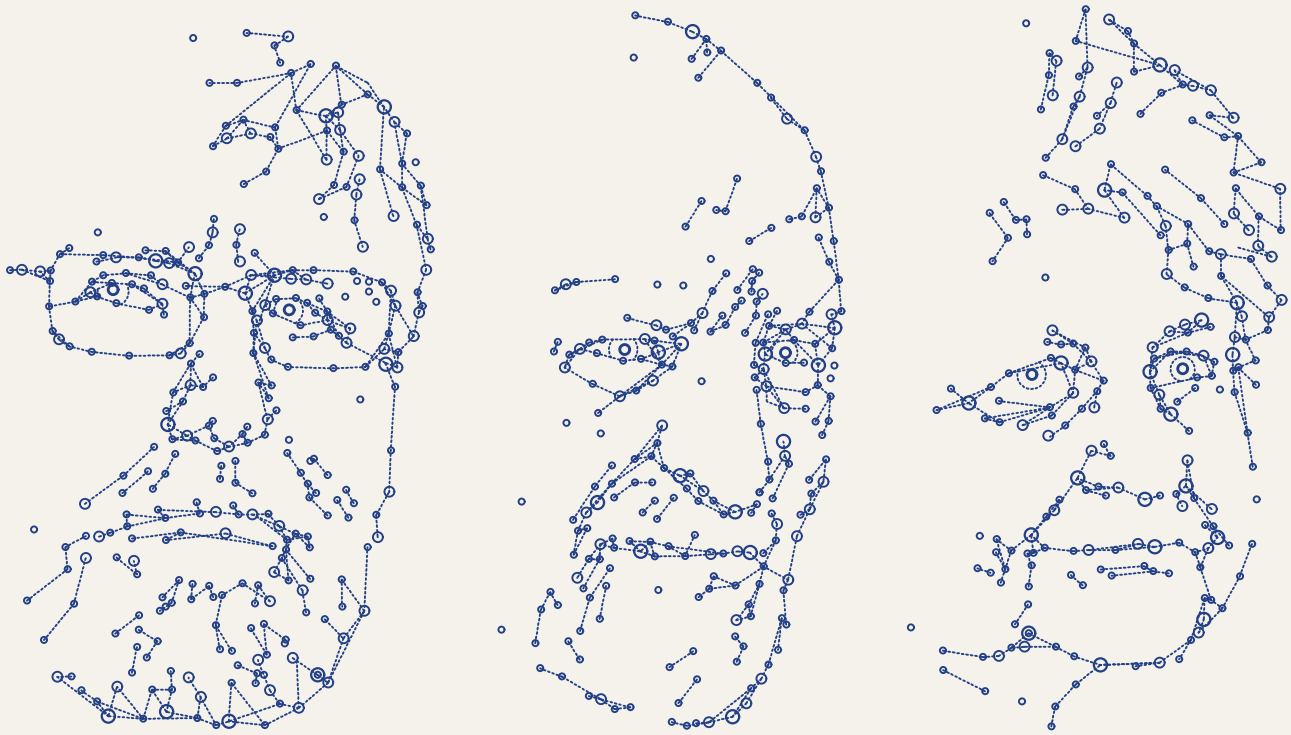
KOGNICE A UMĚLÝ ŽIVOT XX



TŘEŠŤ 30.5.–1.6. 2022

SBORNÍK Z 20. ROČNÍKU KONFERENCE

Editoři:

Gabriela Šejnová Michal Vavrečka Juraj Hvorecký

Kognice a umělý život XX
(recenzovaný sborník)

Vydáno v Praze, 2022

Vydavatel: České vysoké učení technické v Praze
Zpracoval: Český institut informatiky, robotiky a kybernetiky (CIIRC), ČVUT v Praze
Jugoslávských partyzánů 1580/3
160 00 Praha
Tel.: +420 224 354 224

Sborník sestavila Mgr. Gabriela Šejnová

Design obálky: Ing. arch. Pavel Š. Nosál

ISBN 978-80-01-07007-9
DOI: https://doi.org/10.14311/BK.9788001070079

Předmluva

Milé priaznivkyne a milí priaznivci kognitívnej vedy,

Som veľmi rád, že Vás môžem privítať na jubilejnom 20. ročníku konferencie Kognícia a umelý život. Moja radosť je o to väčšia, že sa usporiadanie tejto tradičnej akcie komunity ľudí z prostredia psychológie, informatiky, lingvistiky, robotiky, filozofie a ďalších príbuzných odborov koná po dvojročnej prestávke, zapríčinenej opakovanými návratmi pandémie. Keď som sa na jar tohto roku, trochu neopatrne, pýtal bratislavských organizátorov, či nechceme po vynútenej prestávke konferenciu obnoviť, správne poznamenali, že iniciatíva by mala vzísť z českej strany. Rozhodnutie padlo rýchlo, organizačné zmätky sme riešili s príslušnou dávkou amatérstva a odhodlania zároveň, a výsledkom je, dúfam, že dobre pripravený 20. ročník. Výhodou sa ukázala i možnosť využiť nanovo zrekonštruovaný Zámecký hotel Třešť, v ktorom sa už konferencia v minulosti konala, no tentokrát sa tak udeje v krajšom a, nádejam sa, i akademicky podnetnejšom prostredí. Veľmi ďakujem všetkým, ktorí sa o úspešnú prípravu i konanie konferencie pričinili.

20. výročie akejkoľvek akcie obvykle nabáda k bilancovaniu. Nenachádzam v sebe odvahu hodnotiť celkový prínos tejto konferencie, ale jedna vec je zrejmá. Kognitívna veda prešla za toto obdobie u nás enormným vývojom. Od mimoriadne skromných začiatkov, cez postupné budovanie prvých centier a študijných programov až po dnešok, charakterizovaný štandardným zapojením mnohých pracovísk do medzinárodného výskumu a uznávanými výsledkami. Dovolím si konštatovať, že svojou priateľskou atmosférou, otvorenosťou a vytvorením platformy na výmenu názorov k tomu nemalou mierou prispela aj konferencia Kognícia a umelý život.

Keď hovorím o minulosti, musím spomenúť aj jej smutnú tvár. Za posledné obdobie nás opustili tri zakladateľské osobnosti tohto odboru. Alica Kelemenová, Karel Pstružina a Ivan M. Havel sa každý svojim spôsobom pričinili o to, že kognitívna veda u nás nielen zapustila svoje korene, ale rozvíja sa a rastie netušeným tempom. Práve im je venovaný tento ročník a budem rád, keď si na nich pri prednáškach, diskusiách i čítaní príspevkov spomeniete. Prajem Vám krásny pobyt v Třešti, mnoho úspechov v práci i doma a teším sa na ďalšie pokračovania tejto konferencie.

Mgr. Juraj Hvorecký, PhD

Obsah

Sekce: Plný článek 7

Antecedents of prosocial behaviour: A hybrid choice model investigation of the willingness to play for charity
Eva Ballová Mikušková, Magdalena Adamus 8

Who is afraid of migration?
Ivan Brezina, Vladimíra Čavojová 15

Correction rate in cognitive reflection test as a possible measure of analytical thinking dispositions
Roman Burič 19

Tréning kapacity vizuálnej pracovnej pamäti v prostredí virtuálnej reality
Barbora Cimrová, Martin Marko, Igor Farkaš, Branislav Sobota, Štefan Korečko, Zuzana Rošťáková, Roman Rosipal 23

Zvyšovanie efektivity trénovania a kapacity v atraktorovom neurálnom modeli asociatívnej pamäte
Matej Fandl, Martin Takáč 28

Pozornosť ako biologicky inšpirovaný koncept pre vysvetliteľné, robustné a efektívne strojové učenie
Igor Farkaš, Barbora Cimrová, Štefan Pócoš, Iveta Bečková 34

It Is Easier with Negative Emotions: The Role of Negative Emotions and Emotional Intelligence in Epistemically Suspect Beliefs about COVID-19
Miroslava Galasová 39

Body schema or the body as its own best model
Matěj Hoffmann 45

Kompozitní testování inteligence jako možná cesta k univerzální psychometrii
Petr Hoza, Ondřej Vadinský 52

Ortogonalita vedomia a obsahu a proces uvedomenia
Juraj Hvorecký 58

3D rekonštrukcia tváre v počítači a mysli
Andrej Lúčny 62

Vliv vybraných charakteristik jedince prožívajícího krizi na ochotu komunikovat s chatbotem
Lenka Macháčková, Daniel Dostál 66

Generatívne vlastnosti modelu UBAL
Kristína Malinovská, Igor Farkaš 74

Neurónová sieť s násobiacou vrstvou
Ľudovít Malinovský, Kristína Malinovská 79

Intrinsic motivation based on feature extractor distillation
Matej Pecháč, Igor Farkaš 84

BESST: Brno Extended Speech and Stress Test
Jan Pešán, Vojtěch Juřík 91

Diferenciální evoluce s adaptací velikosti populace v závislosti na diverzitě
Radka Poláková, Petr Bujok 98

Prediktory vizuálnej predstavivosti: senzorická senzitivita, všímavosť a osobnostné dimenzie
Alexandra Ružičková, Lenka Jurkovičová, Jan Páleník, Vojtěch Juřík 106

Úloha stelesnenia a pohľadu v interakcii človeka a robota
Sabína Samporová, Kassandra Friebe, Kristína Malinovská, Matěj Hoffmann 112

Umělý život jako umělecký a estetický problém
Aleš Svoboda 117

COVID-19 pandemic may have changed our attitudes to science, but not our ability to reason scientifically
Jakub Šrol, Vladimíra Čavojová 122

Etické aspekty neurorobotických simulácií
Martin Takáč, Alistair Knott, Mark Sagar 126

Embodied ideas
Silvia Tomašková 133

Přehled obecných přístupů k vyhodnocování inteligence umělých systémů
Ondřej Vadinský 137

Vlci a smečkový algoritmus ve světě membránových agentů
Daniel Valenta, Lucie Ciencialová, Luděk Cienciala 144

Využití membránového systému pro simulaci komunikace v síti Internetu věcí
Šárka Vavrečková 151

Sekce: Rozšířený abstrakt 159

Skúmanie vzdialeností adverzariálnych vstupov k jednotlivým triedam v hlbokých neurónových sieťach
Iveta Bečková, Štefan Pócoš, Igor Farkaš 160

Contextual Plasticity in Sound Localization vs. Source Separation in Real and Virtual Environments
Stanislava Linková, Gabriela Andrejková, Norbert Kopčo 162

A model of the reference frame of the ventriloquism aftereffect based on head-centered, eye-centered and distance-dependent signals
Peter Lokša, Norbert Kopčo 164

Špecifiká teórie mysle u pacientov so schizofréniou a bipolárnou afektívnou poruchou
Ivana Mirdalíková 166

Systém rozpoznávání akcí integrující detekci objektů a jejich pohybů
Anastasia Ostapenko, Michal Vavrečka 168

Is none treatment for mental health problems better than a controversial one?
Klára Petrovická 170

Učení se reprezentace peripersonálního prostoru pomocí neuronových sítí
Zdeněk Straka, Matěj Hoffmann 173

Potenciál augmentované reality během letu
Čeněk Šašinka, Zdeněk Stachoň, Kateřina Chmelařová, Kateřina Jochecová 175

Kolaborativní imerzivní virtuální realita v praxi a výzkumu
Čeněk Šašinka, Jiří Chmelík, Alžběta Šašinková, Zdeněk Stachoň 176

Inkrementální imitační učení pomocí variačního autoenkodéru
Gabriela Šejnová, Karla Štěpánová 177

Personifikovaný robotický chatbot založený na kompozičních dialozích
Michal Vavrečka, Gabriela Šejnová, Petr Schimperk 179

O procitnutí hmoty
Jiří Wiedermann 181

Rejstřík autorů 184

SEKCE:

PLNÝ ČLÁNEK

Antecedents of prosocial behaviour: A hybrid choice model investigation of the willingness to play for charity

Eva Ballová Mikušková, Magdalena Adamus

Institute of Experimental Psychology, Centre of Social and Psychological Sciences, Slovak Academy of Sciences
Dúbravská cesta 9, 841 04 Bratislava, Slovakia
eva.ballova-mikuskova@savba.sk; magdalena.adamus@savba.sk

Abstract

Using a hybrid choice model comprising a Real Effort Task (sliders) and psychological questionnaire methods, we aimed at investigating socio-demographic antecedents of the willingness to play games for charity. We also explored the relationship between the number of correctly solved tasks and socio-demographic and psychological characteristics. A total of 500 participants (aged 18 – 86; M = 44.32, SD = 15.66) answered questions about their religiosity, conservatism/liberalism, prosocial behaviour, collectivism/individualism, future orientation, feeling of helplessness and threat, and personality traits, and at the end were asked to take part in a task in which they could raise money for a charity. The results show that people willing to play the game for charity – compared to those who declined – were significantly more liberal, with more pro-vaccination attitudes (and less anti-vaccination attitudes) and feeling more helpless on climate change issues. In the sample of people willing to play for charity (N = 357; Mage = 43.82, SD = 15.61), the number of correctly solved tasks correlated negatively with age, vertical collectivism, general prosocial behaviour and the threat from climate change.

1 Introduction

Apart of being a considerable health risk, the COVID-19 pandemic raised questions about sources of people's willingness to behave prosocially despite the looming threat (Vladimíra Čavojová et al., 2022). We saw from the very beginning of the pandemic that people cared about those around them. They were concerned not only about their closest family members, but also about neighbours and friends. Not only did they adhere to the containment measures – which in itself is already a prosocial behaviour – but, for instance, some people shopped for those most vulnerable. Meanwhile, others were paying for lunches they never ate to help their favourite restaurants to survive. Generally, people showed they care and helped depending on the situation. For instance, a recent research (Vladimíra Čavojová et al., 2022) showed that many people adopted prosocial behaviours above the required containment measures such as caring for others, providing emotional support, restraining excessive purchases, or limiting unnecessary meetings. Those types of behaviour included calling to check on the family and friends, shopping for the sick, or finding alternative ways of meeting even when face-to-face meetings were allowed under certain conditions. All these types of behaviour may seem simple, however, during the pandemic they required extensive planning and organisation as well as considerable sensitivity to the needs of other people. And all these prompted us to seek an answer to the question what would be the antecedents of such prosocial or other-regarding behaviour.

Our main idea was that people who behave prosocially may share some common characteristics and we posed a question of what those characteristics would be. Previous studies provided evidence that prosocial behaviour could be associated with a series of individual differences in personality traits or value orientation. For instance, Blagov (2020) investigated how different sounding of public health messages appealed to different people depending on their personality and how likely these messages were to affect their behaviour. He found that Agreeableness and Conscientiousness predicted the appeal of compassionate and responsible public health messages.

Apart from personality, also value orientation was strongly related to behaviour during the pandemic. Specifically, individualists tended to show more disregard for experts' recommendations and greater support towards individuals who disobey the recommendations compared to those scoring high in collectivism (Shea & Ueda, 2021). Similarly, Lalot et al. (2021) found that future-oriented individuals adhered to containment measures more, reported a greater sense of compassion and got involved in collective actions, such as donating money or volunteering more often. Serrano-Montilla et al. (2021) also claimed that in their prosocial behaviour people may be also motivated by emotions such as fear or anxiety. In the study by Čavojová et al. (2022), fear of

COVID – as a health-related threat – remained the strongest predictor of other-regarding behaviour during the pandemic. Finally, socio-demographic characteristics also seem to be important drivers of the tendency to behave prosocially. Specifically, people with higher education, more liberal and less religious were more likely to comply with health-promoting behaviour (Vladimíra Čavojová et al., 2022).

Building on the extant literature, in the current study, we thus investigated the relationship between prosocial behaviour and a series of socio-demographic characteristics, religiosity, conservatism/liberalism, prosocial behaviour, collectivism/individualism, future orientation, feeling of helplessness and threat, and personality traits.

However, to obtain a fuller picture of the phenomenon, we decided to go beyond simple self-report methods and combine them with an observation of actual behaviour. Consequently, using a hybrid choice model comprising a Real Effort Task (sliders) and psychological questionnaire methods, we aimed at investigating socio-demographic antecedents of the willingness to play games for charity. Hybrid choice models were developed to add unobservable, mainly attitudinal variables to empirical investigation of choices. In other words, the models allow to embed actual choices observed during the research into a set of individual characteristics such as personality traits, beliefs or values (Kim et al., 2014). To delve deeper on the factors that could induce people to behave prosocially, we also explored the relationship between the number of correctly solved tasks and socio-demographic as well as aforementioned psychological characteristics.

At the same time, most studies investigated prosocial behaviour during the pandemic we found in the extant literature predominantly used self-reported measures. Aware of the concerns about comparability of self-reports and discrete choice models, we aimed to investigate the relationship between the willingness to play and the performance in the slider task and a series of questionnaires about prosocial behaviour. To test the comparability of those two different measurement approaches, in the present study, we employed measures of behaviour related to such diverse forms of prosociality as vaccination, caring and helping behaviour and also pro-environmental behaviour.

Finally, the present study aimed to test the possibility of using a method typically employed in laboratory experiments – Real Effort Task – in a more convenient online environment. This involved a two-step investigation. First, we considered the technical feasibility of using the slider method outside the laboratory. Second, we explored whether results obtained by this observational method provide results consistent with the psychological literature and the self-reported measures purportedly capturing a related phenomenon. The possibility to apply such methods online, would considerably reduce costs of research and also made observational study more widely available since there would be less demand for using laboratory infrastructure.

2 Methods

2.1 Participants and design

The data were collected online (form created in Qualtrics) as part of a larger study. Participants were recruited by the external agency (to be representative of the Slovak population concerning age and gender) and were remunerated for their participation according to an internal scoring system of the external agency by credit points or vouchers. All methods were carried out following APA standards and were approved by the Ethical Board of Masaryk University as a part of the MSCA-IF grant (MSCAfellow3@MUNI).

A total of 500 participants (aged 18 – 86; *M* = 44.32, *SD* = 15.66; 13.4 % elementary or incomplete high school education, 46.2% complete high school education, 40.4 % college or complete college education) answered questions about their religiosity, conservatism/liberalism, prosocial behaviour, collectivism/individualism, future orientation, feeling of helplessness and threat, and personality traits. At the end of the questionnaire section, participants were asked to take part in a task in which they could raise money for a charity. Participants were informed that the money they earn will be transferred by the study organisers from the research budget and that the maximum they could earn is one euro depending on the number of correctly solved tasks.

2.2 Measurements

Participants were asked to indicate their age, sex, level of education, conservativism/liberalism (very conservative = 1, very liberal = 7), and importance of religion (not at all important = 1, very important = 7).

Prosocial behaviour was measured as general prosocial behaviour, behaviour during the COVID-19 pandemic, pro-environmental behaviour, pro-vaccination behaviour and anti-vaccination attitudes.

General prosocial behaviour was measured by 23 items of Prosocial Tendencies Measure (PTM; Babinčák, 2011; Carlo & Randall, 2002) assessing 6 types of prosocial behaviors: altruistic, compliant, emotional, dire, public, and anonymous. Participants responded on a 5-point scale (1 = not at all like me, 5 = absolutely like me) and behaviours were analyzed as a composite score.

Prosocial behaviours regarding the COVID-19 pandemic, pro-environmental behaviour and pro-vaccination behaviour were measured by 5 self-reported items (Vladimíra Čavojová et al., 2022) for each domain. One point has been assigned to answers indicating most selfish answers, 2 points to answers indicating some action/willingness and 3 points to answers indicating action. The mean scores were used with higher score indicating higher prosocial behaviour.

Anti-vaccination attitudes were measured by ten items (V. Čavojová et al., 2022; Wallace et al., 2019) in which participants indicated agreement on a 5-point scale (1 = strongly agree, 5 = strongly disagree). Higher mean scores indicated stronger anti-vaccination attitudes.

Collectivism/Individualism was measured by the 14-item version of Horizontal and Vertical Individualism and Collectivism Scale (HVIC; Singelis et al., 1995; Sivadas et al., 2008) measuring horizontal individualism (HI; the sense of being self-reliant without a tendency to compete with others), vertical individualism (VI; competitively establishing one's status), horizontal collectivism (HC; a tendency to acknowledge interdependence), and vertical collectivism (VC; tendency to establish group hierarchy and compete with members of an out-group). Participants had to agree with items on a 5-point scale from strongly disagree (1) to strongly agree (5) and the mean scores were calculated for each subscale with the higher score indicating the higher preference.

Future orientation—the extent to which people consider the potential distant outcomes in their current behavior—was measured by Consideration of Future Consequences scale (CFC; V. Čavojová & Jurkovič, 2017; Joireman et al., 2012) on two subscales: future and immediate orientation. Participants had to evaluate 14 items on a 7-point scale from strongly disagree (1) to strongly agree (7). The mean scores were calculated for each subscale with the higher score indicating the higher preference.

Feeling of helplessness was measured for three domains: COVID-19 pandemic, climate change, and vaccination. For each type of helplessness participants rated 4 items on a 7-point scale (1 = completely disagree, 7 = completely agree). Scale were modification of scale from Šrol et al. (2021). The mean scores were used to indicate helplessness with the higher score indicating the stronger feelings.

Similarly, *feeling of threat* was measured: participants rated 3 items on a 7-point scale (1 = not threatened at all, 7 = extremely threatened) to indicate how a they felt about COVID-19 pandemic, climate change and vaccination (Kohút et al., 2022) when thinking about their health, quality of life, and economic and social consequences. The mean scores were used to indicate feeling of threat with the higher score indicating the stronger feelings.

Personality traits were measured by the Big Five Inventory 2 short form (Halama et al., 2020; Soto & John, 2017) with 30 items measuring five personality factors: Extraversion, Agreeableness, Conscientiousness, Negative Emotionality, and Open-Mindedness. Participants indicate their agreement with the items using a 5-point scale (1=disagree strongly, 5=agree strongly). The mean scores were calculated for each factor with the higher score indicating the higher preference.

