سه مؤلفه اصلی و انعطاف‌پذیری در وزن‌دهی به آن‌ها، راهی جامع و دقیق برای سنجش شدت شرم ارائه می‌دهد.

درجه شرم با ترکیب سه مؤلفه‌ی درجه ناکافی بودن از دیدگاه عامل، درجه ناکافی بودن از دیدگاه دیگران به همراه اهمیت فرد از نظر عامل، و درجه اجتناب‌پذیری خلاف‌واقع، که هر یک دارای وزن و ضریب خاصی هستند، به صورت زیر تعریف می‌شود.

تعریف ۱۰. درجه شرم:

$$D_{\text{Shame}_i^j} := \mu_1 \times D_{\text{Inq}_{P_{i,r}^a}, \varphi} + \mu_2 \times D_{\text{Social}(i, j, w)} \times D_{j, \text{Inq}_{P_{i,r}^a}, \varphi} + \mu_3 \times D_{\text{CHA}_{P_{i,r}^a}, \varphi}.$$

که در آن $(\mu_1 + \mu_2 + \mu_3 = 1, 0 < \mu_1, \mu_2, \mu_3 < 1)$. فرض کنید در یک پلتفرم آموزش آنلاین، عامل هوش مصنوعی (عامل i) طراحی شده است که به سؤالات دانش‌آموزان پاسخ می‌دهد و آن‌ها را در حل تمرین‌های درسی یاری می‌رساند. در یکی از موقعیت‌ها،

برای روشن‌تر شدن، فرض کنید فردی در یک رقابت حرفه‌ای به دلیل عدم تلاش کافی، عملکرد ضعیفی داشته باشد. اگر این فرد موفقیت را بخشی از هویت خود بداند، درجه ناکافی بودن از دیدگاه خود بالا خواهد بود. همچنین، اگر همکارانش او را مورد قضاوت قرار دهند و نظر آن‌ها برایش مهم باشد، درجه ناکافی بودن از دیدگاه دیگران نیز افزایش می‌یابد. در نهایت، اگر او آگاه باشد که با آمادگی بیشتر می‌توانست نتیجه بهتری بگیرد، درجه اجتناب‌پذیری نیز بالا خواهد بود. در این حالت، درجه شرم او شدید بوده و احتمالاً احساس ناکامی عمیقی را تجربه می‌کند.

در مقابل، اگر فردی در موقعیتی خطایی کوچک مرتکب شود که نه برای خودش و نه برای دیگران اهمیت چندانی نداشته باشد و نتیجه آن نیز خارج از کنترل او باشد، درجه شرم ناچیز خواهد بود و تأثیر احساسی قابل توجهی به دنبال نخواهد داشت.

فرض کنید یک عامل مکالمه‌گر هوش مصنوعی، در بخش پشتیبانی مشتری در وبسایت یک شرکت فعال است و وظیفه دارد به سؤالات کاربران پاسخ دهد و مشکلات آن‌ها را حل کند. حال تصور کنید یکی از کاربران با مشکلی جدی در حساب کاربری‌اش وارد گفتگو می‌شود، اما عامل به اشتباه چند پاسخ کلی و نامربوط ارائه می‌دهد و در نهایت موفق نمی‌شود مشکل را حل کند. اگر این عامل بر اساس یک مدل یادگیرنده طراحی شده باشد که کیفیت تعاملات را به عنوان بخشی از ارزیابی عملکرد خود در نظر بگیرد، و «رضایت کاربران» برایش شاخص موفقیت باشد، در این صورت، درجه ناکافی بودن از دیدگاه خود بالا خواهد بود، چراکه عملکردش را در تطابق با هویت خود به عنوان یک پشتیبان ناکافی می‌بیند. اگر کاربر ناراضی در پایان گفتگو امتیاز منفی بدهد یا بازخورد انتقادی ثبت کند، و عامل طوری طراحی شده باشد که بازخورد کاربران برایش اهمیت داشته باشد (مثلاً در ارزیابی خود یا گزارش‌دهی به سرپرست انسانی)، درجه ناکافی بودن