Willingness to play for charity was measured as an agreement with participation in the game (agree/disagree) and as a number of correctly solved tasks in the game. In the game, participants had to solve simple tasks – set the slider on the scale to the specified value and could earn maximum one euro for charity if successful in all tasks.

3 Results

Descriptive statistics as well as internal consistency (Cronbach's alpha) are reported in Table 1.

Table 1 Descriptive statistics

	M	*SD*	*min*	*max*	*α*
age	44.32	15.66	18	86	-
conservatism (1) - liberalism (7)	3.88	1.22	1	7	-
religiosity	3.95	2.14	1	7	-
number of correctly solved tasks (*n*=357)	33.75	17.75	0	50	-
prosocial behaviour	3.41	0.45	2	5	.88
prosocial behaviour – pandemic	2.01	0.36	1	3	.49
pro-environmental behaviour	2.01	0.30	1	3	.71
pro-vaccination behaviour	1.75	0.50	1	3	.67
anti-vaccination attitudes	2.78	1.01	1	5	.91
horizontal collectivism	3.58	0.73	1	5	.72
vertical collectivism	3.39	0.70	1	5	.59
vertical individualism	2.92	0.88	1	5	.74
horizontal individualism	3.51	0.79	1	5	.63
future orientation	4.77	1.02	1	7	.85
immediate orientation	3.72	1.08	1	7	.82
helplessness –	3.59	1.60	1	7	.90

pandemic					
helplessness – climate	4.15	1.56	1	7	.91
helplessness – vaccination	3.28	1.70	1	7	.85
threat – pandemic	4.62	1.44	1	7	.84
threat – climate	4.46	1.55	1	7	.93
threat – vaccination	3.80	1.97	1	7	.95
extraversion	3.24	0.72	1	5	.74
agreeableness	3.69	0.66	1	5	.72
conscientiousness	3.77	0.68	2	5	.76
negative emotionality	2.79	0.77	1	5	.76
open-mindedness	3.48	0.65	2	5	.67

The results show (Table 2) that people willing to play the game for charity – compared to those who declined – were significantly more liberal, with more pro-vaccination attitudes (and less anti-vaccination attitudes) and feeling more helpless on climate change issues.

Table 2 Comparison of participants willing and not willing to play the game for charity

		M	*SD*	*p*
age	yes	43.82	15.61	.258
	no	45.57	15.77	
conservatism (1) - liberalism (7)	yes	3.98	1.18	**.004**
	no	3.64	1.27	
religiosity	yes	3.95	2.15	.958
	no	3.96	2.11	
prosocial behaviour	yes	3.42	0.46	.282
	no	3.37	0.44	
prosocial behaviour – pandemic	yes	2.02	0.35	.560
	no	2.00	0.38	
pro-environmental behaviour	yes	2.02	0.30	.086
	no	1.97	0.31	
pro-vaccination behaviour	yes	1.78	0.51	**.037**
	no	1.68	0.48	
anti-vaccination attitudes	yes	2.70	0.99	**.008**
	no	2.97	1.06	
horizontal collectivism	yes	3.60	0.76	.324
	no	3.53	0.67	
vertical collectivism	yes	3.40	0.72	.332
	no	3.34	0.65	
vertical individualism	yes	2.90	0.91	.577
	no	2.95	0.79	
horizontal individualism	yes	3.51	0.80	.833
	no	3.49	0.77	
future orientation	yes	4.81	0.99	.142
	no	4.66	1.06	
immediate orientation	yes	3.67	1.11	.114
	no	3.84	0.99	
helplessness – pandemic	yes	3.55	1.59	.433
	no	3.68	1.63	
helplessness – climate	yes	4.27	1.54	**.005**
	no	3.84	1.56	
helplessness – vaccination	yes	3.20	1.67	.105
	no	3.48	1.76	
threat – pandemic	yes	4.65	1.42	.580
	no	4.57	1.48	
threat – climate	yes	4.49	1.56	.450
	no	4.37	1.55	
threat – vaccination	yes	3.70	1.97	.076
	no	4.05	1.96	
extraversion	yes	3.22	0.73	.255
	no	3.30	0.70	
agreeableness	yes	3.72	0.66	.099
	no	3.61	0.64	
conscientiousness	yes	3.80	0.67	.178
	no	3.71	0.70	
negative emotionality	yes	2.79	0.77	.900
	no	2.80	0.77	
open-mindedness	yes	3.51	0.64	.110
	no	3.41	0.65	

In the sample of people willing to play for charity (n = 357; M_{age} = 43.82, SD_{age} = 15.61), the number of correctly solved tasks correlated negatively with age, vertical collectivism, general prosocial behaviour and the threat from climate change (Table 3).

Table 3 Correlations of number of correctly solved tasks in the game and measured variables

	total sample	willing to play
n	*500*	*357*
sex (0=men, 1=women)	.004	.066
age	**-.107***	**-.120***
education	.080	.041
conservatism (1) - liberalism (7)	.071	-.036
religiosity	-.041	-.066
prosocial behaviour	-.043	**-.129***
prosocial behaviour – pandemic	.052	.058
pro-environmental behaviour	.055	.00
pro-vaccination behaviour	.073	.011
anti-vaccination attitudes	**-.106***	-.037
horizontal collectivism	.014	-.029
vertical collectivism	-.044	**-.123***
vertical individualism	-.036	-.029
horizontal individualism	.025	.03
future orientation	.011	-.061
immediate orientation	-.052	-.003
helplessness – pandemic	-.049	-.04

helplessness – climate	.056	-.059
helplessness – vaccination	-.064	-.020
threat – pandemic	-.010	-.047
threat – climate	-.050	**-.125***
threat – vaccination	-.078	-.037
extraversion	-.037	-.002
agreeableness	.01	-.072
conscientiousness	**.096***	.090
negative emotionality	-.053	-.083
open-mindedness	**.092***	.069

4 Discussion and conclusions

The aim of the present study was three-fold. First we intended to investigate antecedents of prosocial behaviour using a hybrid choice model involving a Real Effort Task (sliders). Second, we compared both the willingness to play and the performance in the task with self-reported prosociality along a set of behaviours. Finally, our results allowed us to draw preliminary conclusions about applicability of Real Effort Tasks in particular and Hybrid Choice Models generally in psychological research.

In line with the literature, our results show that people willing to behave prosocially – play for charity – are more liberal and have more favourable attitudes towards vaccines. The relationship between socio-demographic characteristics and self-reported prosocial behaviour was shown also by previous studies. For instance, Čavojová et al. (2022) found out that even after adding personality traits and values to the model, helping behaviour during the pandemic was related to education and liberal orientation. Interestingly, our study shows that people who are prone to behave prosocially declare more helplessness about clime change, i.e., they feel they have little impact on mitigating the changes and their consequences. The fact that such people are more prosocial does not seem straightforward. However, a recent study by Adamus et al. (2022) found that helplessness relation with pro-environmental behaviour is far from being straightforward. Specifically, the authors found that helplessness was positively related to both environmental concern and, although to lesser extent, pro-environmental behaviour. In other words, the more helpless people about climate change felt, the more sustainable their behaviour was. We can only speculate that people who feel helpless may behave prosocially despite the odds and their conviction that their behaviour would change little. And yet they do not want to – for whatever reason – to be negligent.

Additionally, people who did play for charity differed in terms of the number of correctly solved tasks. Specifically, older people with stronger vertical collectivism, prosocial tendencies and threat from climate change tended to solve less tasks correctly. Among the findings only age seems to be straightforward. It is likely that the slider task was more challenging for older participants and thus their performance was slightly poorer compared to younger. The findings about collectivism, climate change threat and particularly prosocial tendencies are puzzling. Based on the extant literature we could expect positive relations between those individual characteristics and prosocial behaviour. Specifically, Čavojová et al. (2022) and Adamus et al. (2022) found collectivism and the combination of collectivism, future orientation and prosocial tendencies to be positively related to helping behaviour during the pandemics and more sustainable behaviour in general. Both the studies, however, employed only self-reported measures of behaviour considered as prosocial. Consequently, we delved deeper on the relationship between the slider task and self-reports of prosocial behaviour in three domains: vaccination, helping during the pandemic and pro-environmental behaviour. To our surprise, the preliminary analysis revealed that neither the willingness to play nor the number of correctly solved slider tasks were related to any of the self-reported measures.

Third, it seems that using Hybrid Choice Models in internet setting is a technically viable alternative for correlational studies or laboratory experiments. It proved to be easy to programme and also financially undemanding. Specifically, we were able to use the slider tasks in a browser with good resolution and appealing graphical design. Programming required less complex skills that typical zTree or oTree methods based on Python. However, the results themselves could be disappointing.

Although we managed to identify characteristics contributing to the willingness to behave prosocially (play for charity) and the number of correctly solved tasks, we found no correlations between the slider task and any of the other measures of prosocial behaviour. Finally, although attitudinal variables predicted self-reported prosocial behaviour relatively well, we found virtually no relation between neither the willingness to play nor the performance in the slider task and any of the unobservable attitudinal variables we included in our study. The results, thus, put a grain of salt into our preliminary optimism concerning the applicability of Hybrid Choice Models in the study on prosocial behaviour. The results went contrary to the predictions of psychological theories and were clearly inconsistent with self-reported measures of prosocial behaviour. Given that both self-reports and Real Effort Tasks are well established methods of capturing prosocial tendencies

and behaviour, our results are both puzzling and troubling.

One of the possible explanations is that Real Effort Tasks (including sliders) are still insufficiently studied in the context of psychological variables such as personality, beliefs or values. The fact that in laboratory settings performance in such tasks is usually measured in the context of experimentally manipulated contextual settings or endowment effects may indicate limited applicability of such tasks in more psychologically complex research. On the other hand, there still remains a possibility that the task itself turned out to be appealing and interesting inclining the participants to play regardless of their prosocial tendencies. Finally, we cannot rule out the experimenter effect: if participants in our study felt they were expected to participate in the task and perform well, this imagined expectation could distort their actual motivations. All these make it necessary to continue more detailed investigations of methods previously applied mainly in the lab. Only collecting more data would help answering the questions about the relationship between measures of prosocial behaviour themselves and their shared relationships with a wide network of socio-demographic and psychological characteristics.

Acklowledgements

The study was supported by the Slovak Research and Development Agency under contract no. APVV-20-0335: “Reducing the spread of disinformation, pseudoscience and bullshit”.

References

Adamus, M., Šrol, J., Čavojová, V., & Ballová Mikušková, E. (2022). *Seeing past the tip of your own nose? How outward and self-centred orientations could contribute to closing the green gap despite helplessness*.

Babinčák, P. (2011). Prosocial Tendencies Measure–Revised (PTM- R) - prvá skúsenosť s krátkou metodikou na meranie prosociálneho správania. In K. Bartošová, M. Čerňák, P. Humpolíček, M. Kukaňová, & A. Slezáčková (Eds.), *Sociální procesy a osobnost. Člověk na cestě životem: Křižovatky a mosty* (pp. 7–12).

Blagov, P. S. (2020). Adaptive and dark personality in the COVID-19 pandemic: Predicting health-behavior endorsement and the appeal of public-health messages. *Social Psychological and Personality Science*, *12*(5), 697–707. https://doi.org/10.1177/1948550620936439

Carlo, G., & Randall, B. A. (2002). The development of a measure of prosocial behaviors for late adolescents. *Journal of Youth and Adolescence*, *31*(1), 31–44. https://doi.org/10.1023/A:1014033032440

Čavojová, V., & Jurkovič, M. (2017). Comparison of experienced vs novice teachers in cognitive reflection and rationality. *Studia Psychologica*, *59*(3), 100–112. https://doi.org/10.21909/sp.2017.02.733

Čavojová, V., Šrol, J., & Ballová Mikušková, E. (2022). How scientific reasoning correlates with health-related beliefs and behaviors during the COVID-19 pandemic? *Journal of Health Psychology*, *27*(3), 534–547. https://doi.org/10.1177/1359105320962266

Čavojová, Vladimíra, Adamus, M., & Ballova-Mikuskova, E. (2022). You before me: How vertical collectivism and feelings of threat predicted more socially desirable behaviour during COVID-19 pandemic. *Current Psychology*, *n.d.* https://doi.org/10.1007/s12144-022-03003-3

Halama, P., Kohút, M., Soto, C. J., & John, O. P. (2020). Slovak adaptation of the Big Five Inventory (BFI-2): Psychometric properties and initial validation. *Studia Psychologica*, *62*(1), 74–87. https://doi.org/10.31577/sp.2020.01.792

Joireman, J., Shaffer, M. J., Balliet, D., & Strathman, A. (2012). Promotion orientation explains why future-oriented people exercise and eat healthy: Evidence from the two-factor consideration of future consequences-14 scale. *Personality and Social Psychol- Ogy Bulletin*, *38*(10), 1272–1287. https://doi.org/10.1177/0146167212449362

Kim, J., Rasouli, S., & Timmermans, H. (2014). Hybrid choice models: Principles and recent progress incorporating social influence and nonlinear utility functions. *Procedia Environmental Sciences*, *22*, 20–34. https://doi.org/10.1016/j.proenv.2014.11.003

Kohút, M., Šrol, J., & Čavojová, V. (2022). How are you holding up? Personality, cognitive and social predictors of a perceived shift in subjective well-being during COVID-19 pandemic. *Personality and Individual Differences*, *186*(111349). https://doi.org/10.1016/j.paid.2021.111349

Lalot, F., Abrams, D., Ahvenharju, S., & Minkkinen, M. (2021). Being future-conscious during a global crisis: The protective effect of heightened Futures Consciousness in the COVID-19 pandemic. *Personality and Individual Differences*, *178*(March), 1–8. https://doi.org/10.1016/j.paid.2021.110862

Serrano-Montilla, C., Alonso-Ferres, M., & Lozano, L. M. (2021). Assessment of the effects of health and

financial threat on prosocial and antisocial responses during the COVID-19 pandemic: The mediating role of empathic concern. *Personality and Individual Differences*, *178*(March), 1–4. https://doi.org/10.1016/j.paid.2021.110855

Shea, B. A. O., & Ueda, M. (2021). Who is more likely to ignore experts ’ advice related to COVID-19 ? *Preventive Medicine Reports*, *23*(January), 1–5. https://doi.org/10.1016/j.pmedr.2021.101470

Singelis, T., Triandis, H., Bhawuk, D., & Gelfand, M. (1995). Horizontal and vertical dimensions of individualism and collectivism: A theoretical and measurement refinement. *Cross-Cultural Research*, *29*(3), 240–275. https://doi.org/10.1177/106939719502900302

Sivadas, E., Bruvold, N. T., & Nelson, M. R. (2008). A reduced version of the horizontal and vertical individualism and collectivism scale: A four-country assessment. *Journal of Business Research*, *61*(3), 201 – 210. https://doi.org/10.1016/j.jbusres.2007.06.016

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*, 117 – 143. https://doi.org/10.1037/pspp0000096

Šrol, J., Ballová Mikušková, E., & Cavojova, V. (2021). When we are worried, what are we thinking? Anxiety, lack of control, and conspiracy beliefs amidst the COVID-19 pandemic. *Applied Cognitive Psychology*. https://doi.org/10.1002/acp.3798

Wallace, A. S., Wannemuehler, K., Bonsu, G., Wardle, M., Nyaku, M., Amponsah-Achiano, K., Dadzie, J. F., Sarpong, F. O., Orenstein, W. A., Rosenberg, E. S., & Omer, S. B. (2019). Development of a valid and reliable scale to assess parents’ beliefs and attitudes about childhood vaccines and their association with vaccination uptake and delay in Ghana. *Vaccine*, *37*(6), 848 – 856. https://doi.org/10.1016/j.vaccine.2018.12.055

Who is afraid of migration?

Ivan Brezina & Vladimíra Čavojová

Institute of Experimental Psychology, Centre for Social and Psychological Sciences, SAS
Dúbravská cesta 9, 841 04 Bratislava
ivan.brezina@savba.sk, vladimira.cavojova@savba.sk

Abstract

In our exploratory study we asked 666 Slovak participants to a) indicate their attitudes about migration, b) estimate their subjective competence in the topic of migration, c) fill a short quiz about migration. Consequently, participants were randomly assigned to two conditions: provide either three arguments for the benefits of migration or three arguments against migration. According to our results, men and women did not differ in their attitudes toward migration ($t(663) = 0.825$, $p = .410$), and the attitudes did not correlate with age ($r = .053$, $p = .169$), but they correlated with education ($r = .205$, $p < .001$): more educated people had more positive attitudes toward migration. Attitudes did not correlate with subjective competence ($r = .027$, $p = .490$), but they correlated positively with actual knowledge ($r = .128$, $p = .001$) and negatively with overestimation ($r = –.151$, $p < .001$): more positive attitudes toward migration were associated with higher score on knowledge quiz and tendency to underestimate one´s knowledge. We also analysed arguments qualitatively, and as a result, we divided them into five negative categories (labour market, security, disease & hygiene, fear of cultural differences, economic burden) and five positive categories (labour force, population growth/demographics, cultural enrichment, economic growth, helping & reciprocity).

1 Introduction

Migration remains hotly debated political topic in many European countries since the crisis of 2015, often polarizing attitudes of public and not rarely stripping other social issues of their prominence.

While religiosity varies significantly among populations of Central Europe (Poles and Czechs rank on different extremes of the scale), public attitudes towards religious, ethnic and other minorities show both less diversity within the region and more negative inclination as compared to Western Europe.
“The continental divide” in respective attitudes and values across Europe, as identified by Pew Research Centre (2018), could be illustrated by the finding that in nearly every Central and Eastern European country polled, *fewer than half* of adults claim they would be willing to accept Muslims into their family; while in nearly every Western European country surveyed, *more than half* say they would accept a Muslim into their family. When, for example, asked how comfortable would respondents feel if a colleague at work with whom the participants are in daily contact was a Muslim – Czechia was the only country in EU where people were more likely to say they would be uncomfortable than comfortable. In the context of the same question, Hungary ranked the second last in EU (both Poland and Slovakia ranked within the last third of the countries), and the least welcoming when the colleague was of Buddhist religion (Special Eurobarometer 469, 2018).

In the following paper we will explore and describe the content categories the distinct thematic layers of attitudes toward migrants through arguments that participants of representative sample elaborated in virtual online discussion with the motivation of financial reward (pro and contra migration). Participants´ demography, knowledge about the topic, overestimation of their knowledge, as well as their attitudes towards migration will be analysed together with the content of their (pro and con) arguments.

2 Methods

2.1 Participants

The representative sample comprised 666 Slovak nationals (52.7% were men) aged between 19 and 84 ($M = 41.84$, $SD = 13.90$). Of these, 30.2% had completed lower secondary education, 45.5% had attained upper secondary education and 24.4% had completed higher education. The participants were recruited by the external participant recruitment agency and they were rewarded with points (within the remuneration system of external agency), which can be exchanged for various products. All the data were collected online.

2.2 Materials

At the beginning of the survey, after informed consent, participants were asked to indicate their basic demographic variables, such as gender, age, education. It also served as a quota variable for the external agency. Consequently, participants were randomly assigned to two conditions: provide either three arguments for the benefits of migration or three arguments against migration. Financial reward for the 10 best arguments was offered.

Attitudes toward migration. Participants were asked to indicate their level of agreement with the eight statements (e.g. "Migration can effectively solve the problem with aging of European population.") on a scale from 1 (totally disagree) to 7 (totally agree). All items were recoded so that higher scores indicated more positive attitudes toward migration, while lower scores indicated a negative attitude toward migration. The scale measures attitudes toward economic migrants as well as refugees, so before creating a total mean score, we performed reliability analysis and factor analysis. Internal consistency of the scale was good (Cronbach's α = .845) and all items loaded into one single factor.

Subjective competence was be measured by a single question: Please rate your knowledge about the migration on the scale from 1 = „*I don't understand this issue at all*" to 10 = „*I understand the issue completely*".

Actual knowledge was measured by 10 single-choice questions about migration (e.g. "Which country has the largest proportion of migrants to its own population?" Answers: USA, Spain, Qatar, Germany, Lebanon, Uganda). Each correct answer received 1 point and we used the sum of correct answers as the total score of actual knowledge.

Overconfidence was operationalized as overestimation (i.e. specific type of overconfidence, (Moore & Healy, 2008) and was measured by a single question after completing the knowledge quiz: "You have just completed the quiz with 10 questions. Please, estimate the number of your correct answers." Overestimation was calculated by subtracting actual number of correct answers from the estimated number of correct answers. A positive number indicates overconfidence (higher estimated than the actual number of correct answers), while a negative number indicates underconfidence (higher actual than the estimated number of correct answers); at the same time, it indicates the magnitude of over/underestimation.

3 Results

3.1 Qualitative analysis

As a consequence of two-step qualitative coding process realized by two independent researchers, we identified 6 content categories of arguments in favour of migration (in order of incidence): 1. Cultural enrichment (18%), 2. Labour force (17%), 3. Altruism and helping (10%), 4. Population growth (7%), 5. Economic growth (5%), and 6. Reciprocity (2%).

In context of arguments against migration, we identified the following content categories (in order of incidence): 1. Fear of cultural differences (32%), 2. Security (26%), 3. Economic burden (21%), 4. Labour market (8%), 5. Health and hygiene (3%). The presented categories consisted of subcategories represented by elementary codes. In the following paragraph we will analyse these categories more in detail.