درجه‌ی اجتناب‌پذیری کنش برابر با:

$$D_{CHA_{i,r}^a, \varphi} = \frac{3}{5} = 0.6$$

حال با استفاده از وزن‌دهی $0.6 = \mu_1$ برای درجه‌ی غیراخلاقی بودن و $0.4 = \mu_2$ برای درجه‌ی اجتناب‌پذیری، شدت گناه به صورت ترکیب خطی این دو مؤلفه محاسبه می‌شود:

$$D_{Guilt_i, \varphi} = 0.6 \times 0.643 + 0.4 \times 0.6 \approx 0.626$$

فرض کنید درجه اهمیت این دو معیار برای عامل j به ترتیب 0.9 و 0.8 باشد، و مجموع اهمیت کل معیارها برابر با $2/60$ ، پس درجه‌ی ناکافی بودن کنش عامل i از دیدگاه عامل j برابر خواهد بود با:

$$D_{j, Inq_{i,r}^a, \varphi} = \frac{1/70}{2/60} \approx 0.653$$

با فرض اینکه یک عامل j ناظر بر کنش عامل i باشد و از پیامد آن آگاه باشد (برای مثال بعد از ثبت نظر توسط دانش‌آموز کنش عامل i مورد بازدید قرار می‌گیرد) و اهمیت بالایی برای عامل i داشته باشد (مثلاً $D_{Social}(i, j, w) = 0.85$) و با استفاده از وزن‌های $0.3 = \mu_1$ ، $0.4 = \mu_2$ و $0.3 = \mu_3$ شدت شرم نیز به صورت زیر محاسبه می‌شود:

$$\begin{aligned} D_{Shame_i^j, \varphi} &= 0.3 \times 0.686 + \\ &+ 0.4 \times (0.85 \times 0.653) + \\ &+ 0.3 \times 0.6 = 0.607 \end{aligned}$$

تحلیل‌های عددی فوق نشان می‌دهد که عامل i ، به دلیل تصمیم نادرست خود، هم گناه و هم شرم قابل‌توجهی را تجربه می‌کند. این مقادیر می‌توانند به عنوان هشدارهایی برای یادگیری، اصلاح رفتار، اولویت‌بندی سؤالات بحرانی، یا حتی ارائه‌ی عذرخواهی از سوی سیستم به کاربر، به کار گرفته شوند و نمونه‌ای دقیق از کارایی

دانش‌آموزی در نزدیکی امتحانات نهایی، سؤالی مرتبط با تمرین پایانی برای سیستم ارسال می‌کند؛ اما عامل به اشتباه، اهمیت سؤال را کم‌تأثیر تشخیص داده و پاسخ‌دهی را به تأخیر می‌اندازد. در نتیجه، دانش‌آموز موفق به ارسال تمرین نمی‌شود و مردود می‌گردد. کنش انجام‌شده (عدم پاسخ‌گویی به موقع) منجر به وقوع پیامدی منفی (φ) می‌شود، که طبق نمادگذاری به صورت $HP_{i,r}^a$ نمایش داده می‌شود.

فرض کنید در این سیستم پنج هنجار در نظر گرفته شده باشد: پاسخ‌گویی سریع به درخواست‌های مهم، تعهد به موفقیت تحصیلی دانش‌آموز، تحلیل درست اهمیت پرسش‌ها، حفظ احترام متقابل، و پاسخ‌دهی منصفانه. از این میان، تنها سه هنجار اول توسط کنش نقض شده‌اند. با استفاده از تابع D_{Norm} ، درجه اهمیت این سه هنجار به ترتیب برابر با 0.9 ، 0.95 ، و 0.85 فرض می‌شود و مجموع اهمیت کل هنجارها نیز برابر با $4/20$ است. بنابراین، درجه‌ی غیراخلاقی بودن برابر خواهد بود با نسبت مجموع اهمیت هنجارهای نقض‌شده $2/70$ به مجموع اهمیت همه‌ی هنجارها، یعنی:

$$D_{Imm_{i,r}^a, \varphi} = \frac{2/70}{4/20} \approx 0.643$$

هم‌چنین فرض کنید سه معیار ارزیابی در این سیستم تعریف شود: رضایت کاربر، موفقیت آموزشی، و رفتار حرفه‌ای. در این میان، دو معیار اول برآورده نشده‌اند. فرض کنید درجه اهمیت این دو معیار برای عامل i به ترتیب 0.9 و 0.85 باشد، و مجموع اهمیت کل معیارها برابر با $2/55$ ، پس درجه‌ی ناکافی بودن برابر خواهد بود با:

$$D_{Inq_{i,r}^a, \varphi} = \frac{1/75}{2/55} \approx 0.686$$

همچنین، فرض کنید از میان پنج کنش ممکن، سه کنش می‌توانستند از وقوع پیامد φ جلوگیری کنند؛ در نتیجه،