The most frequent content category of arguments against migration, (1) Cultural differences, was composed of several elementary codes/subcategories. The most numerous subcategory was related to religion (68 participants). The arguments targeted the negative aspects of religious diversity either in abstract manner or a direct one. The more specific arguments played the card of religious intolerance, and referred exclusively to Muslim religion (no other religion was explicitly mentioned), often highlighting the negative consequences of coexistence with individuals worshiping a different God. The second content subcategory related to immigrants´ supposed refusal to accommodate to new culture (62), frequently underlying newcomers lack of will to assimilate. Following set of arguments underscored the importance of differences in work ethics, morals, and sense of responsibility (30) – this subcategory was, interestingly, very frequently employed in the first argument that participants elaborated (out of the three in total), less so in other two. Lack of formal education and low standard of immigrants´ respect of women rights formed another codes used by fifteen and eight participants, respectively.

In the category of (2) Security, two elementary codes or subcategories were identified: terrorism and crime. Their proportion in relation to first, second and third argument elaborated by each participant was 16:43 (terrorism : crime) participants in the first one, 15:42 in second argument, and 8:37 in the last one, suggesting that terrorism provided less "ammunition" for the arguments and became less prominent when participants had to think, elaborate and produce the third and last argument.

Interestingly, out of all participants instructed to argue against migration (N = 338), only 4 (1.2 %) refused to provide their input; none of those mentioning unwillingness to elaborate argument contrasting with

their own attitude. The other way around the results were strikingly different. Within the group of respondents arguing in favour of migration (N=328), 87 participants (26,7%) did not provide their argument, with 73 persons (22,3%) stating that migration is a negative phenomenon. The situation where a quarter of participants refuses to provide argument against their own attitude is quite illustrative, especially when taking into account that the virtual discussion was explicitly anonymous and participants were economically incentivised (authors of 10 best arguments evaluated by the independent were awarded by 30 euro cash).

3.2 Quantitative analysis

The correlations between demographic characteristics (age, gender, and education) and attitudes toward migration, subjective competence, actual knowledge and overconfidence are in the Table 1.

Table 1. Correlational analysis of examined variables

	Gender	Age	Education
Subjective competence	-.129**	.133**	.124**
Attitudes towards migration	-.034	.053	.205**
Actual knowledge	-.198**	.092*	.192**
Overestimation	-.008	.079*	-.098*

In reference to correlations between selected variables, we have identified positive relation between education and pro-migration attitudes ($r(666) = .205$, $p < .001$), as well as negative relation between overconfidence and positive attitudes, meaning the more overconfident participant was, the more negative attitude towards migration he/she showed ($r(666) = -.151$, $p < .001$). We also found no significant correlation between education and diversity of arguments (defined by number of topics participant covered) neither against ($r(338) = -.005$, $p = .931$), nor in favour of migration ($r(328) = .020$, $p = .721$). Similarly, and interestingly, none of abovementioned content categories was significantly related to education level. We identified no significant relation between age of participants and attitudes towards migration ($r(666) = .053$, $p = .169$).

4 Discussion

The method employed allowed us to investigate the topic free from theoretical framework, and in line with the aim of thematic exploration of attitudes towards/fear of migration. As compared to self-report study, the process of generating arguments, we believe, does liberate one´s imagination and creativity while it offers a valuable insight into the socially desirable part of public discourse and persuasion. Content categories we analysed, both in context of arguments in favour and against, represented a genuine set of thematic anchors migration in public discourse is attached to. Interestingly, reliance on any of content categories in participants´ arguments (economy, security, religion, reciprocity, health, etc.) did not prove to relate to their education level.

Contrary to our expectations, education also did not relate to diversity of arguments, but as expected, it did correlate with pro-migration attitudes in line with the findings of other authors (Mayda, 2006; O'Rourke & Sinnott, 2006). These studies document that more educated individuals are less likely to have anti-immigrant sentiments, and the association between education and pro-immigrant sentiments is stronger in richer countries where natives are more skilled than the immigrants. In another study with similar aims, authors (d'Hombres & Nunziata, 2016) used data from the ESS and focused on 12 European countries finding a positive effect of education on pro-immigration attitudes in the order of 6-11 percentage points, on average. According to Hainmueller and Hiscox (2007) a large component of the link between education and attitudes toward immigrants is driven by differences among individuals in their cultural values and beliefs: more educated respondents are significantly less racist and place greater value on cultural diversity than do their less educated counterparts; they are also more likely to believe that immigration generates more benefits for the host economy. Wike and colleagues (Wike et al., 2016) report on cross-cultural pattern in Europe, where those with less education say increasing diversity makes their country a worse place to live.

Absence of significant relation between age and attitudes toward migrants could be considered as the most surprising of our results. Studies consistently demonstrate that hostility towards refugees and migrants is less prevalent among younger, politically liberal and more educated people (Dempster & Hargrave, 2017).

Conclusion

The aim of our paper was to explore the attitudes and arguments for and against migration in more depth. Besides identifying main groups of arguments for and against migration, two main findings arise from our investigation. First, the topic of migration is so emotionally strong and polarizing for many people that quarter (22.3 %) of participants refused to even think about the benefits of migration if their attitude was firmly against migration. Not only this shows strong confirmation bias in action, it illustrates why it is so

difficult to start rational debate when people are refusing to look at the other side of an issue even for the sake of argument.

Second, participants with the strongest negative attitudes toward migration tended to overestimate their knowledge the most. This is yet another reason why the debate about such complex and polarizing topics is so difficult.

It would be beneficial for the both sides in the debate to acknowledge that the opponent might have some legitimate fears and concerns while holding contrary opinion; on the other hand, inability to search for common ground and lack of openness prevent any meaningful discussions.

Acknowledgments

The study was supported by the Slovak Research and Development Agency as part of the research project APVV-20-0335 and by the scientific grant agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic as part of the project VEGA 2/0053/21.

References

d'Hombres, B., & Nunziata, L. (2016). Wish you were here? Quasi-experimental evidence on the effect of education on self-reported attitude toward immigrants. *European Economic Review*, *90*, 201–224. https://doi.org/10.1016/J.EUROECOREV.2016.02.007

Dempster, H., & Hargrave, K. (2017). *Understanding public attitudes towards refugees and migrants*. www.facebook.com/ChathamHouse

Hainmueller, J., & Hiscox, M. J. (2007). Educated Preferences: Explaining Attitudes Toward Immigration in Europe. *International Organization*, *61*(2), 399–442. https://doi.org/10.1017/S0020818307070142

Mayda, A. M. (2006). Who is against immigration? A cross-country investigation of individual attitudes toward immigrants. *Review of Economics and Statistics*, *88*(3), 510–530. https://doi.org/10.1162/REST.88.3.510

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517. https://doi.org/10.1037/0033-295X.115.2.502

O'Rourke, K. H., & Sinnott, R. (2006). The determinants of individual attitudes towards immigration. *European Journal of Political Economy*, *22*(4), 838–861. https://doi.org/10.1016/J.EJPOLECO.2005.10.005

Pew Research Center. (2018). *Eastern and Western Europeans Differ on Importance of Religion, Views of Minorities, and Key Social Issues*. https://www.pewforum.org/2018/10/29/eastern-and-western-europeans-differ-on-importance-of-religion-views-of-minorities-and-key-social-issues/

Special Eurobarometer 469. (2018). *Integration of Immigrants in the European Union* . https://www.europeanmigrationlaw.eu/documents/EuroBarometer-IntegrationOfMigrantsintheEU.pdf

Wike, R., Stokes, B., & Simmons, K. (2016). *Europeans Fear Wave of Refugees Will Mean More Terrorism, Fewer Jobs | Pew Research Center*. https://www.pewresearch.org/global/2016/07/11/europeans-fear-wave-of-refugees-will-mean-more-terrorism-fewer-jobs/

Correction rate in cognitive reflection test as a possible measure of analytical thinking dispositions

Roman Burič

Institute of Experimental Psychology, Centre of Social and Psychological Sciences, Slovak Academy of Sciences, Dúbravská cesta 9, 841 04 Bratislava, Slovakia;
roman.buric@savba.sk

Abstract

Cognitive Reflection Test (CRT) was designed to measure peoples' ability to inhibit compelling intuition and tap into analytical thinking instead. However, recent studies suggest the overall accuracy in CRT is largely dependent on the intuitive responses, and the response change after analytical thinking engagement is rather rare. This would mean the current CRT measures the ability to generate logically correct intuitions, rather than analytical thinking engagement. In this study, I tested an alternative measurement of cognitive reflection by isolating intuitive and analytical responses in CRT and computing the correction rate in cases of incorrect intuitions. This way, the correction rate reflects the response correction after engaging in analytical thinking. However, the results did not show the incremental validity of the measure when predicting reasoning accuracy. Nevertheless, intuitive accuracy in CRT emerged as its strongest correlate. Results suggest that more attention should be drawn to intuitive reasoning and predictors of logically correct intuitions.

1 Introduction

According to the classical model of dual-process theories, people automatically generate an intuitive type 1 response when reasoning, which is inhibited and replaced by a type 2 analytical response if necessary (Evans, 2007; Kahneman, 2013). This tendency to analytical thinking engagement is commonly measured with Cognitive Reflection test (CRT, Frederick, 2005). However, recent research shows that if participants correctly solve problems that cue a faulty intuition, it does not happen after the correction of the faulty intuition, but the intuitive answer is in most cases already correct (Bago & De Neys, 2017, 2019; Burič & Konrádová, 2021; Burič & Šrol, 2020). Such results were also achieved when using CRT (Bago & De Neys, 2019; Burič & Konrádová, 2021). The question arises as to what extent is the cognitive reflection truly a reflection if the intuitive answer in CRT is already correct.

2 A new way of measuring cognitive reflection

The CRT should, by definition (Frederick, 2005), measure the extent to which reasoners are prone to reflect upon their incorrect intuitive response and subsequently replace it with a logically correct response. The cognitive reflection itself then should be operationalized as a proportion of cases, in which participant generated incorrect intuitive response, but decided to change the response after deliberation took place.

To measure such response change, one needs to isolate the intuitive and deliberative responses. This leads to four possible directions of the response change - both answers incorrect (00), the first answer incorrect and the second correct (01), the first one correct and the second incorrect (10), and both answers correct (11). The ratio of the 01 direction among the cases in which the intuitive response was incorrect (01 + 00), should represent the degree of cognitive reflection.

To test the validity of such a new measure, multiple variables will be measured. The original version of the CRT was previously shown to correlate positively with numeracy, analytical thinking dispositions, negatively with intuitive thinking dispositions, and was shown to predict bias susceptibility in multiple cognitive biases. Therefore, I assume the new measure will show the same pattern of relationships with the numeracy and thinking dispositions, which should serve as an estimate of construct validity. To test incremental validity, I assume the correction rate in CRT should predict accuracy in the belief-bias task above the standard CRT and other individual difference measures.

3 Method

3.1 Participants

The data were collected by an external agency specializing in participants recruitment. The sample consisted of 150 Slovak participants. The online software Qualtrics was used to run the study.

3.2 Materials

3.1.1 Cognitive reflection test

I used five CRT items, including three from the original CRT (Frederick, 2005) with altered content and numbers (e.g., "If it takes3 printers 3 minutes to print 3 magazines, how long would it take 100 printers to print 100 magazines?"). Two other items were taken from previous research (Burič & Konrádová, 2021; Burič & Šrol, 2020). I shortened the items to be as similar in length so they could be presented via a two-response paradigm with the same time limit.

3.1.2 Berlin numeracy test

The four-item, multiple-choice format version of the Berlin numeracy test was used to measure numeracy (Cokely et al., 2012).

3.1.3 Need for cognition

To tap participants' analytic thinking disposition, a short 5–item Need for Cognition scale (NFC; Epstein et al., 1996) was used (e.g.: "I prefer complex to simple problems"). Participants were asked to rate the items of both tests on a scale from 1 ("completely uncharacteristic of me") to 5 ("completely characteristic of me").

3.1.4 Faith in intuition

To measure participants' intuitive thinking disposition, a 5–item Faith in intuition scale (NFI; Epstein et al., 1996) was used (e.g. "I believe in trusting my hunches"). Participants were asked to rate the items of both tests on a scale from 1 ("completely uncharacteristic of me") to 5 ("completely characteristic of me").

3.1.5 Conflict syllogisms

To measure reasoning accuracy and thus also the susceptibility to belief-bias, I included conflict syllogisms into the study. In the syllogistic reasoning task, participants are presented with two premises and a conclusion – and are asked to decide whether the conclusion necessarily follows from the premises or not. Conflict syllogisms are constructed in the way that the compelling intuitive response is in conflict with the logical structure of the problem, and therefore is biased. This measure was previously shown to be a reliable measure of belief bias (Bago & De Neys, 2017; Burič & Šrol, 2020).

3.3 Two-response paradigm

I isolated the intuitive response from the one based on logical principles in CRT via the so-called two-response paradigm. This paradigm is often used by scholars studying sound reasoning (Bago & De Neys, 2017; Burič & Šrol, 2020; Thompson et al., 2011; Thompson & Johnson, 2014) to separate the two types of a response when studying cognitive biases.

The two-response paradigm is based on the features of the dual processes. The dual process theories distinguish between type 1 and type 2 processes. Type 1 processes are intuitive, fast and autonomous. Type 2 processes are slower and require deliberate control (De Neys, 2018). To obtain both of them, in the two-response paradigm, participants were presented with all of the reasoning tasks twice. In the first attempt, several restrictions were used to prevent participants from engaging in type 2 processes. They were instructed to provide the first response that comes to mind and had a limited time to respond (participants had to respond immediately after reading the task). In some cases, researchers also burdened participants working memory with a secondary cognitive load (Bago & De Neys, 2017), but the time limit alone was shown to be sufficient to limit type 2 processes engagement (Burič & Šrol, 2020). In the second response, participants had an unlimited time to answer, so they could think the problems through and engage in type 2 processes.

4 Results

First, I examined the direction of answer change analysis. The results are summarized in Table 1. In line with previous studies, the most frequent direction of change was 00, meaning the response was incorrect in both the initial, and final phases. The second most common direction was 11 – the response was correct already at the initial phase. This is also in line with previous studies showing that in most cases, in which the response is correct, it is correct already intuitively (Bago & De Neys, 2017; Burič & Šrol, 2020). The 01 and 10 were the least frequent directions of change, which again show that the response change is relatively rare and the participants usually stick with their initial response.

Table 1. Frequency of the direction of response change

Direction of response change	%
00	51,6
11	21,7
01	12,7
10	14,0

To examine the relationships between the measured variables on reasoning accuracy, I run the correlation

analysis (Table 2), followed by a hierarchical regression analysis (Table 3).

Table 2. Correlation analysis

	1.	2.	3.	4.	5.	6.	7.
1.Reasoning accuracy	1						
2.CRT intuitive accuracy	.352**	1					
3. CRT final accuracy	.312**	.437**	1				
4.CRT correction rate	.094	.094	.435**	1			
5. Need for cognition	.040	.097	.216**	.072	1		
6. Faith in intuition	-.088	-.061	-.077	.009	-.156*	1	
7.Numeracy	.240**	.275**	.372**	.046	.154*	.004	1

**. Correlation is significant at the .01 level, *. Correlation is significant at the .05 level.

As can be seen from Table 2, only three variables were shown to be associated with the reasoning accuracy – intuitive and final response in CRT, and numeracy. Interestingly, the intuitive CRT response showed the strongest correlation with reasoning accuracy – even stronger than the final CRT response. However, the NFC, FI nor the CRT correction rate did not show such associations.

In a regression analysis, I wanted to examine the predictive power of the measured variable, but to look more closely at the intuitive CRT response. As I also wanted to examine the incremental validity of the new cognitive reflection measure, I added these variables in separate steps of the analysis. The results are summarized in Table 3.

Table 3. Regression analysis with reasoning accuracy as a dependent variable

	β	*t*	*p*
Step 1			
Numeracy	.143	1.37	.175
Need for cognition	.016	.15	.881
Faith in intuition	-.127	-1.26	.210
CRT final accuracy	**.212**	**2.01**	**.048**
Step 2			
Numeracy	.132	1.26	.210
Need for cognition	.010	.10	.925
Faith in intuition	-.123	-1.22	.226
CRT final accuracy	.177	1.63	.106
CRT intuitive accuracy	.139	1.33	.187
Step 3			
Numeracy	.132	1.25	.214
Need for cognition	.010	.09	.925
Faith in intuition	-.123	-1.21	.229
CRT final accuracy	.178	1.47	.145
CRT intuitive accuracy	.139	1.32	.190
CRT correction rate	-.002	-.02	.985

As can be seen, the final CRT accuracy emerged as the only significant predictor in the first step of the analysis. Overall, the model explained 8% of the variance. Even though the intuitive CRT accuracy showed the strongest relationship with the reasoning accuracy in the correlation analysis, such results were not supported after adding the variable in the second step of the regression analysis. After adding the CRT correction rate in the final step, this

variable did not emerge as a significant predictor and did not explain any additional variance.

5 Discussion

The main aim of this paper was to reflect the results of the latest studies that questioned the classical models of dual-process theories (Bago & De Neys, 2017, 2019; Burič & Konrádová, 2021; Burič & Šrol, 2020). These studies showed that the reasoning accuracy is to a large extent determined by the intuitive accuracy and the response is then rarely changed. If this is the case, then the foundations upon which the CRT was constructed are put in question as well. I tried to construct a new way of measuring cognitive reflection, that would take such findings into the account. This new measure would be based on the correction rate in the CRT, meaning it would indeed measure the participants' ability to reflect upon their response and correct their initial, intuitive answer.

However, the results did not show any predictive power of the measure on the reasoning accuracy. Such results might be a consequence of too few cases that could be analyzed. As each of the 150 participants solved 5 CRT problems twice, this results in 750 responses overall, in which direction of change could be analyzed. However, many initial responses could not be analyzed due to the time limit set in the first attempt in the two-response paradigm. This led to only 37% of cases, in which participants responded in both the initial and final stage and the direction of response change thus could be analyzed. Out of those, just 28 of cases (4.7%) were the 01 cases. This could lower the statistical power of the analysis, as p values are highly sensitive to number of analyzed cases.

Although, the attention should be also drawn to correlation analysis, in which the intuitive CRT response was the strongest correlate of reasoning accuracy. In regression analysis, it also lowered the predictive power of the final CRT response, even though none of these variables emerged as significant predictors. These results again show that the role of intuitive accuracy plays a bigger role in sound reasoning than initially thought.

Acknowledgment

This work was funded by Doctogrand No. APP0329 and APVV-20-0335.

References

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109. https://doi.org/10.1016/j.cognition.2016.10.014

Bago, B., & De Neys, W. (2019). The Smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*(3), 257–299. https://doi.org/10.1080/13546783.2018.1507949

Burič, R., & Konrádová, Ľ. (2021). Mindware instantiation as a predictor of logical intuitions in the Cognitive Reflection Test. *Studia Psychologica*, *63*(2), 114–128. https://doi.org/10.31577/sp.2021.02.822

Burič, R., & Šrol, J. (2020). Individual differences in logical intuitions on reasoning problems presented under two-response paradigm. *Journal of Cognitive Psychology*, 1–18. https://doi.org/10.1080/20445911.2020.1766472

Cokely, E. T., Galesic, M., Schulz, E., Garcia-Retamero, R., & Ghazal, S. (2012). Measuring Risk Literacy: The Berlin Numeracy Test. *Judgment and Decision Making*, *7*(1), 23.

De Neys, W. (Ed.). (2018). *Dual process theory 2.0* (1 Edition). Routledge.

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive-experiential and analytical-rational thinking styles. *Journal of Personality and Social Psychology*, *71*(2), 390–405. https://doi.org/10.1037//0022-3514.71.2.390

Evans, J. St. B. T. (2007). On the resolution of conflict in dual process theories of reasoning. *Thinking & Reasoning*, *13*(4), 321–339. https://doi.org/10.1080/13546780601008825

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Kahneman, D. (2013). *Thinking, Fast and Slow* (1st edition). Farrar, Straus and Giroux.

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, *20*(2), 215–244. https://doi.org/10.1080/13546783.2013.869763

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Tréning kapacity vizuálnej pracovnej pamäti v prostredí virtuálnej reality

Barbora Cimrová (a, b), Martin Marko (a, b), Igor Farkaš (a),
Branislav Sobota (c), Štefan Korečko (c)
Zuzana Rošťáková (d), Roman Rosipal (d)

(a) Fakulta matematiky fyziky a informatiky, Univerzita Komenského v Bratislave
(b) Centrum experimentálnej medicíny, Slovenská akadémia vied, Bratislava
(c) Fakulta elektrotechniky a informatiky, Technická univerzita v Košiciach
(d) Ústav merania Slovenskej akadémie vied, Bratislava
barbora.cimrova@fmph.uniba.sk

Abstrakt

V naše štúdii sme skúmali možnosť kognitívneho tréningu v prostredí virtuálnej reality (VR). Zamerali sme sa na pracovnú pamäť, ktorá patrí medzi kľúčové kognitívne funkcie, dôležité napríklad pre riešenie problémov, rozhodovanie, učenie sa novým zručnostiam a podobne. Je známe, že jej kapacita sa môže vplyvom rôznych tréningových stratégií meniť. Podarilo sa nám navrhnúť a implementovať protokol s narastajúcou kognitívnou záťažou prostredníctvom komplexnej aplikácie v prostredí VR. Cieľom bolo zlepšenie schopnosti filtrácie irelevantných distrakčných podnetov vo vizuálno-priestorovej pracovnej pamäti. Úspešnosť sme preverovali na behaviorálnej úrovni pomocou štandardného testu detekcie zmeny a na fyziologickej úrovni prostredníctvom snímania mozgovej aktivity participantov a vyhodnotenia neurálnych korelátov kapacity vizuálnej pracovnej pamäti. Naše výsledky potvrdili, že tréning indukoval monotónne zvyšujúcu sa kognitívnu záťaž, no napriek desaťdňovému tréningu sme v skúmaných behaviorálnych mierach nenašli očakávaný efekt. Výsledky analýzy elektroencefalografických mier však naznačujú, že k zlepšeniu schopnosti filtrácie mohlo u experimentálnej skupiny dôjsť už po prvej fáze kognitívneho tréningu.

1 Tréning kognitívnych funkcií vo VR

Virtuálna realita (VR), ktorú môžeme definovať ako typ rozhrania užívateľa a počítača poskytujúca virtuálnu simuláciu prostredia alebo aktivity v reálnom čase, predstavuje pokrokovú technológiu, ktoré umožňuje dizajnérom vytvoriť bohaté a imerzívne virtuálne prostredie so širokým rozšírením možností, ktoré sú v bežnom živote obmedzené (Adamovich a spol., 2009). V ostatných rokoch zaznamenáva jej využitie exponenciálny rast jednak v počte užívateľov, ale rovnako aj v šírke spektra eventuálnych aplikácií (Xiong a spol., 2021). Jednou zo zaujímavých otázok je prenositeľnosť tréningu vo VR do schopností v reálnom živote. V našej štúdii sme sa preto zamerali na preskúmanie tréningu kognitívnych schopností vo VR.

Pracovná pamäť tvorí zložku tzv. fluidnej inteligencie (Li a spol., 2021) a je známe, že tréningom je možné dosiahnuť zvýšenie jej kapacity (Jones a spol., 2021). Nedostatočnou schopnosťou filtrácie irelevantných podnetov môže dôjsť k zníženiu kapacity pracovnej pamäti jej zahltením. Vizuálna priestorová pracovná pamäť je tiež jednou z mála kognitívnych funkcií, ktoré majú známy elektrofyziologický neurálny korelát. Konkrétne ide o kontralaterálnu oneskorenú negativitu snímanú prostredníctvom elektroencefalografu (EEG) z mozgovej hemisféry opačnej oproti polovici zorného poľa, v ktorom je prezentovaný podnet, ktorý si má meraný subjekt po určitú dobu udržať vo vizuálnej priestorovej pracovnej pamäti. Veľkosť takejto negatívnej výchylky EEG koreluje s počtom položiek, ktoré si subjekt dokázal v pamäti udržať (Vogel, 2005). Existencia takýchto známych korelátov umožňuje posúdiť aj oveľa jemnejšie zmeny, ako by boli postrehnuteľné behaviorálnymi mierami.

2 Metódy

2.1 Participanti

Na štúdii sa zúčastnilo 30 zdravých dobrovoľníkov, študentov vysokej školy. Medzi vylučovanie kritériá patrili neurologické a psychiatrické diagnózy, užívanie návykových látok, poruchy zraku, ľavorukosť či ambidextria (obojručnosť). Do experimentálnej skupiny sme zaradili 14 participantov, z toho 10 mužov (priemerný vek ± SD bol 21,2 rokov ± 1,2 roky). V kontrolnej

skupine bolo 16 účastníkov, z čoho 13 boli muži (priemerný vek ± SD bol 22 rokov ± 1,8 rokov).

2.1 Tréning v prostredí VR

Pre naše účely sme použili na mieru vytvorenú hru implementovanú do prostredia virtuálnej reality v systéme CAVE, ktorý pozostáva z 20-ich LCD obrazoviek usporiadaných do plochy okolo hráča a špeciálnych okuliarov, vďaka ktorým sa okolo participanta vytvorí imerzívne virtuálne 3D prostredie (Korečko, Hudák, Sobota, 2019). Hra nazvaná Tower defence pozostávala z blokov s adaptívne sa stupňujúcou náročnosťou, čo umožnilo postupné zlepšovanie sa v hre a teda personalizovaný tréning kognitívnych schopností (Korečko a spol., 2018). Každý blok sa skladal z desiatich opakovaní.

Úlohou participanta bolo pomocou ovládača v ruke (joystick) zamerať a zostreliť (označiť) približujúce sa cieľové objekty („nepriateľské drony“), ktoré nalietavali z rôznych strán. To bolo možné až keď boli vo vzdialenosti „na dostrel“, pričom krátko predtým na určitý čas zmizli a ich polohu si bolo nutné pamätať. Náročnosť odpovedala rýchlosti a počtu cieľových objektov, objektov, ktoré nemali byť zostrelené (tzv. distrakčné objekty, „priateľské drony“) a participant mal ich prítomnosť. Samotný protokol tréningu bol zostavený z desiatich tréningových sedení v priebehu dvoch týždňov.

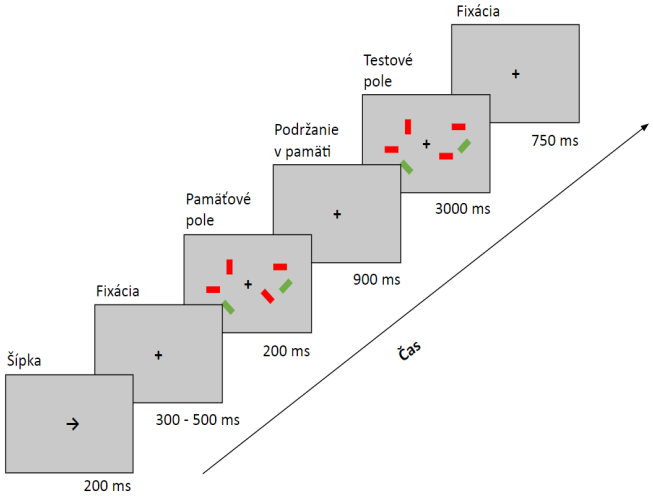
2.3 Kognitívne testovanie

Efekt tréningu sme posudzovali podľa výkonu v úlohe meranej mimo prostredia CAVE, ktorá bola administrovaná trikrát: v prvý deň pred začatím tréningu, po piatom dni tréningu a po desiatom dni, teda po ukončení trénovania. Použili sme úlohu detekcie zmeny (UDZ), ktorá je považovaná sa štandardný marker kapacity priestorovej pracovnej pamäti (Repovš & Baddeley, 2006).

UDZ pozostávala zo 640 opakovaní. Každé opakovanie sa začínalo zobrazením šípky, ktorá naznačovala, na ktorú stranu má subjekt zamerať pozornosť, a to bez presunutia zraku z fixačného bodu uprostred obrazovky – čím bola zabezpečená lateralizovaná prezentácia vždy do jednej hemisféry nevyhnutná pre výpočet CDA. Nasledovalo pamäťové pole s dvoma až štyrmi cieľovými a žiadnym alebo dvoma distrakčnými podnetmi v každej polovici zorného poľa. Po intervale 900 ms, kedy bolo nutné držať v pamäti orientáciu cieľových podnetov (z cieľovej polovice zorného poľa), nasledovalo testové pole. Úlohou participanta bolo stlačením tlačidla odpovedať, či došlo k zmene orientácie cieľového podnetu.

2.4 Snímanie aktivity mozgu

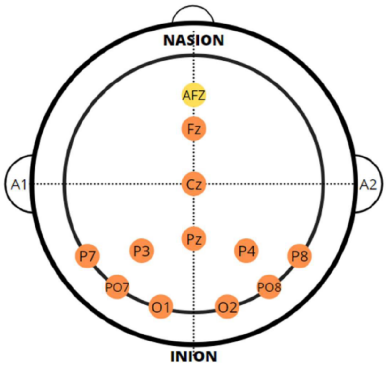
Na posúdenie jemnejších zmien, ktoré by zatiaľ nemuseli byť merateľné behaviorálnymi mierami, sme počas UDZ zaznamenávali aj elektroencefalografickú aktivitu (EEG) z posteriórnych oblastí mozgu. Snímacie elektródy boli umiestnené podľa medzinárodného systému 10-20 na oblastiach zobrazených na obr. 2.



Obr. 1. Schéma úlohy na detekciu zmeny. Cieľové (červené) a distrakčné (zelené) podnety sa zobrazia náhodne v jednej zo 4 možných orientácií. Šípka určuje cieľovú polovicu zorného poľa. V tomto príklade šikmý červený obdĺžnik v pravom zornom poli zmenil orientáciu na horizontálnu, takže došlo k zmene.

2.5 Analýza mozgovej aktvity

Záznamy korelát kapacity vizuálnej pracovnej pamäti, teda počtu položiek držaných v mysli po dobu nevyhnutnú na splnenie úlohy, CDA, sme vypočítali pre každého participanta, ako rozdiel priemernej aktivity z hemisféry kontralaterálnej (na opačnej strane) k zornému poľu, v ktorom boli prezentované cieľové podnety a priemernej aktivity z hemisféry ipsilaterálnej (na rovnakej strane), ako cieľové podnety.



Obr. 2. Schematické znázornenie umiestnenia elektród na hlave participanta. Snímacie elektródy (znázornené oranžovou) boli umiestnené nad záhlavnou (okcipitálnou) a temennou (parietálnou) oblasťou a v stredovej (mediálnej) línii. Zemniaca elektróda bola v prefrontálnej stredovej oblasti (AFz) a referenčné elektródy boli umiestnené na ušných lalôčikoch (A1 a A2).

3 Výsledky

3.1 Behaviorálne dáta z tréningu

Zo samotnej tréningovej hry sme získavali parametre odrážajúce výkon účastníkov, medzi ktoré patrilo skóre definované ako súčet správnych zásahov a (správne) nezasiahnutých distraktorov, ďalej presnosť definovaná ako rozdiel medzi relatívnou mierou zásahov a relatívnou mierou zasiahnutých distraktorov (falošné alarmy), potom schopnosť detegovať signál, vyjadrená rozdielom z-transformovaných hodnôt pre relatívnu mieru zásahu a falošné alarmy a nakoniec parameter výkonu, ktorý zohľadňuje stupňujúcu sa náročnosť a je vypočítaný ako súčin náročnosti úlohy a presnosti participantov.

Priemerná presnosť v úlohe bola relatívne vysoká (90.3%), avšak v priebehu tréningu sa mierne znižovala. Podobný trend ukázali aj výsledky pre schopnosť detegovať signál, čo je možné odôvodniť zvyšujúcou sa náročnosťou úlohy počas tréningu, ktorá bola definovaná ako celkový počet zobrazených cieľov a distraktorov (T+D) v jednotlivých úrovniach úlohy. Čo je dôležité, absolútne skóre v úlohe sa u participantov počas desiatich tréningových sedení postupne zvyšovalo, čo odzrkadľuje (očakávaný) efekt tréningu. Pre zohľadnenie náročnosti sme analyzovali aj parameter výkonu, ktorý bol vypočítaný ako súčin náročnosti úlohy a presnosti participantov. Výkon v úlohe naprieč tréningovými sedeniami rástol.

Na základe analýzy behaviorálnych mier sledovaných počas tréningu v Tower Defence môžeme skonštatovať, že sa nám podarilo vytvoriť tréningový protokol s adaptívne sa zvyšujúcou kognitívnou záťažou pre participantov. Z povahy indukovanej záťaže je možné predpokladať, že tréning predstavoval čoraz väčšiu kognitívnu výzvu, resp. tréningový potenciál, pre relevantné funkcie pozornosti a pracovnej pamäti, na ktoré sa tréning zameriaval (detekcia signálu, filtrovanie distraktorov, aktualizácia a udržiavanie vizuálno-priestorovej reprezentácie objektov v pamäti). Napriek rastúcej kognitívnej záťaži v tréningu boli participanti schopní reagovať na čoraz väčšie množstvo podnetov, za čoraz kratší čas, pričom sa im darilo udržať relatívne vysokú mieru presnosti a dobrý výkon. Tieto výsledky naznačujú, že kognitívny tréning mal potenciál stimulovať neurokognitívne okruhy podporujúce pracovnú pamäť a kontrolu pozornosti.

3.1 Kapacita pracovnej pamäti v UDZ

Analýza kapacity vizuálnej pracovnej pamäti pre kontrolnú aj experimentálnu skupinu dobrovoľníkov bola vykonaná pomocou modelu so zmiešanými efektami (LMEM). Presnosť v UDZ sa znížila pri vyššom počte cieľov (2 verzus 4 ciele, $F(1,308) = 1157{,}766$; $p < .001$) a v prítomnosti distraktorov (0 verzus 2 distraktory, $F(1,308) = 14{,}518$; $p < .001$). Tieto hlavné efekty sú v predpokladanom smere a poukazujú na skutočnosť, že pri vyššom počte relevantných a irelevantných podnetov bola úloha náročnejšia, teda validujú dizajn úlohy.

V súvislosti s hlavnou hypotézou, t.j. či tréning bude viesť k zvýšeniu kapacity vizuálnej pracovnej pamäti, štvorfaktorová analýza ANOVA nepreukázala žiadny štatisticky významný vzťah: faktor opakovaného merania nebol významný ani ako hlavný faktor ($F(1,308) = 0{,}731$; $p = 0{,}482$), ani v interakcii s ostatnými faktormi, ktoré boli zahrnuté v modeli. Predpokladaná interakcia faktorov skupina × meranie nebola štatisticky významná ($F(2,308) = 0{,}435$; $p = 0{,}648$) a významnosť sa nepreukázala ani pri zohľadnení počtu cieľov a distraktorov ($F(2,308) = 0{,}738$; $p > 0{,}479$) pre interakcie vyššieho rádu).

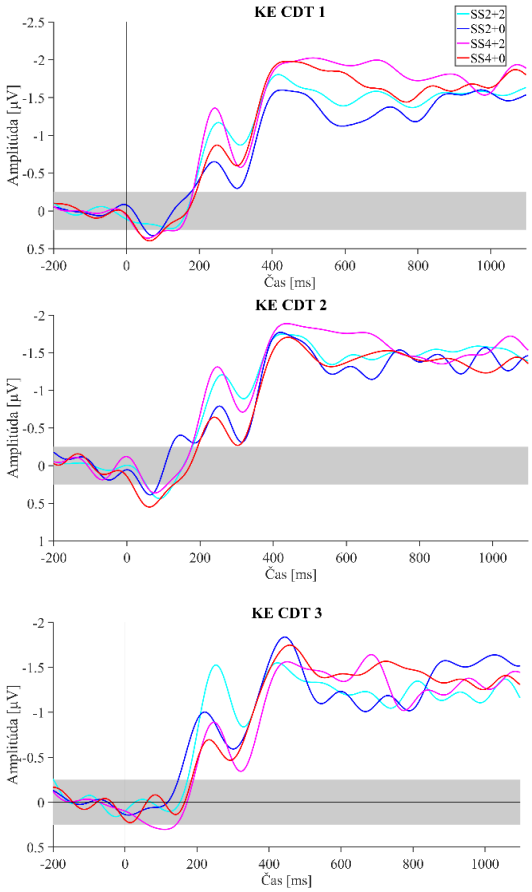
Tieto výsledky ukázali, že presnosť participanta v CDT úlohe závisí od počtu podnetov a distraktorov, čo potvrdzuje validitu tejto experimentálnej úlohy. Očakávané zlepšenie v presnosti vizuálnej pracovnej pamäti v tréningovej skupine naprieč troma meraniami sa však nepreukázalo. Podobne ako v kontrolnej skupine, skupina s tréningom nevykazovala zlepšenie vizuálnej pracovnej pamäti v čase. Možnými vysvetleniami absencie účinku tréningu môže byť krátkosť (resp. nedostatočná intenzita) tréningového programu, rozdiely medzi tréningovou úlohou a CDT, menší rozsah výskumnej vzorky, alebo možnosť, že behaviorálna úroveň posudzovania účinkov tréningu nebola dostatočne senzitívna, aby sa prejavili.

3.1 Elektrofyziologické dáta

Namerané EEG záznamy z oboch skupín sme spracovali v prostredí programu Brain Vision Analyzer (BVA2.0) a

MATLAB (R2019b). Ukážky výsledných CDA kriviek pre obe skupiny sú na obr. 3 a 4, kde vidieť charakteristickú oneskorenú negativitu s latenciou v rozsahu 400 až 900 ms od zobrazenia podnetov, ktoré si mal participant podržať v pracovnej pamäti.

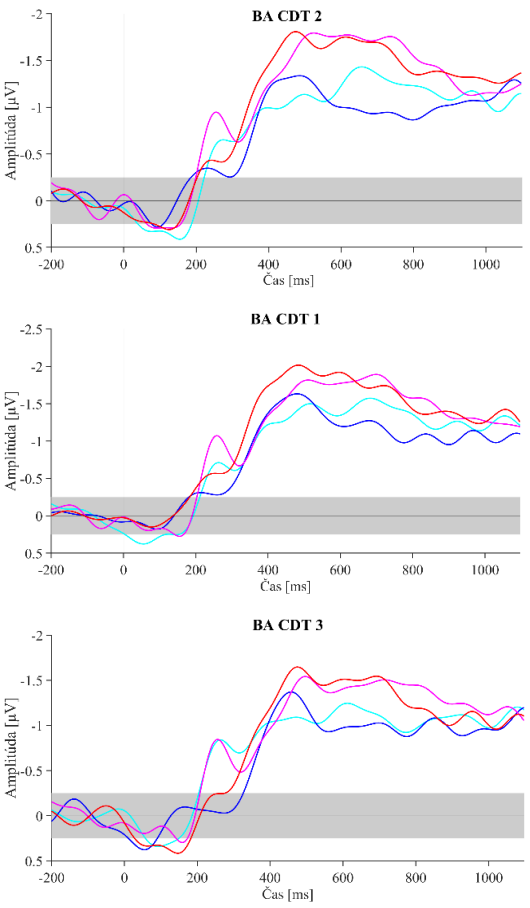
Amplitúda tejto vlny odráža počet položiek držaných v pracovnej pamäti z kontralaterálneho zorného poľa. Farbou sú kódované situácie podľa počtu pamäťových a distrakčných položiek. V kontrolnej (BA) skupine (obr.4) vidíme, že vzájomné priebehy kriviek (rozdielne farby - rôzny počet podnetov a distraktorov) sa naprieč meraniami (prvý, druhý a tretí graf = deň merania) takmer nemenia. Naopak, v experimentálnej (KE) skupine (obr. 3) vidíme trend naznačujúci, že uprostred (CDT2 - druhý graf) a na konci experimentu (CDT3 - posledný graf) sa vlny SS4/2 a SS4/0 (štyri podnety s a bez distraktorov), ako aj SS2/2 a SS2/0 (dva podnety s a bez distraktorov) viac k sebe približujú v porovnaní s meraním na začiatku experimentu (CDT1 - prvý graf). Vidíme, že zníženie rozdielu možno pozorovať už po prvej fáze tréningu.



Obr. 3. Výsledné krivky kontralaterálnej oneskorenej aktivity (CDA) v experimentálnej skupine meranej v Košiciach.

4 Záver

Podarilo sa nám nadizajnovať a implementovať tréningový protokol vo virtuálnej realite prostredia CAVE pre trénovanie schopnosti filtrácie dôležitej pre funkciu priestorovej pracovnej pamäti. Napriek 10-dňovému tréningu sme však nepozorovali očakávané zlepšenie presnosti detegovania zmien v kognitívnom teste UDZ: presnosť detekcie signálov sa medzi meraniami významne nezmenila. Výsledky analýzy elektroencefalografických mier CDA však poukazujú v experimentálnej skupine na trend naznačujúci zlepšenie schopnosti filtrácie už po prvej fáze kognitívneho tréningu v prostredí CAVE.



Obr. 4. Výsledné krivky kontralaterálnej oneskorenej aktivity (CDA) v kontrolnej skupine meranej v Bratislave.

Literatúra

Jones, J. S., Adlam, A. L. R., Benattayallah, A., & Milton, F. N. (2021). The neural correlates of working memory training in typically developing children. *Child Development*, https://doi.org/10.1111/cdev.13721

Korečko Š., Hudák M., Sobota B. (2019). LIRKIS CAVE: Architecture, Performance and Applications. *Acta Polytechnica Hungarica*, 16(2), 199-218.

Korečko Š., Hudák M., Sobota S., Marko M., Cimrová B., Farkaš I., Rosipal R. (2018). Assessment and training of visuospatial cognitive functions in virtual reality: proposal and perspective. In: *Proceedings of 9th International Conference on Cognitive Infocommunications (CogInfoCom),* Budapest, str. 39-43, Danvers: IEEE.

Li, G., Chen, Y., Le, T. M., Wang, W., Tang, X., & Li, C. S. R. (2021). Neural correlates of individual variation in two-back working memory and the relationship with fluid intelligence. *Scientific Reports*, 11(1), 1-13.

Repovš, G., & Baddeley, A. (2006). The multi-component model of working memory: Explorations in experimental cognitive psychology. *Neuroscience*, 139(1), 5–21.

Vogel, E. K., McCollough, A. W., & Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature*, 438(7067), 500–503.

Xiong, J., Hsiang, E. L., He, Z., Zhan, T., & Wu, S. T. (2021). Augmented reality and virtual reality displays: emerging technologies and future perspectives. *Light: Science & Applications*, 10(1), 1-30.

Zvyšovanie efektivity trénovania a kapacity v atraktorovom neurálnom modeli asociatívnej pamäte

Matej Fandl
Fakulta matematiky, fyziky a informatiky UK
Mlynská dolina, Bratislava
matej.fandl@fmph.uniba.sk

Martin Takáč
Fakulta matematiky, fyziky a informatiky UK
Mlynská dolina, Bratislava
takac@ii.fmph.uniba.sk

Abstrakt

Hopfieldove siete sú známy model asociatívnej pamäte. Ich moderné varianty vieme interpretovať ako siete so skrytou vrstvou latentných neurónov, ktoré slúžia ako detektory naučených pamäťových vzorov. Takéto siete dovoľujú jednoduchý spôsob trénovania, kde pre každý vzor, ktorý chceme do pamäte uložiť, pridáme jeden latentný neurón. Problém takéhoto riešenia je rast časovej a priestorovej zložitosti modelu s počtom zapamätaných vzorov. Predkladaná práca opisuje proces hľadania modelu asociatívnej pamäte inšpirovaného architektúrou moderných Hopfieldových sietí, v ktorom bude proces trénovania viesť k deľbe práce medzi neurónmi na skrytej vrstve a tvorbe distribuovaných reprezentácií.