سطح بالاتری از طبیعی بودن و همدلی ارتقا می‌بخشد، بلکه در طراحی سیستم‌های هوشمند اخلاق‌مدار که قادر به ارزیابی اعمال خود بر اساس هنجارها و انتظارات اجتماعی هستند، نقشی کلیدی ایفا می‌کند. به عنوان مثال، یک عامل مصنوعی که بتواند «درجه گناه» خود را پس از یک خطا محاسبه کند، می‌تواند اقدامات جبرانی مؤثرتری را برگزیند. همچنین، در سیستم‌های چندعاملی، درک متقابل حالات احساسی صورتی‌شده عامل‌ها می‌تواند به شکل‌گیری اعتماد، هماهنگی بهتر و کارایی بیشتر در تصمیم‌گیری‌های جمعی و دستیابی به اهداف مشترک منجر شود. این پژوهش راه را برای تحقیقات آتی در زمینه پیاده‌سازی این مدل‌ها در معماری‌های عامل‌های شناختی و آزمودن آن‌ها در موقعیت‌های کاربردی پیچیده هموار می‌سازد.

مراجع

- [1] J. P. Tangney, J. Stuewig, and D. J. Mashek, "Moral emotions and moral behavior," *Annual Review of Psychology*, vol. 58, pp. 345–372, 2007.
- [2] P. M. Niedenthal, J. P. Tangney, and I. Gavanski, "“if only i weren’t” versus “if only i hadn’t”: Distinguishing shame and guilt in counterfactual thinking," *Journal of Personality and Social Psychology*, vol. 67, no. 4, pp. 585–595, 1994.
- [3] E. Lorini and F. Schwarzenruber, "A logic for reasoning about counterfactual emotions," *Artificial Intelligence*, vol. 175, no. 3–4, pp. 814–847, 2011.
- [4] N. Guiraud, D. Longin, E. Lorini, S. Pesty, and J. Rivière, "The face of emotions: A logical formalization of expressive speech acts," in *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, pp. 1031–1038, International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2011.
- [5] E. Lorini, "A dynamic logic of knowledge, graded beliefs and graded goals and its application to emotion modelling," in *Logic, Rationality, and Interaction. LORI 2011* (H. van Ditmarsch, J. Lang, and S. Ju, eds.), vol. 6953 of *Lecture Notes in Computer Science*, (Berlin, Heidelberg), pp. 165–178, Springer, 2011.

چارچوب پیشنهادی مقاله برای سنجش کمی احساسات در سیستم‌های چندعاملی را به نمایش می‌گذارند.

۴ نتیجه‌گیری

در این مقاله، چارچوبی صوری برای اندازه‌گیری شدت احساسات خلاف واقع، به‌ویژه گناه و شرم، ارائه شد. این چارچوب با تعریف مفاهیم بنیادین هنجارها و معیارهای ارزیابی آغاز گردید و سپس با معرفی فرمول‌های منطقی، احساس گناه را به‌عنوان پذیرش مسئولیت نقض هنجار همراه با امکان اجتناب، و شرم را به‌عنوان پذیرش قضاوت منفی دیگران درباره عدم کفایت، با آگاهی از این قضاوت و امکان اجتناب، تحلیل گردید.

برای کمی‌سازی شدت این احساسات، مفاهیمی همچون «درجه اهمیت» هنجارها و معیارها از دیدگاه عامل، «درجه اهمیت نظر دیگران»، «درجه غیراخلاقی بودن» (معیاری برای میزان نقض هنجارها)، «درجه ناکافی بودن» (میزان عدم تطابق با معیارها از دیدگاه خود عامل و نیز از دیدگاه دیگران) و «درجه اجتناب‌پذیری خلاف واقع» (نسبت کنش‌های جایگزین ممکن برای جلوگیری از پیامد نامطلوب) معرفی شدند. در نهایت، «درجه گناه» به‌عنوان ترکیبی از درجه غیراخلاقی بودن و درجه اجتناب‌پذیری خلاف واقع تعریف گردید. به طور مشابه، «درجه شرم» با ترکیب درجه ناکافی بودن از دیدگاه عامل، درجه ناکافی بودن از دیدگاه دیگران (با لحاظ اهمیت آن دیگران برای عامل) و درجه اجتناب‌پذیری خلاف واقع مشخص شد.

در حوزه‌هایی نظیر محاسبات احساسی و سیستم‌های چندعاملی خودمختار، این چارچوب می‌تواند به توسعه نسل جدیدی از عامل‌های هوشمند کمک شایانی کند که قادر به درک، شبیه‌سازی و واکنش متناسب به احساساتی همچون گناه و شرم باشند. چنین قابلیت‌هایی نه تنها تعاملات انسان-ماشین را به