1 Úvodná kapitola

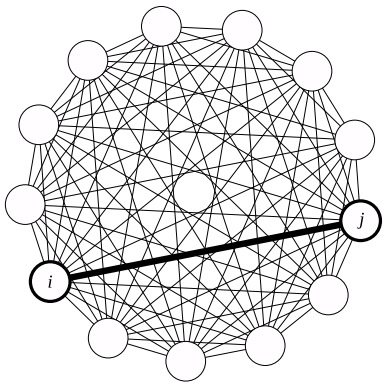
Asociatívna (obsahovo adresovateľná) pamäť je taký typ pamäte, ktorej obsah získavame nie dopytovaním sa pomocou konkrétneho kľúča, alebo adresy, ale prezentáciou zašumeného, či nekompetného vzoru. Obsah z pamäte teda získavame pomocou obsahu samotného. Podľa typu takejto pamäte, jej výstupom po prezentácii vzoru je buď rekonštruovaný vstup (autoasociatívna pamäť), alebo iný vzor, ktorý je asociovaný s prezentovaným vzorom (heteroasociatívna pamäť). Schopnosť tvoriť si asociácie a pracovať s nimi je vlastnosť kritická v ľudskej kognícii a snaha simulovať túto schopnosť v umelých systémoch viedla k tvorbe rôznorodých modelov asociatívnej pamäte (Steinbuch, 1961; Hopfield, 1982, 1984; Acevedo-Mosqueda a spol., 2013; Krotov a Hopfield, 2016; Ramsauer a spol., 2020; Krotov, 2021).

Cieľom našej práce je vyvinúť model asociatívnej pamäte, ktorý v sebe skombinuje možnosť efektívneho trénovania a vysokej kapacity. Využitie takéhoto modelu vidíme pri problémoch, ktoré vyžadujú klasifikáciu a rekonštrukciu veľkého počtu rôznych vzorov. Pre diskusiu o použití veľkej asociatívnej pamäte pri identifikácii sekvencií imunitného repertoáru (Widrich a spol., 2020), ale aj iných problémoch, viď Krotov a Hopfield (2020).

V článku opíšeme Hopfieldove siete, tak originálny model (Hopfield, 1982), ako aj moderné varianty, ktoré riešia jeho problémy s kapacitou. Tieto modely slúžia ako základná inšpirácia našej práce.

2 Hopfieldove siete

Hopfieldove siete (Obr. 1) sú plne rekurenté jednovrstvové neurónové siete vykazujúce atraktorovú dynamiku (O'Reilly a Munakata, 2000; Strogatz, 2018). Tieto siete sú známym, dobre preskúmaným modelom asociatívnej pamäte.



Obr. 1: Hopfieldova sieť s 13 neurónmi. Všetky neuróny sú vzájomne prepojené symetrickými synapsami. Na obrázku sú zvýraznené neuróny i a j.

2.1 Originálny model

Na vstupe originálneho modelu (Hopfield, 1982) je binárny vektor $\boldsymbol{x} \in \{-1, 1\}^d$. Stav siete $\boldsymbol{\xi} \in \{-1, 1\}^d$ zodpovedá aktuálnym aktiváciám jednotlivých neurónov. Úprava stavu i-teho neurónu ξ_i závisí od stavov všetkých ostatných neurónov v sieti a synaptických váh, ktoré ich prepájajú:

$$\xi_i = sign\left(h_i\right) \tag{1}$$

$$h_i = \sum_{j=1}^{d} w_{ij}\xi_j - \theta_i \tag{2}$$

w_{ij} je váha synapsy prepájajúcej neuróny i a j, θ_i určuje excitačný prah neurónu i. Tento má zvyčajne hodnotu 0. Váhy jednotlivých synáps určujú, aký je očakávaný vzťah aktivácií príslušnej dvojice neurónov:

$$w_{ij}^k = \begin{cases} x_i^k x_j^k & i \neq j \\ 0 & i = j \end{cases} \tag{3}$$

$$w_{ij} = \sum_{k=1}^{N} w_{ij}^k, \tag{4}$$

kde $k \in \langle 1, N \rangle$ je index pamäťového vzoru. Pre konkrétny vstup k a dvojicu neurónov u ktorých očakávame rôzne hodnoty aktivácie dostávame zápornú váhu. Inak je váha kladná. V každom stave vieme vypočítať energiu siete E nasledovne:

$$E = -\frac{1}{2}\boldsymbol{\xi}^T \boldsymbol{W}\boldsymbol{\xi} = -\frac{1}{2}\sum_{i,j}^{d} \xi_i w_{ij} \xi_j \tag{5}$$

Položením aktivácií jednotlivých neurónov hodnotám rovným hodnotám prislúchajúcich komponentov vstupného vektora ($\xi_i = x_i$) sa začne spontánna aktivita siete. Stav $\boldsymbol{\xi}$ sa približuje k jednému z bodových atraktorov, ktorý je lokálnym energetickým minimom. Akonáhle sa stav prestane meniť, považujeme ho za rekonštruovaný vzor. Pri aplikovaní pravidla (1) si môžeme zvoliť synchrónnu a asynchrónnu dynamiku. Od použitia konkrétnej dynamiky závisí úspešnosť rekonštrukcie. Pri synchrónnej verzii môže úprava stavu skončit v limitnom cykle a teda nikdy nedosiahne bodový atraktor. Pre detaily viď (Hopfield, 1982).

2.1.1 Vlastnosti originálneho modelu

Pozitívami originálneho modelu sú jednoduchosť, lokalistické (párové) interakcie, učenie bez učiteľa a rýchlosť samotného učenia. Naučiť sieť nový vzor k pomocou pravidla (4) znamená iba pripočítať hodnoty matice váh $\boldsymbol{W}^k$ k aktuálnej matici váh $\boldsymbol{W}$. Ide teda o jednokrokové (one-shot) učenie.

Hlavnou nevýhodou je nízka kapacia siete ($C \cong 0.14d$ ak akceptujeme malú rekonštrukčnú chybu). Aj v prípadoch, kedy táto kapacita stačí, môže v sieti dôjsť k tvorbe falošných atraktorov, čo sú lokálne minimá stavového priestoru zložené z viacerých blízkych pamäťových vzorov. Ak sa vstup siete nachádza v spádovej oblasti falošného atraktora, stav siete konverguje doň a rekonštrukcia je neúspešná. Aj keď tvorba falošných atraktorov je často považovaná za nedostatok, Gorman a spol. (2017) ukázali, že ich v niektorých prípadoch vieme považovať za prototypy trénovacej množiny pozostávajúcej zo zašumených vzorov.

2.2 Moderné Hopfieldove siete

Krotov a Hopfield (2016) vytvorili moderný binárny variant Hopfieldovej siete nazývaný Hustá Asociatívna Pamäť (Dense Associative Memory, DAM), ktorý adresuje vyššie spomínané problémy originálneho modelu. Autori navrhli novú energetickú funkciu

$$E = -\sum_{k=1}^{N} F\left(\boldsymbol{x^k}^T \boldsymbol{\xi}\right), \tag{6}$$

kde F je interakčná funkcia, napríklad $Fa = a^n$, umožňujúca zvýraznenie rozdielov medzi podobnými pamäťovými vzormi. Pre $n = 2$ dostaneme originálny Hopfieldov model. Čím vyššie n zvolíme, tým výraznejšie prispievajú rozdiely medzi jednotlivými vzormi k celkovej energii siete E. Z upravenej energetickej funkcie odvodili pravidlo na výpočet nového stavu siete, ktoré vedie k exponenciálnej kapacite vzhľadom na dimenzionalitu vstupného priestoru d.

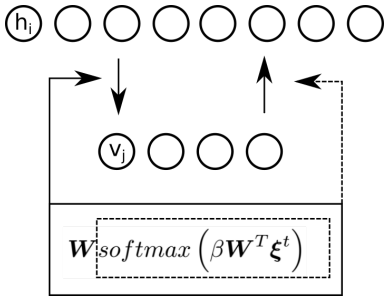
Demircigil a spol. (2017) opísali použitie DAM s exponenciálnou interakčnou funkciou $F(x) = e^x$, neskôr Ramsauer a spol. (2020) z rovnice (6) a tejto interakčnej funkcie vytvorili moderný variant Hopfieldovej siete (Continuous modern Hopfield Network, CMHN), ktorý pracuje s vektormi spojitých hodnôt, teda $\boldsymbol{x} \in \mathbb{R}^d$. Pravidlo na výpočet nového stavu CMHN má tvar

$$\boldsymbol{\xi}^{t+1} = \boldsymbol{W} softmax\left(\beta \boldsymbol{W}^T \boldsymbol{\xi}^t\right), \tag{7}$$

kde β sa dá chápať ako inverzná teplota a slúži na zvýraznenie alebo potlačenie rozdielov medzi jednotlivými pamäťovými vzormi pri rekonštrukcii.

Spomínané moderné modely dosahujú vysokú kapacitu vďaka tomu, že interakcie medzi jednotlivými neurónmi už nie sú lokalistické (Krotov a Hopfield, 2020), ale v interakcii je viacero neurónov. Krotov a Hopfield (2020) poukázali na interpretáciu moderných Hopfieldových sietí ako sietí s jednou skrytou vrstvou a lokalistickymi interakciami.

Obr. 2 znázorňuje túto interpretáciu. Aktivácie neurónov na viditeľnej vrstve predstavujú jednotlivé príznaky (dimenzie vstupného priestoru) a sú transformované na aktivácie neurónov na skrytej vrstve, ktoré slúžia ako detektory. Hodnota aktivácie neurónu na skrytej vrstve predstavuje pravdepodobnosť, s ktorou tento neurón reprezentuje aktuálny vstup. Tieto aktivácie sú potom použité pre výpočet aktivácií neurónov na viditeľnej vrstve pomocou tých istých váh. Aktivácia každého neurónu na viditeľnej vrstve je lineárnou kombináciou komponentov váhových vektorov neurónov na skrytej vrstve.



Obr. 2: Interpretácia mechanizmu pravidla (7) ako siete s jednou skrytou vrstvou. Aktivácie neurónov $\boldsymbol{v}$ na viditeľnej vrstve sú transformované na aktivácie neurónov na skrytej vrstve $\boldsymbol{h}$ (prerušovaná čiara). Tie sú potom transformované rovnakými váhami na aktivácie neurónov na viditeľnej vrstve (plná čiara).

Zaujímavé sú korešpondencie CMHN s inými modelmi. Ramsauer a spol. (2020) ukázali ekvivalenciu s mechanizmom pozornosti v neurónových sieťach typu transformer (Vaswani a spol., 2017). Takáč a spol. (2020) zasa ponúkli bayesovskú interpretáciu samoorganizujúcich sa máp (SOM). SOM sú v tejto interpretácii použiteľné aj na rekonštrukciu vzorov a ich mechanizmus rekonštrukcie korešponduje s mechanizmom výpočtu nového stavu v CMHN (7.)

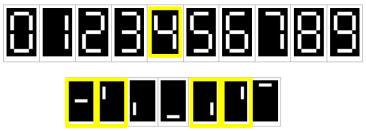
2.2.1 Učenie v spojitých moderných Hopfieldových sieťach

Pri trénovaní tejto siete máme viacero možností. Môžeme nastaviť $\boldsymbol{W} = \boldsymbol{X}$, teda vytvoriť pre každý vzor, ktorý chceme do pamäte uložiť, jeden latentný neurón. Váhové vektory takýchto neurónov budú rovné vstupnému vektoru, ktorý tento neurón reprezentuje. Každý latentný neurón v tomto prípade reprezentuje jeden konkrétny exemplár. Tento spôsob učenia má tú výhodu, že ide o jednokrokové učenie podobne ako v prípade originálneho Hopfieldovho modelu. Sieť pri rekonštrukcii v závisloti od nastavenia inverznej teploty buď konverguje k najbližšiemu vzoru (pre vysoké hodnoty β), alebo kombinácii viacerých najbližších vzorov. Ďalšou výhodou tohto spôsobou je inkrementálnosť učenia. Natrénovaný model vieme jednoducho doučiť aj mimo pôvodnej trénovacej fázy bez toho, aby sme sa stretli s katastrofickým zabúdaním, teda s poškodeňim až zničením už naučených infromácií. Nevýhodou tohto prístupu je časová a výpočtová zložitosť rekonštrukcie rastúca s počtom uložených pamäťových vzorov.

Iná možnosť trénovania, ktorú máme, je zvoliť fixný počet latentných neurónov a trénovať váhy známymi metódami, akými je napríklad spätné šírenie chyby (keďže ide o učenie s učiteľom, cieľovým výstupom siete je jej vstup). Táto metóda trénovania môže viesť k tvorbe efektívnejších reprezentácií na skrytej vrstve, problém ale je, že proces trénovania je časovo a výpočtovo náročný. Stratíme tým možnosť jednokrokového, prípadne málokrokového, trénovania. Lillicrap a spol. (2020) opisujú ďalšie problémy trénovania neurónových sietí spätným šírením chyby, vrátane problémov týkajúcich sa biologickej plauzibilnosti algoritmu.

3 Deľba práce neurónov na skrytej vrstve

Deľbou práce neurónov na skrytej vrstve vieme dosiahnuť distribuované reprezentácie vstupných vzorov. To znamená, že jednotlivé latentné neuróny nebudú detekovať konkrétny exemplár ako v prípade jednoduchého nastavenia $\boldsymbol{W} = \boldsymbol{X}$ spomínaného v predchádzajúcej sekcii. Budú detekovať časti vzorov z ktorých sa rekonštruované vzory dajú vyskladať. Obr. 3 ilustruje idealizovaný prípad distribuovanej reprezentácie na jednoduchej vstupnej množine digitálnych čísel.



Obr. 3: Príklad distribuovanej reprezentácie čísla 4. Prvý riadok ukazuje trénovaciu množinu jednoduchých digitálnych čísel. Druhý riadok ukazuje časti, z ktorých je každé číslo z pôvodnej množiny vyskladateľné. Žltou farbou je vyznačený konkrétny príklad.

Pozitívna hypotetická vlastnosť modelu s deľbou práce na skrytej vrstve je generalizácia - schopnosť rekonštruovať vzory, ktoré neboli použité pri trénovaní, ale ktoré sa skladajú z častí, pre ktoré si model už vytvoril detektory.

Zatiaľ čo trénovaním spätným šírením chyby vieme takúto deľbu práce medzi latentnými neurónmi dosiahnuť, stratíme ním možnosť rýchleho a inkrementáleho učenia, ako aj lokalistické interakcie (Lillicrap a spol., 2020). Vo zvyšku príspevku opíšeme prístup k hľadaniu pravidla, ktoré kombinuje dobré vlastnosti oboch spomenutých spôsobov.

4 Cieľové vlastnosti nášho modelu

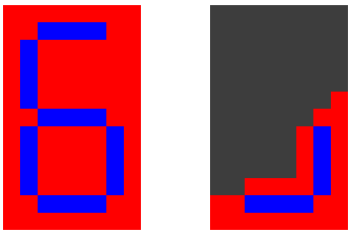
Predchádzajúce sekcie opisujú štartovaciu čiaru nášho výskumu. Máme model asociatívnej pamäte s jednou skrytou vrstvou. Neuróny na skrytej vrstve môžu slúžiť ako detektory konkrétnych trénovacích vzorov (exemplárov), alebo častí, z ktorých sú tieto trénovacie vzory zložené. Naším cieľom je nájsť takú kombináciu aktivačnej dynamiky a trénovacieho pravidla, ktorou dosiahneme efektívnu deľbu práce medzi neurónmi na tejto skrytej vrstve. Vlastnosti, ktoré by v ideálnom prípade mal náš model ideálne spĺňať, sú

- spojité stavy,
- miera biologickej plauzibilnosti,
- učenie bez učiteľa,
- rýchle učenie,
- inkrementálne učenie,
- rýchla rekonštrukcia (ideálne v jednom kroku),
- vysoká kapacita (podobne ako DAM a CMHN),
- generalizácia.

Prístup opísaný nižšie je založený na princípoch známych z teórie dynamických systémov - pozitívnej a negatívnej spätnej väzbe.

5 Kompetícia vektorov a komponentov

Kombinácia pozitívnej a negatívnej spätnej väzby vo forme laterálnej excitácie a inhibície sa javí ako kľúčová, pretože vedie k stabilite učiacich pravidiel (dĺžka váhových vektorov nerastie donekonečna) a kompetícii, ktorá vedie k deľbe práce. Inhibíciu a negatívnu spätnú väzbu ako dôležitý mechanizmus v neurálnych modeloch využívajú aj O'Reilly a Munakata (2000), O'Reilly a spol. (2012) a Krotov a Hopfield (2019).



Obr. 4: Maskovaný detektor. Na ľavej strane obrázka je vstupný vzor, digitálne čislo 6. Na pravej strane je detektor pravej spodnej časti vzoru, ktorá sa nachádza v čísle 6. Červené pixely predstavujú negatívnu hodnotu komponentu, modré pixely predstavujú pozitívnu hodnotu komponentu, šedé pixely predstavujú číslo 0. Nulové komponenty v našom prístupe slúžia ako maska, ktorá označuje časti, ktoré detektor neberie do úvahy.

Chceme, aby použitie mechanizmu pre výpočet nového stavu CMHN (7) vyskladalo výsledný stav z častí, ktoré sú zachytené vo váhach príslušných latentných neurónov. Cieľom je tvorba maskovaných detektorov častí vzorov ktoré sa opakujú vo vstupnej množine. Príklad jedného takého detektora je znázornený na Obr. 4. Základom pre tento prístup je kompetitívne učiace pravidlo používané v samoorganizujúcich sa mapách

$$\boldsymbol{w}_i^{t+1} = \frac{\boldsymbol{w}_i^t + \alpha h(i, i*)\boldsymbol{x}}{\|\boldsymbol{w}_i^t + \alpha h(i, i*)\boldsymbol{x}\|}, \tag{8}$$

kde α je rýchlosť učenia a $h(i, i*)$ je funkcia susednosti dôležitá kvôli zachovaniu topologickej organizácie.

Na určovanie podobnosti vektorov nevolíme euklidovskú vzdialenosť, ale kosínovú podobnosť. Kosínova podobnosť má pre nás dôležitú vlastnosť - nulové komponenty neprispievajú k jej veľkosti. Vďaka tomu nám podobnosť váhového vektora so vstupným vektorom ukáže iba mieru, do akej sa vo vstupnom vektore nachádza časť detekovaná aktuálnym neurónom.

Keďže v našom modeli je zachovanie topologickej organizácie nežiadúce, v (8) nahradíme funkciu susednosti. Na jej mieste použijeme dva nové členy ktoré zabezpečia pozitívnu spätnú väzbu. V komponentovej forme

$$w_{ij}^{t+1} = \frac{w_{ij}^t + \alpha p_j \gamma_{ij} x_i}{\sqrt{\sum_{k=1}^{d} \left(w_{kj}^t + \alpha p_j \gamma_{kj} x_k\right)^2}}, \tag{9}$$

p_j určuje pravdepodobosť, že latentný neurón j reprezentuje aktuálny vstup $\boldsymbol{x}$ a γ_{ij} je člen, ktorý moduluje rýchlosť rastu komponentu i aktuálneho neurónu.

Jednotlivé váhové vektory sú vo vzájomnej kompetícii. Mieru úpravy váh latentného neurónu určuje podobnosť jeho váhového vektora s aktuálnym vzorom. Komponenty váhových vektorov súťažia podľa podmienok určených výpočtom hodnoty členu γ_{ij}. Pozitívna spätná väzba v tomto prípade znamená, že podobnejšie vektory a ich komponenty rastú rýchlejšie. Spolu s normalizáciou vektorov to vedie k približovaniu sa nerelevantných komponentov smerom k nule, teda k tvorbe masky ako na Obr. 4.

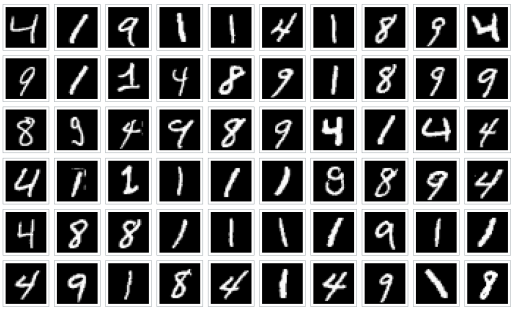
5.1 Schopnosť tvorby maskovaných detektorov

Hodnotiť efektívnosť tohto učiaceho pravidla môžeme na dvoch úrovniach:

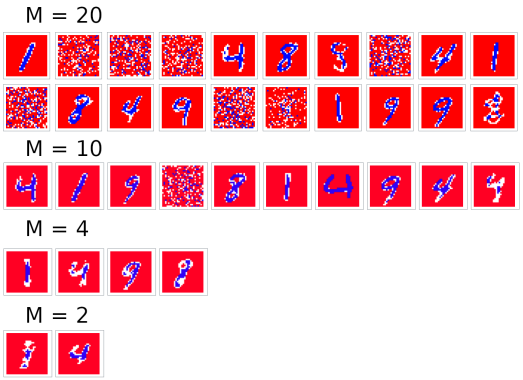
1. Vedie k deľbe práce medzi neurónmi? Učia sa rôzne neuróny detekovať rôzne vzory?
2. Vedie k tvorbe detektorov častí vstupných vzorov?

Obr. 6 ukazuje výsledné váhové vektory neurónov na skrytej vrstve pri trénovaní siete pravidlom (9) na podmnožine datasetu MNIST (Deng, 2012) obsahujúcej 85 exemplárov čísel 1, 4, 8 a 9 (časť tejto podmnožiny sa dá vidieť na Obr. 5). Pre výpočet členov rovnice 9 sme zvolili $\gamma_{ij} = e^{-(w_{ij}-x_i)^2}$ a $\boldsymbol{p} = softmax(\beta \boldsymbol{W}^T \boldsymbol{x})$. Parametre: $\alpha = 0.01$, $\beta = 0.3$ rastúca o 10% po každej iterácii, 100 iterácií.

Z výsledných váh je vidieť, že zatiaľ čo pravidlo na pohľad dobre dosahuje deľbu práce medzi neurónmi



Obr. 5: Ukážka trénovacích dát použitých v experimente. Podmnožina datasetu MNIST obsahujúca iba čísla 1, 4, 8 a 9.



Obr. 6: Vizualizácia vektorov váh neurónov na skrytej vrstve pri použití nášho kompetitívneho pravidla. Trénovanie prebehlo na podmnožine datasetu MNIST skladajúcej sa z čísel 1, 4, 8 a 9. Farebné kódovanie: červená - negatívna hodnota, modrá - pozitívna hodnota, biela - hodnota blízka nule. M je počet neurónov na skrytej vrstve.

- rôzne neuróny sa učia detekovať rôzne vzory - takmer vôbec nevedie k tvorbe maskovaných detektorov. Jednotlivé váhové vektory pripomínajú konkrétne exempláre z trénovacej množiny a nie ich časti. Ak je počet neurónov na skrytej vrstve výrazne vyšší ako počet kategórií číslic ($M = 20, M = 10$), niektoré neuróny sa nešpecializujú vôbec. Toto je vlastnosť, ktorá by hypoteticky mohla viesť k inkrementálnemu učeniu. Pri dodatočnom trénovaní modelu za účelom naučenia sa nového pamäťového vzoru by sa mohli špecializovať práve tieto neuróny. Pri $M = 4$, teda pri počte rovnom počtu kategórií číslic v trénovacej množine, sa každý neurón špecializoval na jednu kategóriu. Počeť neurónov nižší ako počet kategórií číslic vedie k zmiešaným detektorom s vyšším množstvom nulových komponentov.

Rekonštrukcia vzorov pomocou takýchto váh bola vo všeobecnosti úspešná iba vzhľadom na príslušnosť ku kategórii. Na Obr. 7 je príklad rekonštrukcie vzorov použitých pri trénovaní. Aj napriek tomu, že kategória rekonštruovaných vzorov je správna, vzory samotné sú po rekonštrukcii poškodené. Táto konfigurácia učiaceho pravidla je zaujímavá vzhľadom na deľbu práce medzi neurónmi, ale pre potreby rekonštrukcie a tvorbu vhodných detektorov sa javí ako nedostatočná.

Obr. 7: Príklad rekonštrukcie vzorov siete trénovanej kompetitívnym učiacim pravidlom (9). V hornom riadku sú vstupné vzory, v príslušných stĺpcoch spodného riadku sú rekonštrukcie. Rekonštrukcia je výsledkom jedného rekonštrukčného kroku aktivačnej dynamiky spojitých moderných Hopfieldových sietí (7) s parametrom $\beta = 3,5$. Rekonštruované vzory majú zvýraznený kontrast pre lepšiu viditeľnosť.

5.2 Záver

V práci sme opísali jeden z prístupov, ktorý volíme v snahe dosiahnuť efektívnu deľbu práce v modeli asociatívnej pamäte inšpirovanom modernými variantmi Hopfieldových sietí. Trénovanie je založené na princípe samoorganizácie pomocou pozitívnej a negatívnej spätnej väzby. Tú dosahujeme modifikáciou kompetitívneho učiaceho pravidla, ktoré má v sebe explicitne zahrnuté členy ovplyvňujúce kompetíciu medzi váhovými vektormi neurónov na skrytej vrstve a medzi jednotlivými komponentmi týchto vektorov. Laterálna inhibícia, ktorá sa javí ako kritický mechanizmus pre dosiahnutie nášho cieľa, je implementovaná normalizáciou jednotlivých váh. Prezentované výsledky ukazujú, že zatiaľ čo pravidlo zdanlivo vedie k rozdeleniu si zodpovedností medzi jednotlivými neurónmi, natrénovať detektory častí vstupných vektorov ani dosiahnuť akceptovateľnú rekonštrukciu sa nám zatiaľ nepodarilo.

Literatúra

Acevedo-Mosqueda, M., Yanez-Marquez, C. a Acevedo-Mosqueda, M. (2013). Bidirectional associative memories: Different approaches. *ACM Computing Surveys*, 45(2).

Demircigil, M., Heusel, J., Löwe, M., Upgang, S. a Vermet, F. (2017). On a model of associative memory

with huge storage capacity. *J. Stat. Phys. 168 (2), 288-299 (2017)*.

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142.

Gorman, C., Robins, A. a Knott, A. (2017). Hopfield networks as a model of prototype-based category learning: A method to distinguish trained, spurious, and prototypical attractors. *Neural Networks*, 91:76–84.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, 79(8):2554–2558.

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, 81(10):3088–3092.

Krotov, D. (2021). Hierarchical associative memory. *CoRR*, abs/2107.06446.

Krotov, D. a Hopfield, J. (2020). Large associative memory problem in neurobiology and machine learning.

Krotov, D. a Hopfield, J. J. (2016). Dense associative memory for pattern recognition. *Advances in Neural Information Processing Systems 29 (2016), 1172-1180*.

Krotov, D. a Hopfield, J. J. (2019). Unsupervised learning by competing hidden units. *Proceedings of the National Academy of Sciences of the United States of America*, 116(16):7723–7731.

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J. a Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(6):335–346.

O'Reilly, R. C. a Munakata, Y. (2000). *Computational explorations in cognitive neuroscience*. MIT Press, Cambridge, Mass. [u.a.].

O'Reilly, R. C., Munakata, Y., Frank, M. J., Hazy, T. E. a Contributors (2012). *Computational Cognitive Neuroscience*. Online Book, 4th Edition, URL: `https://CompCogNeuro.org`.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., Gruber, L., Holzleitner, M., Pavlović, M., Sandve, G. K., Greiff, V., Kreil, D., Kopp, M., Klambauer, G., Brandstetter, J. a Hochreiter, S. (2020). Hopfield networks is all you need.

Steinbuch, K. (1961). Die lernmatrix. *Kybernetik*, 1(1):36–45.

Strogatz, S. H. (2018). *Nonlinear Dynamics and Chaos*. CRC Press.

Takáč, M., Knott, A. a Sagar, M. (2020). SOM-based system for sequence chunking and planning. V *Artificial Neural Networks and Machine Learning – ICANN 2020*, str. 672–684. Springer International Publishing.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. a Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Widrich, M., Schäfl, B., Pavlović, M., Ramsauer, H., Gruber, L., Holzleitner, M., Brandstetter, J., Sandve, G. K., Greiff, V., Hochreiter, S. a Klambauer, G. (2020). Modern hopfield networks and attention for immune repertoire classification. Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. a Lin, H. (zost.), V *Advances in Neural Information Processing Systems*, vol. 33, str. 18832–18845. Curran Associates, Inc.

Pozornosť ako biologicky inšpirovaný koncept pre vysvetliteľné, robustné a efektívne strojové učenie

Igor Farkaš, Barbora Cimrová, Štefan Pócoš, Iveta Bečková

Fakulta matematiky, fyziky a informatiky
Univerzita Komenského v Bratislave
{farkas,cimrova,pocos,beckova}@fmph.uniba.sk

Abstrakt

Strojové učenie zožalo vďaka hlbokým neurónovým sieťam významné úspechy, čo sa týka riešenia rozmanitých úloh ako sú klasifikácia obrázkov, jazykové úlohy, alebo rozhodovanie v hrách. Na druhej strane, známe sú nedostatky týchto metód umelej inteligencie ako napríklad nízka efektivita trénovania, netransparentnosť alebo absencia robustnosti natrénovaných modelov. V príspevku predstavíme koncept pozornosti z pohľadu psychológie a neurovedy, kde zahŕňa širšie spektrum schopností s rozmanitými mechanizmami v mozgu, ako aj z pohľadu strojového učenia, kde zavedenie pozornosti prispelo k zlepšeniu presnosti a vysvetliteľnosti modelov neurónových sietí, avšak nie efektivity trénovania a robustnosti. Dá sa teda predpokladať, že potenciál v tomto smere nebol ešte vyčerpaný.

1 Úvod

Strojové učenie, a s tým súvisiaca umelá inteligencia, sa teší v ostatnej dekáde vysokej popularite vďaka hlbokým neurónovým sieťam, pretože tie umožnili nachádzať úspešné riešenia rôznych úloh ako sú klasifikácia obrázkov, úlohy v prirodzenom jazyku či hranie hier (Schmidhuber, 2015). Výsledky týchto výpočtových modelov v mnohých prípadoch dosahujú úroveň človeka, a niekedy ho aj prekonávajú (Mnih a spol., 2015). O umelých neurónových sieťach je známe, že sú architektonicky inšpirované sieťami v mozgu človeka a ich procesy učenia pripomínajú učenie u ľudí (na príkladoch). Učenie na príkladoch sa javí ako najlepší spôsob, ako dosiahnuť zložité správanie, ktoré formálne predstavuje matematické zobrazenie vstupov na výstupy (napr. obrázkov na predikovanú kategóriu). Tento konekcionistický prístup stojí vo fundamentálnom kontraste so symbolovými modelmi na báze logiky a ontológií, kde ťažisko spočíva v expertíze dizajnéra, ktorý de facto vytvorí hotový znalostný systém. Hlboké neurónové siete učiace sa s učiteľom využívajú čisto empirický prístup (tzv. end-to-end), pri ktorom sa sieť učí úlohy priamo z pôvodných vstupov (a nerieši sa explicitne extrakcia príznakov). Neurónové siete majú veľa pozitív a je zjavné, že v ostatných rokoch hrajú prvé husle v strojovom učení, pričom významnú úlohu zohrávajú aj vo výpočtovej kognitívnej vede (Farkaš, 2011). Cieľom tohto príspevku je však poukázať na súčasné nedostatky neurónových sietí a možnosť ich odstránenia alebo aspoň zmiernenia, pomocou mechanizmov pozornosti.

2 Nedostatky umelých neurónových sietí

Najmä v súvislosti s hlbokými modelmi umelých neurónových sietí, ktoré mávajú veľa skrytých vrstiev, a teda aj astronomický počet voľných (trénovateľných) parametrov (t.j. váh medzi neurónmi), vznikli tri hlavné problémy: (1) nízka efektivita trénovania, (2) nízka transparentnosť, a (3) absencia robustnosti. Stručne si vysvetlíme každý z týchto nedostatkov.

2.1 Nízka efektivita trénovania

Neurónové siete bežne potrebujú veľa opakovaní príkladov, na ktorých sa učia. Počet potrebných opakovaní obyčajne závisí od veľkosti siete, zložitosti úlohy, ako aj ďalších faktorov (napr. hyperparametre siete). Dĺžka trénovania výrazne narastá u hlbokých modelov, ktoré súčasne potrebujú obrovské množstvo príkladov, a tie sú našťastie v súčasnosti už dostupné. Ukazuje sa, že to pomáha tomu, aby sieť predišla preučeniu (t.j. zameraniu sa na detaily v trénovacích dátach), a tým pádom slabej generalizácii (t.j. predikcii na testovacích dátach).

Existujú aj snahy, ako znížiť časovú náročnosť trénovania, lebo tá už sa stáva aj ekologickým problémom (trénovanie neurónovej siete je dosť energeticky náročné). Na druhej strane, sú známe aj metódy učenia na pár príkladoch (few-shot learning), alebo len jednom (one-shot learning), ale tie majú tiež svoje obmedzenia a predstavujú len malú časť použiteľných prístupov.

Dlhé trvanie učenia má dva hlavné dôvody. Po prvé, sieť v podstate začína pri trénovaní od nuly (čisto empirický prístup), zatiaľ čo u človeka sa predpokladajú už nejaké predispozície alebo znalosti získané z predchádzajúcich skúseností. Známe sú rôzne heuristiky ako sieť správne inicializovať, aby sa „dobre učila", ale toto problém efektívnosti nerieši. Po druhé, učenie zložitejších klasifikačných úloh predstavuje tzv. ne-

konvexný problém, ktorého dostatočne dobré riešenie (lokálne minimum chybovej funkcie) hľadáme iteratívnym spôsobom, čo je v podstate pohyb dosť naslepo vo vysokorozmernom priestore trénovateľných parametrov.

2.2 Nízka transparentnosť

Nízka transparentnosť neurónových sietí priamo vyplýva z ich architektúry a reprezentácie znalostí (pomocou reálnych čísel), ktoré sú ukryté vo váhach medzi neurónmi a aktivitách neurónov. Vzhľadom na úspešnosť týchto modelov je dôležité hľadať spôsoby, ako neurónovým sieťam lepšie porozumieť. Vysvetliteľná umelá inteligencia sa stala dôležitou vetvou výskumu, zameranou na pochopenie rôznych metód umelej inteligencie (Barredo Arrieta and others, 2020). Súčasťou tejto agendy je aj vysvetlenie toho, prečo neurónová sieť dáva na výstupe to, čo dáva (Montavon a spol., 2018). Vysvetlenia sú dôležité pre rôzne cieľové skupiny, či už expertov, užívateľov alebo pacientov, s čím súvisia aj rôzne úrovne vysvetlenia (expert rozumie aj matematickým formulám, zatiaľ čo bežný človek uprednostní vysvetlenie v prirodzenom jazyku alebo obrázkoch).

2.3 Absencia robustnosti

Absencia robustnosti natrénovaných neurónových sietí je najnovšie identifikovaný problém, ktorý bráni v nasadzovaní týchto modelov do rôznych kľúčových aplikácií. Tento problém prakticky znamená, že sieť sa dá ľahko oklamať. Samozrejme, nie hocijako, ale oveľa triviálnejšie, než človek. Geniálna myšlienka autorov (Szegedy a spol., 2014) tejto idey spočívala v návrhu takých špeciálnych vstupov pre úspešne natrénovanú sieť, pre ktoré dáva úplne zlé predikcie, častokrát s vysokou mierou presvedčenia.

Najbežnejším príkladom je klasifikácia obrázkov do tried (pričom na počte tried nezáleží). Natrénovaná sieť s vysokou presnosťou predikuje správne triedy na testovacích dátach, no napriek tomu sa dá ľahko oklamať obrázkami, ktoré boli len málo, no veľmi špecificky, pozmenené. To naznačuje, tieto vstupy sú dosť zriedkavé na to, aby sa neprejavili na testovacej chybe, no zároveň dosť bežné, aby sa dali vhodnými metódami nájsť. Skúmanie robustnosti neurónových sietí patrí medzi aktívne oblasti výskumu (Bečková a spol., 2020; Pócoš a spol., 2022).

Absencia robustnosti je asi najväčší problém, pretože otázka bezpečnosti je v modernom technologickom svete kľúčová. Nutná dĺžka trénovania sa dá zvládnuť, a v prospech toho hrá aj zrýchľujúci sa hardvér. Nízka transparentnosť sa možno nikdy nebude dať úplne prekonať, a možno riešenie bude spočívať v získaní dôveryhodnosti systému umelej inteligencie, ak bude správne fungovať (ani človek nedokáže vždy jasne zdôvodniť svoje rozhodnutie). Avšak absencia robustnosti nie je tolerovateľná, aj preto, že človek ponúka spoľahlivejšie, robustnejšie riešenie, z čoho vyplýva ďalšia potreba inšpirovať sa biologickými systémami. Pozornosť je jednou z ciest.

3 Mechanizmy pozornosti

Pozornosť je pojem známy v bežnom jazyku ale aj vo vedeckom skúmaní, najmä v psychológii a neurovede. Počiatky jeho skúmania siahajú na koniec 19. storočia, keď svetovo známy americký filozof William James ju opísal nasledovne: „Každý vie, čo je pozornosť. Je to ovládnutie mysle, v jasnej a živej forme, jedným zo zdanlivo niekoľkých súčasne možných objektov alebo myšlienkových pochodov." (James, 1890).

Odvtedy sa však chápanie pozornosti výrazne posunulo, až do takej miery, že súčasnú perspektívu niektorí autori, napríklad Hommel a spol. (2019), opisujú veľmi pesimisticky: „Nikto nevie, čo pozornosť je." V článku argumentujú, že existujú tri hlavné problémy v chápaní konceptu pozornosti u ľudí: Po prvé, koncept pozornosti vyvoláva mylné predstavy o jednom koherentnom súbore kognitívnych alebo nervových operácií, v závislosti od úrovne analýzy, ktoré všetky prispievajú k tomu, čo nazývame „pozornosť". Ako druhý problém uvádzajú to, že pozornosť sa uvádza ako problém, ktorý sa snažíme vysvetliť, ale aj ako samotné vysvetlenie (napr. pozornosť ako výsledok kapacitných obmedzení mozgu na jednej strane, verzus pozornosť ako schopnosť vysporiadať sa s týmito obmedzeniami). A po tretie, predpokladá sa, že pozornosť predstavuje konkrétny súbor kognitívnych alebo nervových operácií od iných, zdanlivo odlišných operácií, ako sú tie, ktoré súvisia s rozhodnutiami, zámermi, motiváciou, emóciami ale najmä plánovaním a vykonávaním akcií.

Je teda zložité nájsť jednotiaci konceptuálny rámec, ktorý by zastrešil všetky významy pozornosti, no niektorí autori sa o to snažia (Lindsay, 2020).

3.1 Koncept pozornosti v psychológii a neurovede

Vedecké skúmanie pozornosti má svoj pôvod v psychológii, kde dôsledné experimentovanie so správaním môže viesť k presným prejavom tendencií a vlastností pozornosti pri rôznych podmienkach. Cieľom kognitívnej vedy a kognitívnej psychológie je premeniť tieto pozorovania na modely mentálnych procesov, ktoré by mohli vytvárať takéto vzorce správania. Takýchto teoretických a výpočtových modelov bolo vytvorených veľa, s rôznymi predpokladanými základnými mechanizmami (Driver, 2001; Borji a Itti, 2013).

V oblasti neurovedy zase dostupnosť dát z neurofyziologických meraní aktivít neurónov v mozgu u zvierat, spolu s neinvazívnymi metódami merana

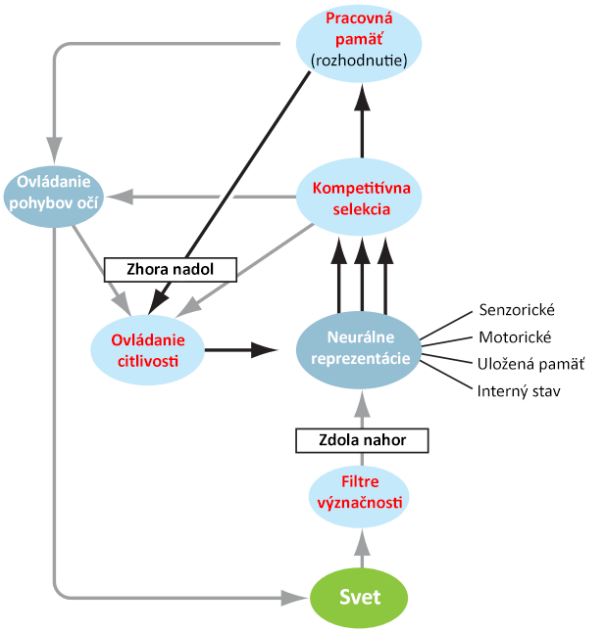
ľudskej mozgovej aktivity (ako napr. EEG, fMRI a MEG), umožnili priamo pozorovať základné neurálne koreláty kognitívnych procesov. To priamo umožňuje návrh výpočtových neurálnych modelov, ktoré dokážu replikovať empirické dáta a ponúkať tak mechanistické vysvetlenie rôznych prejavov pozornosti.

Spektrum toho, čo označujeme ako pozornosť je naozaj rozmanité. Pri nahliadnutí do učebníc kognitívnej psychológie (napr. Eysenck a Keane 2000) nachádzame rôzne príklady schopností: (1) vybrať vonkajšie udalosti pre ďalšie interné spracovanie (sústredená pozornosť); (2) ignorovať zavádzajúce informácie a/alebo irelevantné lokácie (selektívna pozornosť); (3) automaticky spracovávať nepodstatné informácie (mimovoľná pozornosť); (4) selektívne integrovať informácie patriace k jednej udalosti v rámci zmyslových modalít a medzi nimi (integrácia informácií); (5) uprednostniť spracovanie udalostí z konkrétnej lokácie (priestorová pozornosť); (6) systematicky vyhľadávať cieľovú udalosť (vizuálne vyhľadávanie); (7) vykonávať viacero úloh súčasne (rozdelená pozornosť); (8) ovládať priestorové parametre pohybov očí (selektívna pozornosť na akciu); (9) uprednostniť jeden cieľ pred ostatnými (pozornosť zameraná na cieľ); (10) uprednostniť jeden objekt, pamäťovú položku alebo vedomú reprezentáciu pred ostatnými (objektovo sústredená pozornosť); a (11) konsolidovať informácie pre neskoršie použitie a sústrediť sa na očakávanie možnej udalosti počas určitého času (trvalá pozornosť).

Každopádne, tieto rôzne typy pozornosti je možné kategorizovať, čo trochu zvyšuje ich pochopenie. Lindsay (2020) uvádza nasledovné typy pozornosti: (1) Pozornosť ako nabudenie alebo ako bdelosť, (2) senzorická pozornosť v rôznych modalitách (s dominanciou vizuálnej), zameraná na príznaky, alebo na lokáciu, (3) pozornosť pri exekutívnom riadení, a (4) interakcie pozornosti s pamäťou. S typmi pozornosti sa spájajú rôzne aspekty, ako napr. skrytá/otvorená (angl. covert/overt) pozornosť, procesy zdola nahor verzus zhora nadol.

Zrozumiteľnú schému funkčných mechanizmov pozornosti vyjadruje obr. 1. Ide o permanentnú interakciu s prostredím, v rámci ktorej sa uplatňujú štyri komponenty: (1) pracovná pamäť, (2) ovládanie senzitivity, (3) kompetitívna selekcia a (4) automatické filtrovanie význačných stimulov. Každý proces výrazne a zásadne svojím spôsobom prispieva k pozornosti, pričom vôľou riadené zameranie pozornosti zahŕňa prvé tri procesy, fungujúce zhora nadol, ktoré fungujú v rekurentnej slučke. Opačným smerom pôsobí automatická detekcia význačných stimulov.

Pracovná pamäť je špecifická forma pamäti, cez ktorú prechádza spracovanie akejkoľvek senzorickej informácie nielen z okolitého sveta, ale aj vnútorného sveta. Obsah pracovnej pamäte (s kapacitnými obmedzeniami) je tak okamžite spracovateľný v danom kontexte a stáva sa predmetom pozornosti. To, ktorá informácia získa prístup do pracovnej pamäti, je



Obr. 1. Funkčné mechanizmy pozornosti (podľa Knudsen 2007).

výsledkom kompetitívnej selekcie (súťaženia o miesto v pracovnej pamäti). Pracovná pamäť sa týka všetkých modalít a má v nich svoje špecifiká. Je široko distribuovaná v mozgu, s centrom riadenia v prefrontálnej kôre.

Ovládanie citlivosti modulované zhora nadol hrá kľúčovú úlohu pri optimalizácii informácie, ktorá je v centre pozornosti. To sa dosahuje buď zameraním pohľadu na objekt, čím sa zvyšuje priestorové rozlíšenie, alebo zvýšením pomeru signál-šum. Toto sa týka všetkých senzorických modalít, pamäti, aj vnútorných stavov. Neurálna evidencia týchto modulačných procesov pochádza najmä z (invazívnych) elektrofyziologických meraní najmä na mozgoch opíc, kde vidieť zmeny citlivosti neurónov, pričom ich zníženie, resp. zvýšenie (t.j. miery aktivity neurónu) sa potom dá interpretovať ako neurálna implementácia miery pozornosti. Informácia sa môže dostať do mozgu aj zdola nahor, bez zasahovania mechanizmov zhora nadol. Príkladom sú podnety s vysokou význačnosťou, ktoré skrátka automaticky upútajú pozornosť človeka alebo zvieraťa. Takýto prístup k pracovnej pamäti riadený vonkajšími podnetmi, bežne označovaný ako pozornosť zdola nahor, odráža účinky filtrov význačnosti (salientnosti) na mnohých úrovniach v centrálnom nervovom systéme, ktoré selektujú vlastnosti tých podnetov, ktoré budú pravdepodobne dôležité. Tieto filtre sú realizované rôznymi neurálnymi mechanizmami.

Prezentovaný pohľad navodzuje konceptualizáciu pozornosti ako inherentnej súčasti všetkých perceptuálnych a kognitívnych procesov človeka či zvieraťa, ktorá funguje v permanentnej slučke, v rámci ktorej dochádza k adaptívnemu a flexibilnému riade-

Tab. 1. Typické prístupy k mechanizmom pozornosti v strojovom učení (prevzaté z Niu a spol. (2021)).

Kritérium	Pozornosť
jemnosť pozornosti	spojtá/diskrétna globálna/lokálna
forma vstupných príznakov	na položku na lokáciu
vstupné reprezentácie	vzájomná, na seba, spoločná, hierarchická
výstupné reprezentácie	jeden výstup, viac hláv, viacrozmerná

niu obsahu práve spracovávanej informácie. Táto informácia môže pritom pochádzať z vonkajšieho prostredia (vlastnosť objektu, lokácia v priestore) alebo môže byť interne generovaná (obsah dlhodobej pamäti, pravidlo relevantné pre rozhodovanie). Chun a spol. (2011) ponúkajú taxonómiu mechanizmov pozornosti práve z tejto perspektívy. Súčasne argumentujú proti existencii jednotiaceho modelu pozornosti vzhľadom na rozmanitosť mechanizmov, ktoré stoja za jej prejavmi.

3.2 Koncept pozornosti v strojovom učení

Mechanizmy pozornosti v umelých systémoch nie sú novou záležitosťou, no napriek trom dekádam výskumu v tejto oblasti, najmä v umelých neurónových sieťach, tieto stále vo väčšine prípadov nedosahujú úroveň človeka. Mechanizmy pozornosti boli aplikované v mnohých úlohách (pozri napr. prehľad v Niu a spol. (2021)), no najvýznamnejšie dve oblasti predstavuje prirodzený jazyk (Galassi a spol., 2021), ktorý zahŕňa rôzne úlohy a počítačové videnie (Guo a spol., 2022), kde najčastejšou úlohou je klasifikácia obrazových dát.

Typická implementácia pozornostného mechanizmu spočíva v tom, že sieť sa trénuje s učiteľom (najčastejšie pomocou algoritmu spätného šírenia chyby) tak, aby dokázala správne riešiť úlohy vďaka zameraniu pozornosti na časť vstupu. Za touto naučenou schopnosťou sa skrýva iteratívne nastavenie matíc parametrov, ktoré určujú algebraické transformácie vektorov aktivít na rôznych vrstvách siete (spolu s nelinearitami neurónov). Mechanizmy pozornosti pritom počas testovania pôsobia typicky zdola nahor.

V oblasti strojového učenia tiež existujú snahy o unifikáciu mechanizmov pozornosti, možno aj preto, že spektrum existujúcich mechanizmov a použitých reprezentácií je oveľa užšie ako v mozgu. Niu a spol. (2021) vo svojom prehľade výskumu prezentovali jednotiaci model pozornosti (obr. 4 v článku), ktorý sa týka hlbokých neurónových sietí. Mechanizmy pozornosti rozdelili podľa štyroch kritérií, ako znázorňuje tab. 1.

Pri klasifikácii obrázkov model zameria pozornosť na časť obrázka, ktorá výrazne prispeje k predikcii správnej kategórie, alebo pri jazykovom preklade na relevantné slovo zdrojovej vety. Pri rozlišovaní spôsobu váhovania jednotlivých komponentov hovoríme o spojitej (alebo globálnej) pozornosti, ak ku každému komponentu prislúcha kladná váha, hoci aj veľmi malá. V prípade „ostrého" výberu komponentov hovoríme o diskrétnej pozornosti.

Ak vstupné príznaky sú vektory, hovoríme o pozornosti na položku, pričom pozornosť je smerovaná na jednotlivé vstupné vektory. Menej častým prípadom je pozornosť na lokáciu, ktorá sa využíva ak nemáme viac vstupov, no chceme nájsť nejakú informatívnu oblasť vstupu (používa sa najmä na obrázky).

Výpočet pozornostných váh závisí od zdroja informácií. Ak počítame pozornosť jedného vektora vzhľadom na druhý, ide o vzájomnú pozornosť (napr. pri preklade vety do iného jazyka). V prípade, že v tomto procese figuruje iba jeden vektor, ide o pozornosť na seba (klasifikácia obrázka). V niektorých aplikáciách je možné miešať pozornosť reprezentácií z rôznych príznakov. Tu hovoríme o spoločnej pozornosti. Napokon, pozornosť môže byť aj hierarchická, napríklad keď máme viac úrovní reprezentácie a v každej použijeme nejakú formu pozornosti.

Čo sa týka výstupu pozornostného mechanizmu, ten môžeme reprezentovať rôzne. Bežný spôsob je použiť jeden výstup (zvyčajne vektor) ktorý sumarizuje niekoľko vstupných vektorov. Rozšírením tejto myšlienky je použitie viacerých pozornostných hláv, ktoré podporujú bohatšiu reprezentáciu informácie v modeli. Zaujímavou alternatívou je aplikovanie viacrozmernej pozornosti, ktorá pri vhodnej aplikácii ponúka možnosť zmysluplnej reprezentácie vstupu viacerými možnými spôsobmi.

V aktuálnom prehľadovom článku Guo a spol. (2022) ponúkajú alternatívnu taxonómiu prístupov s využitím mechanizmov pozornosti v oblasti počítačého videnia. Tieto mechanizmy boli využité v rôznych úlohách, t.j. okrem klasifikácie obrazu pri detekcii objektov, sémantickej segmentácii, porozumenia videu, generovania obrazu, 3D videnia, ako aj multimodálnych úlohách. Pozornosť v týchto prístupoch možno algoritmicky zamerať na rôzne aspekty (lokácia, čas, farebný kanál, alebo aj vetvu v spracovaní, napr. lokálnu/globálnu), ako aj ich kombinácie. Z článku je zrejmé, že rozmanitosť modelov v ostatných rokoch narástla významne.

4 Záver

Je zrejmé, že mechanizmy pozornosti sa stali kľúčovou súčasťou modelov neurónových sietí s cieľom vylepšiť ich vlastnosti. Toto sa čiastočne podarilo, pretože pozornosť pomáha zvyšovať presnosť modelov a prispieva ich k ich vysvetliteľnosti (aj keď na pomerne nízkej úrovni). Napriek rozmanitosti prístupov pretrvávajúcim

problémom ostáva najmä absencia robustnosti. V tomto smere by mohla pomôcť ďalšia inšpirácia z biológie. Je možné, že vyriešenie tohto problému si bude vyžadovať aj iné koncepty než pozornosť.

Pri snahe o porovnanie mechanizmov pozornosti v psychológii a v strojovom učení môžeme pozorovať, že existuje čiastočný prekryv medzi oboma oblasťami, keď niektoré koncepty majú aj svoje náprotivky. Napr. Lindsay (2020) uvádza len dva príklady: otvorená vizuálna pozornosť u človeka pripomína diskrétnu priestorovú pozornosť v umelom systéme, a skrytá vizuálna pozornosť odpovedá spojitej vizuálnej pozornosti zameranej na lokáciu alebo na nejakú črtu. Jedným zo základných rozdielov je to, že v živých systémoch dominuje mechanizmus pozornosti zhora nadol, a to v rámci permanentnej slučky s prostredím. Toto absentuje pri klasifikácii obrázkov, ale aj pri iných úlohách. V každom prípade, zakomponovanie mechanizmu pozornosti sa zdá byť nutnou, a možno nepostačujúcou zložkou pri dosiahnutí vysvetliteľného a robustného umelého systému s efektívnym učením.

Poďakovanie

Tento výskum bol podporený Slovenskou spoločnosťou pre kognitívnu vedu.

Literatúra

Barredo Arrieta and others, A. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115.

Bečková, I., Pócoš, Š. a Farkaš, I. (2020). Computational analysis of robustness in neural network classifiers. Farkaš, I., Masulli, P. a Wermter, S. (zost.), V *Artificial Neural Networks and Machine Learning – ICANN 2020*, str. 65–76. Springer.

Borji, A. a Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207.

Chun, M., Golomb, J. a Turk-Browne, N. (2011). A taxonomy of external and internal attention. *Annual Reviews of Psychology*, 62:73–101.

Driver, J. (2001). A selective review of selective attention research from the past century. *British Journal of Psychology*, 92:53–78.

Eysenck, M. a Keane, M. (2000). *Cognitive Psychology: A Student's Handbook*. Psychology Press, Philadelphia, 7. vyd.

Farkaš, I. (2011). Konekcionizmus v náručí výpočtovej kognitívnej vedy. Kvasnička, V. a spol. (zost.), V *Umelá inteligencia a kognitívna veda III*, str. 19–62. Vydavateľstvo STU v Bratislave.

Galassi, A., Lippi, M. a Torroni, P. (2021). Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(10):4291–4308.

Guo, M., Xu, T., Liu, J. a spol. (2022). Attention mechanisms in computer vision: A survey. *Computational Visual Media*, 8:331–368.

Hommel, B., Chapman, C. S., Cisek, P., Neyedli, H. F., Song, J.-H., a Welsh, T. N. (2019). No one knows what attention is. *Attention, Perception & Psychophysics*, 81:2288–2303.

James, W. (1890). *Principles of Psychology*. New York: Holt.

Knudsen, E. (2007). Fundamental components of attention. *Annual Review of Neuroscience*, 30(1):57–78.

Lindsay, G. (2020). Attention in psychology, neuroscience, and machine learning. *Frontiers in Computational Neuroscience*, 14.

Mnih, V. a spol. (2015). Human-level control through deep reinforcement learning. *Nature*, 518:529–542.

Montavon, G., Samek, W. a Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15.

Niu, Z., Zhong, G. a Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62.

Pócoš, Š., Bečková, I. a Farkaš, I. (2022). Examining the proximity of adversarial examples to class manifolds in deep networks. arXiv: 2204.05764 [cs.LG].

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. a Fergus, R. (2014). Intriguing properties of neural networks. V *International Conference on Learning Representations*.

It Is Easier with Negative Emotions: The Role of Negative Emotions and Emotional Intelligence in Epistemically Suspect Beliefs about COVID-19

Miroslava Galasová

Institute of Experimental Psychology, CSPS SAS
Dubravska cesta 9, 841 04 Bratislava
miroslava.galasova@gmail.com

Abstract

The coronavirus pandemic has been accompanied by various emotions, and evidence suggests that mainly negative emotions might positively correlate with epistemically suspect beliefs. Trait emotional intelligence, on the other hand, could moderate the relationship between negative emotions and epistemically suspect beliefs because it might modulate the processing of emotions. Therefore, the aim of the current research is twofold – to examine relationships between negative emotions and epistemically suspect beliefs about COVID-19 and examine a moderating role of the trait of emotional intelligence in this relationship. Participants (N = 254, M_{age} = 46) participated in an online survey and answered items related to trait emotional intelligence, epistemically suspect beliefs about COVID-19 and current mood. Results showed a positive relationship between negative emotions and epistemically suspect beliefs. However, the hypothesis about the moderating role of trait emotional intelligence was not supported. Eventually, understanding the role of emotions and their processing might lead to the development of effective strategies for mitigating epistemically suspect beliefs.

1 Introduction

People succumb to epistemically suspect beliefs when they believe in things or events that have not been or cannot be corroborated by a reliable scientific method (Šrol, 2021). In a worse case, they have been examined and come out disproved. An example of such belief is the advice of the former U.S. president, Donald Trump, who suggested drinking disinfectants can kill the coronavirus (Mostajo-Radji, 2021). Except drinking disinfectants can also kill the person together with the virus. Further examples of epistemically suspect beliefs have come from hospitals that cured patients overdosed with Ivermectin – an antiparasitic drug that has been allegedly effective in treating COVID-19 (Porubcin et al., 2022).

Importantly, victims of epistemically suspect beliefs are not only regular people. Epistemically suspect beliefs lure on experts too. A woman with COVID-19 ended up in a hospital after her general practitioner administered her intravenously a dose of veterinary Ivermectin, although this procedure did not follow evidence-based medicine (Porubcin et al., 2022). Well, is it true that desperate people do desperate things?

The coronavirus pandemic brought uncertainty and instability to the lives of many people. Fear, anger, anxiety, and other negative emotions flooded people around the globe (Metzler et al., 2021; Brooks et al., 2020). Unfortunately, negative emotions also seem to relate to conspiracy theory beliefs which create a branch of epistemically suspect beliefs (Douglas et al., 2020; Galasová & Čavojová, 2022; Šrol et al., 2022; Van Mulukom et al., 2022). The current study verifies previous suggestions and recent findings, which indicate negative emotions positively correlate with epistemically suspect beliefs about COVID-19. Furthermore, the study examines a moderating role of trait emotional intelligence, which might interfere with the intensity and processing of perceived negative emotions.

1.1 Epistemically Suspect Beliefs and Negative Emotions

Epistemically suspect beliefs could be divided into three branches – conspiracy theory beliefs, pseudoscientific beliefs, and paranormal beliefs – which consistently share positive associations with each other (Šrol, 2021). In the present study, the COVID-19 Epistemically Suspect Beliefs Scale aims mainly at the branches of conspiracy theory beliefs and pseudoscientific beliefs (Teličák & Halama, 2020). The pseudoscientific beliefs have already been introduced (e.g., Donald Trump's advice about disinfectants). On the other hand, often spread conspiracy theories about COVID-19 were related to its origin (e.g., *Coronavirus has been artificially created in laboratories.*), vaccination (e.g., *Vaccination is misused to implant microchips in people.*), and other topics (e.g., higher control of people).

As it may be perceived from the examples above, conspiracy theory beliefs might evoke feelings of threat, loss of control, or anxiety (Šrol et al., 2022; Van Mulukom et al., 2022). Moreover, even Douglas et al. (2020) point out that conspiracy theories flourish mainly during critical times when the mood of the society is affected by various adverse circumstances. Certain relationships between negative emotions and conspiracy theory beliefs have already been directly or indirectly supported (e.g., Galasová & Čavojová, 2022; Van Mulukom et al., 2022). Thus, we might expect that negative emotions will be positively associated with conspiracy theory beliefs. Consequently, I expect positive correlations between negative emotions and pseudoscientific beliefs, too.

1.2 Trait Emotional Intelligence

Trait emotional intelligence relates to a capacity to recognize, understand, manage, and regulate the emotions of oneself and others (Petrides, 2009). This concept builds on previous models of emotional intelligence (e.g., Mayer & Salovey, 1997), which understand emotional intelligence as a trainable ability. However, the concept of trait emotional intelligence describes emotional intelligence as a set of relatively stable traits divided into four factors and 15 facets. The four factors are emotionality, self-control, sociability, and well-being. Specific facets are listed in Figure 1. Since trait emotional intelligence might affect the intensity and processing of the perceived emotions, I hypothesized that the level of trait emotional intelligence or its specific factors could moderate the relationship between negative emotions and epistemically suspect beliefs about COVID-19.

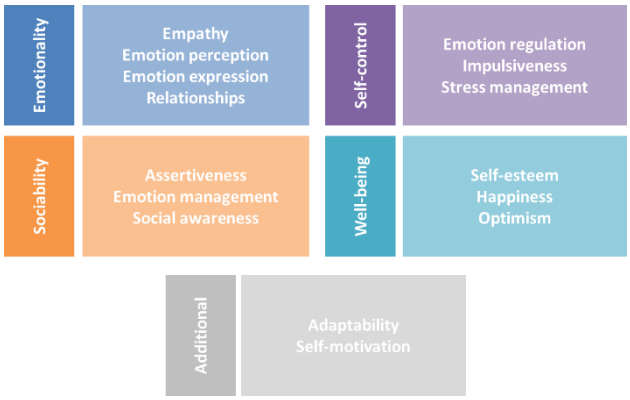


Fig. 1. The four factors and 15 facets of trait emotional intelligence (adapted from Kaliská et al., 2015).

2 Method

2.1 Design and Procedure

The current study was a part of larger cross-sectional research that primarily focused on mutual relationships and the effect of negative emotions on the susceptibility to cognitive biases and epistemically suspect beliefs. The data were collected via an online survey which started with informed consent. If participants agreed with the terms of the research, they proceeded with several basic socio-demographic questions. Then, they filled in scales on emotional intelligence, epistemically suspect beliefs about COVID-19 and their current mood.

2.2 Participants

The hired external research recruitment agency reached 254 participants (women = 63%, M_{age} = 46, SD = 16.5). Only 33 participants finished elementary or secondary education without the leaving examination. The majority of the sample (n = 121) has finished secondary education with the leaving examination, and 100 participants have obtained at least a bachelor's or higher university degree. Participants participated in the research voluntarily and anonymously. They could quit the research at any time. The recruitment agency remunerated each successfully finished participation by a small amount of money (approx. 2.00 EUR per participant).

2.3 Materials

2.3.1 International Positive and Negative Affect Schedule (I-PANAS)

The I-PANAS scale measures positive and negative affects among participants (Thompson, 2007). Participants were asked to assess their current positive (pleasantness, excitement, happiness, content) and negative (unpleasant-ness, anger, sadness, and disgust) emotions. Participants answered on a five-point Likert scale (1 = *not at all*; 5 = *very much/extremely*). For the purpose of this study, I worked with specific ratings of negative emotions and an average rating of the four negative emotions.

2.3.2 Trait Emotional Intelligence Questionnaire – short form (TEIQue-SF)

The TEIQue scale, in its short form, contains 30 items and is considered a unidimensional scale. The instrument also

has a Slovak validated version (Petrides & Furnham, 2006; Kaliská et al., 2015). Participants answered items on a 7-point Likert scale (1 = *completely disagree*; 7 = *completely agree*). Half of the items had reversed scoring. The total score was computed as the sum of ratings from all items divided by the number of items. Higher scores indicate higher emotional intelligence. The reliability of the short Slovak version for adults was excellent (ω = .89). The reliability of the four factors was also appropriate - well-being (ω = 0.78), emotionality (ω = 0.72), self-control (ω = 0.67), and sociability (ω = 0.70).

2.3.3 COVID-19 Epistemically Suspect Beliefs Scale (COVID-19 ESB)

Teličák and Halama (2020) recently developed a scale of epistemically suspect beliefs about COVID-19 and verified it among the Slovak population. The present scale consisted of eleven statements, of which five represented pseudoscientific beliefs (e.g., *Wearing protective masks is dangerous for children and older people*) and six items represented conspiracy theory beliefs about COVID-19 (e.g., *Coronavirus has been artificially created in laboratories*). Participants expressed an agreement or disagreement with these items on the five-point Likert scale (1 = *definitely disagree*; 5 = *definitely agree*). The scale had excellent reliability (ω = .93). The raw score was computed as the mean rating of all items. Higher scores indicated participants' tendency to adhere to epistemically suspect beliefs about COVID-19.

3 Results

3.1 Descriptive statistics and correlations

Provided descriptive statistics show that participants, in general, did not succumb to epistemically suspect beliefs (ESB) about COVID-19 (M = 2.40, SD = 1.08, Table 1). Nevertheless, about 46% of all participants were prone to believing them. Interestingly, I observed significant differences between the two subscales, the pseudoscientific and conspiracy theory beliefs (Wilcoxon test, Z = 2211, p < .001). While participants succumbed to conspiracy theory beliefs, they did not tend to believe in pseudoscientific statements about COVID-19 that much.

The data in Table 1 show that **negative emotions (unpleasantness, anger, and disgust) as well as the Total negative emotion (TNE) significantly and positively correlated with COVID-19 ESB**. Nevertheless, Table 2 reveals slight differences between pseudoscientific and conspiracy theory beliefs in relation to specific negative emotions. Pseudoscientific beliefs correlated significantly and positively with all negative emotions, but conspiracy theory beliefs correlated significantly only with anger, disgust and Total negative emotions. Importantly, all mentioned correlations were quite weak.

The Trait emotional intelligence and its factors appeared to have no relationships with COVID-19 ESB and its subscales, except for the factor of well-being. Well-being as the only factor of TEI showed a negative and significant correlation with COVID-19 ESB and its conspiracy theory beliefs subscale. But again, the relationship was weak.

Table 1. **Descriptive statistics and Spearman's correlations of negative emotions and Trait emotional intelligence with the COVID-19 ESB scale**

#	Variable	*M*	*SD*	*r*
1	C-19 ESB	2.40	1.08	-
2	Unpleasant	1.85	1.14	**0.14***
3	Sadness	2.01	1.19	0.12
4	Anger	1.71	1.09	**0.19****
5	Disgust	1.86	1.16	**0.15***
6	TNE	1.86	1.03	**0.15***
7	TEI	4.90	0.73	-0.08
8	Well-being	5.19	0.98	**-0.13***
9	Emotionality	5.09	0.91	-0.06
10	Self-control	4.76	0.88	-0.02
11	Sociability	4.62	0.94	0.01

Note. The table contains only correlations for the COVID-19 ESB scale. TNE = total negative emotions (average score), TEI = trait emotional intelligence, * = p < .05, ** = p < .01.

Table 2. **Spearman's correlations of the two COVID-19 ESB subscales, negative emotions and Trait emotional intelligence**

#	Variable	**PSB** (*Mdn* = 1.80)	**CTB** (*Mdn* = 2.83)
1	Unpleasant	**0.16***	0.11
2	Sadness	**0.14***	0.11
3	Anger	**0.19****	**0.17****
4	Disgust	**0.14***	**0.15***
5	TNE	**0.16***	**0.14***
6	TEI	-0.02	-0.11
7	Well-being	-0.10	**-0.14***
8	Emotionality	0.00	-0.10
9	Self-control	0.05	-0.06

10	Sociability	0.04	0.00

Note. Correlations of the pseudoscientific subscale (PSB), conspiracy theory beliefs subscale (CTB), and other variables. TNE = total negative emotions (average score), TEI = trait emotional intelligence, * = $p < .05$, ** = $p < .01$, † = $p = .05$

3.2 Testing Trait emotional intelligence in the role of the moderator

In the next step, I analyzed the role of trait emotional intelligence in the relationship between Total negative emotions and COVID-19 ESB. Moderation analysis conducted in JAMOVI showed that the Trait emotional intelligence was not a significant moderator in this relationship (B = -0.12, 95% CI [-0.31, 0.06], p = .19). Interestingly, however, the Simple slope analysis showed that the effect of Total negative emotions on the COVID-19 ESB was significant when the Trait emotional intelligence was either low (B = 0.26, 95% CI [0.10, 0.41], p = .001) or average (B = 0.17, 95% CI [0.03, 0.30], p = .01). On the other hand, the effect of Total negative emotions on epistemically suspect beliefs about COVID-19 was not significant when trait emotional intelligence was high. **Despite this finding, the trait emotional intelligence did not significantly moderate the relationship between negative emotions and COVID-19 ESB**.

3.2.1 Testing Well-being as the moderator

The moderation analysis focused on the well-being factor was consistent with the previous findings on Trait emotional intelligence. **Thus, well-being was not a significant moderator in the relationship between negative emotions and COVID-19 ESB** (B = -0.05, 95% CI [-0.19, 0.08], p = .42). Nevertheless, I observed the same trend indicating the effect of negative emotions was significant when well-being was low (B = 0.19, 95% CI [0.05, 0.34], p < .01) or average (B = 0.14, 95% CI [0.002, 0.28], p < .05). When the well-being was high, the effect of negative emotions on COVID-19 ESB was not significant (B = 0.09, 95% CI [-0.14, 0.32], p = .45).

3.2.2 Testing Emotionality as the moderator

Emotionality, as the second factor of trait emotional intelligence, **appeared as the significant moderator of the relationship between total negative emotions and epistemically suspect beliefs about COVID-19** (B = -0.21, 95% CI [-0.35, -0.07], p < .01). The Single slope analysis (Fig. 2.) showed again that negative emotions had significant effect on COVID-19 ESB when emotionality was low (B = 0.36, 95% CI [0.18, 0.53], p < .001) or average (B = 0.16, 95% CI [0.04, 0.29], p < .05). No significant effect was observed when the emotionality was high (B = -0.03, 95% CI [-0.22, 0.16], p = .77).

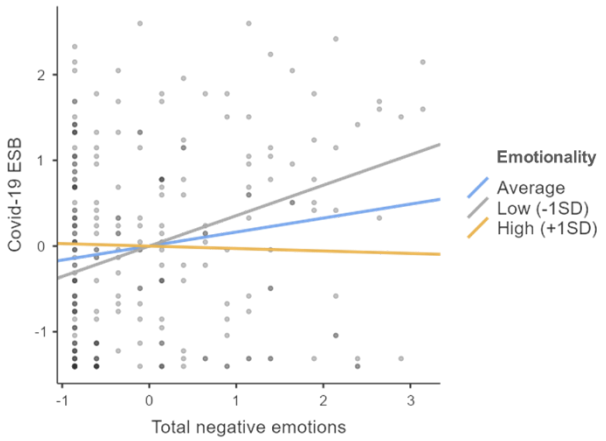


Fig. 2. The Single slope analysis of the Emotionality factor shows that with stronger negative emotions, participants tended to cling to COVID-19 EBS more as their emotionality decreased.

3.2.3 Testing Self-control as the moderator

The third factor of trait emotional intelligence, **self-control, did not moderate the relationship between total negative emotions and COVID-19 ESB significantly** (B = 0.10, 95% CI [-0.05, 0.25], p = .18). Interestingly, regardless the level of self-control (low (B = 0.17, 95% CI [0.02, 0.32], p = .03), average (B = 0.26, 95% CI [0.13, 0.40], p < .001), or high (B = 0.35, 95% CI [0.14, 0.57], p < 0.01)), the effect of total negative emotions on epistemically suspect beliefs about COVID-19 was significant.

3.2.4 Testing Sociability as the moderator

Eventually, the last factor of the trait emotional intelligence, sociability, also **was not the significant moderator of the relationship between total negative emotions and epistemically suspect beliefs about COVID-19** (B = -0.06, 95% CI [-0.19, 0.07], p = .35). The Single slope analysis copied the previous trends in which total negative emotions had significant effect on COVID-19 ESB when sociability was low (B = 0.26, 95% CI [0.10, 0.43], p < .01) or average (B = 0.20, 95% CI [0.07, 0.33], p < .01), but no significant effect was observed when the sociability was high (B = 0.14, 95% CI [-0.05, 0.33], p = .14).

4 Discussion

People in critical times are affected by various adverse events and, as it seems, related negative emotions do associate with susceptibility to conspiracy theory beliefs (Douglas et al., 2020; Galasová & Čavojová, 2022; Šrol et al., 2022). Even the current study results are in line with previous findings and suggestions. Negative emotions of unpleasantness, anger, and disgust positively correlated with epistemically suspect beliefs about COVID-19. Moreover, I observed a significant difference between the two subscales, conspiracy theory and pseudoscientific beliefs. While participants were generally resistant to pseudoscientific beliefs, they showed a higher inclination to conspiracy theory beliefs about COVID-19. Slight differences were also in correlations. Pseudoscientific beliefs positively and significantly correlated with all negative emotions. Conspiracy theory beliefs, on the other hand, positively and significantly correlated only with anger, disgust, and total negative emotions. Yet, all associations were weak indeed. Therefore, a mechanism that makes people more susceptible to epistemically suspect beliefs obviously is not dependent only on negative emotions. As the research suggests, the ability to think analytically is also an important asset that mitigates susceptibility to epistemically suspect beliefs (Šrol, 2021).

Concerning emotional intelligence as a moderator of the relationships between negative emotions and epistemically suspect beliefs about COVID-19, I found no significant result. Thus, emotional intelligence as such does not seem to moderate the relationship between negative emotions and epistemically suspect beliefs about COVID-19. However, the subscale emotionality appeared as a significant moderator of this relationship. The subscale emotionality describes a capacity to perceive, recognize, and understand the emotions of oneself and others. People with high emotionality can consistently recognize their emotions and the emotions of others. They understand the needs of others and support a prosocial environment (Kaliská et al., 2015). Moreover, prosocial behavior also seems to play a role in conspiracy theory beliefs. Recent research shows that people with antisocial tendencies succumbed to conspiracy theory beliefs more than prosocially oriented people (Šrol et al., 2022). However, even prosocial behavior can be suppressed when fear steam-roll rationality of the people (Adamus et al., 2022)

Eventually, current findings of the COVID-19 pandemic show that people from the United Kingdom spent considerably more time online than before the pandemic (Ofcom, 2021). I believe this trend is also typical for other nations around the globe. However, by spending time online, people are exposed to threats related to the spread of misinformation and epistemically suspect beliefs. Research consistently shows that content with negative affect gets shared more than content with positive affect (e.g., Soroka & McAdams, 2015). In addition, conspiracy theory beliefs are supposed to evoke negative emotions (Douglas et al., 2020). Therefore, we should be looking for actions that mitigate such impacts and decrease the polarization of society. An effective way how these negative consequences of an online world and social networks can be mitigated might lie in artificial intelligence. For example, an artificial intelligence application shows promising results in detecting misinformation by detecting emotional appeal from titles and text of news (Paschen, 2019). Interventions like these applied to social media can help administrators reduce misinformation content more successfully and time effectively. However, since the interaction is what counts, and we know that negativity lure more, implementation of such improvements might take a lot of time and consideration from leading social networks.

Acknowledgment

The study was supported by the Scientific grant agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic as part of the project VEGA 2/0053/21: *Examining unfounded beliefs about controversial social issues*.

Ethical Approval and Conflict of Interests

The Ethical Committee of the Centre of Social and Psychological Sciences SAS (CSPS SAS) approved the procedure of the current study. The author states that there is no conflict of interest.

References

Adamus, M., Čavojová, V., & Mikušková, E. (2022). Fear trumps the common good: Psychological antecedents of vaccination attitudes and behaviour. *Acta Psychologica*, *227*. DOI: 10.1016/j.actpsy.2022.103606

Brooks, S., Webster, R., Smith, L., Woodland, L., Wessely, S., Greenberg, N., & Rubin, G. (2020). The Psychological Impact of Quarantine and How to Reduce It: Rapid Review of the Evidence. *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3532534

Douglas, K., Cichocka, A., & Sutton, R. (2020). Motivations, emotions and belief in conspiracy theories. In M. Butter & P. Knight (Eds.), *Routledge Handbook of Conspiracy Theories* (1st ed., pp. 181 - 191). Routledge. Retrieved 13 May 2022, from DOI: 10.4324/9780429452734.

Galasová, M., & Čavojová, V. (2022). Negative emotions, emotional intelligence, and conspiracy mentality [accepted manuscript]. In E. Aigelová, L. Viktorová, & M. Dolejš (Eds.), *Proceedings of the Czech & Slovak Psychological Conference PhD Existence*, Univerzita Palackého v Olomouci.

Kaliská, L., Nábělková, E., & Salbot, V. (2015). *Dotazník črtovej emocionálnej inteligencie (TEIQue–SF/TEIQue-CSF): manuál k skráteným formám* (1st ed.). Banská Bystrica: Belianum.

Mayer, J. D., & Salovey, P. (1997). What is emotional intelligence? In P. Salovey, & D. J. Sluyter (Eds.), *Emotional development and emotional intelligence: Educational implications* (pp. 3-31). New York, NY. Basic Books.

Metzler, H., Rimé, B., Pellert, M., Niederkrotenthaler, T., Di Natale, A., & Garcia, D. (2021). Collective Emotions during the COVID-19 Outbreak. DOI: 10.31234/osf.io/qejxv

Mostajo-Radji, M. (2021). Pseudoscience in the Times of Crisis: How and Why Chlorine Dioxide Consumption Became Popular in Latin America During the COVID-19 Pandemic. *Frontiers In Political Science*, *3*. DOI: 10.3389/fpos.2021.621370

Ofcom. (2021). *Online Nation*. Retrieved from https://www.ofcom.org.uk/__data/assets/pdf_file/0013/220414/online-nation-2021-report.pdf

Paschen, J. (2019). Investigating the emotional appeal of fake news using artificial intelligence and human contributions. *Journal Of Product &Amp; Brand Management*, *29*(2), 223-233. DOI: 10.1108/jpbm-12-2018-2179

Petrides, K. V. (2009). Psychometric properties of the Trait Emotional Intelligence Questionnaire (TEIQue). In C. Stough, D. H. Saklofske, & J. D. A. Parker (Eds.), *Assessing emotional intelligence: Theory, research, and applications* (pp. 85–101). Springer Science + Business Media. DOI: 10.1007/978-0-387-88370-0_5

Petrides, K. V., & Furnham, A. (2006). The role of trait emotional intelligence in a gender-specific model of organizational variables. *Journal of Applied Social Psychology*, *36*, 552–569. DOI: 10.1111/j.0021-9029.2006.00019.x

Porubcin, S., Rovnakova, A., Zahornacky, O., & Jarcuska, P. (2022). Intravenous veterinary Ivermectin in a COVID-19 patient causing neurotoxicity. *Idcases*, *27*, e01446. DOI: 10.1016/j.idcr.2022.e01446

Soroka, S., & McAdams, S. (2015). News, Politics, and Negativity. *Political Communication*, *32*(1), 1-22. DOI: 10.1080/10584609.2014.881942

Šrol, J. (2021). Individual differences in epistemically suspect beliefs: the role of analytic thinking and susceptibility to cognitive biases, *Thinking and Reasoning*, *28*(1), pp. 125-162. DOI: 10.1080/13546783.2021.1938220. M.

Šrol, J., Čavojová, V., & Ballová Mikušková, E. (2022). Finding Someone to Blame: The Link Between COVID-19 Conspiracy Belief, Prejudice, Support for Violence, and Other Negative Social Outcomes. *Frontiers In Psychology*, *12*. DOI: 10.3389/fpsyg.2021.726076

Teličák, P., & Halama, P. (2020) *Covid-19 škála nepodložených presvedčení* [Covid-19 scale of epistemically suspect beliefs]. Unpublished manuscript. Institute of the experimental psychology, CSPS SAS.

Thompson, E. R. (2007). Development and Validation of an Internationally Reliable Short-Form of the Positive and Negative Affect Schedule (PANAS). *Journal of CrossCultural Psychology*, *38*(2), 227–242. DOI: 10.1177/0022022106297301

Van Mulukom, V., Pummerer, L., Alper, S., Bai, H., Cavojova, V., & Farias, J. et al. (2022). Antecedents and consequences of COVID-19 conspiracy beliefs: a systematic review. *Social Science & Medicine*, *301*. DOI: 10.1016/j.socscimed.2022.114912

Body schema or the body as its own best model

Matej Hoffmann

Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague
Email: matej.hoffmann@fel.cvut.cz

Abstract

Rodney Brooks (1991) put forth the idea that during an agent's interaction with its environment, representations of the world often stand in the way. Instead, using the world as its own best model, i.e. interacting with it directly without making models, often leads to better and more natural behavior. The same perspective can be applied to representations of the agent's body. I analyze different examples from biology—octopus and humans in particular—and compare them with robots and their body models. At one end of the spectrum, the octopus, a highly intelligent animal, largely relies on the mechanical properties of its arms and peripheral nervous system. No central representations or maps of its body were found in its central nervous system. Primate brains do contain areas dedicated to processing body-related information and different body maps were found. Yet, these representations are still largely implicit and distributed and some functionality is also offloaded to the periphery. Robots, on the other hand, rely almost exclusively on their body models when planning and executing behaviors. I analyze the pros and cons of these different approaches and propose what may be the best solution for robots of the future.

1 Introduction

In artificial intelligence and robotics, models of the world have been and largely still are the key means of realizing interaction of a mechanism with its environment. This position was attacked by Brooks (1990) stating: "The key observation is that the world is its own best model. It is always exactly up to date. It always contains every detail there is to be known. The trick is to sense it appropriately and often enough." Brooks (1991) added "When we examine very simple level intelligence we find that explicit representations and models of the world simply get in the way."

If this were the case for the world, how about for the body of an agent—human, animal, or robot? Our body seems to be even more "always there" than our environment. The representationalist stance typical of robotics and (good old-fashioned) artificial intelligence (Haugeland, 1985) is also applied to the body. Indeed, traditional robots heavily rely on internal models of their bodies. These are in particular the models of their kinematics—joints and links, their dimensions and orientations—and their dynamics which deal with masses and forces needed to generate motion (see 10.2.3 in (Hoffmann, 2021) for more details). With traditional robots, the interaction with the world is mediated by these models. In cognitive science, this approximately corresponds to the "body in the brain" approach which emphasizes representations of the body in the cerebral cortex—see for example Section 5.1 in (De Vignemont, 2018). This contrasts with the "brain in the body" or "body in the world" perspective, also called the sensorimotor approach that is in line with Brooks' perspective (see Section 4.2 in (De Vignemont, 2018) or a discussion in the Introduction to (Ataria et al., 2021)).

This work draws on (Hoffmann, 2021, 2022; Hoffmann and Müller, 2017).

2 Biology – from octopus to humans

The octopus constitutes an interesting case. Belonging to cephalopods, highly derived molluscs, it is the most intelligent among them and with the largest nervous system. Cephalopods, the most advanced invertebrate class, feature, on one hand, the highest centralization of the nervous system. On the other hand, next to the central nervous system (CNS) composed of the brain and two optic lobes, there is a large peripheral nervous system (PNS) of the body and the arms. Despite the high level of centralization and in contrast to vertebrate and insect brains, there is no obvious somatotopic arrangement in either motor or sensory areas (see (Zullo and Hochner, 2011) for more details). The octopus also has a unique embodiment—a flexible body and eight arms with virtually infinite degrees of freedom. From an engineering perspective, modeling and controlling such a body (*plant* in engineering jargon) using inverse kinematics and dynamics would be a nightmare. However, Yekutieli et al. (2005) speculate that the octopus reaches toward a target using the following strategy: (1) Initiating a bend in the arm so that the suckers point outward. (2) Orienting the base of the arm in the direction of the target or just above it. (3) Propagating the bend along the arm at the desired speed by a wave of muscle activation that equally activates all muscles along the arm. (4) Terminating the reaching movement when the suckers touch the target by stopping the bend propagation and thus catching the target. A big part of the complexity is

thus "off-loaded" from to the peripheral nervous system and the body itself.

In humans, central representations of the body in the cerebral cortex certainly exist. There has been more than a century of empirical observations and theorizing, leading to concepts like body image (system of perceptions, attitudes, and beliefs pertaining to one's own body) and body schema (system of sensory-motor capacities that function without awareness or the necessity of perceptual monitoring) (definitions taken from the Introduction to (Ataria et al., 2021)). The most well-known body maps are the somatotopic representations (the "homunculi") in the primary motor and somatosensory cortices (Leyton and Sherrington, 1917; Penfield and Boldrey, 1937). Yet, the somatosensory homunculi are only an "entry point" or "relay station" to downstream cortical processing rather than accurate representations or models of the body (e.g., (Longo and Haggard, 2010)). Downstream areas in the posterior parietal cortex (like Brodmann area 5) are thought to be involved in higher-level more integrated representations related to the configuration of the body in space, for example, but detailed understanding is still missing. Reaching in primates bears some similarity to that in the octopus. A reaching movement has some high-level characteristics like the direction of a hand's movement in space, the extent of the movement (amplitude), the overall duration (movement time), and other parameters such as anticipated level of resistance to the movement (Schöner et al., 2018). Also, movement generation involves cooperation between the CNS and PNS. The exact mechanisms of motor control in humans and other primates are still debated. Compared to invertebrates, motor control in vertebrates, specifically mammals and in particular primates, becomes more "cortical" and the motor cortex has the possibility of more direct control over the details of a particular movement, which is likely correlated with the need for dexterous manipulation (see 10.2.2 in Hoffmann (2021) for more details).

3 Body models for controlling movements

Body models can be classified according to different characteristics, such as fixed vs. adaptive, amodal vs. modal, explicit vs. implicit, serial vs. parallel, modular vs. holistic, or centralized vs. distributed (Hoffmann, 2021). For this article, we focus on the dimensions shown in Fig. 1.

3.1 Explicit and veridical versus implicit and action-oriented

Traditional robot body models are explicit; it is clear what in the model corresponds to what in the body (e.g., a certain parameter to the length of the left forearm). They are also objective and veridical; the parameters should be the true physical values of the quantities (lengths, angles, masses, etc.). This is illustrated by the iCub humanoid robot (Metta et al., 2010) and its models positioned at the far left in Fig. 1 A. In the biological realm, representations in general are not like that and this should hold for representations of the body as well. "What the nervous system needs to do, in general, is to transform the input into the right action" (Webb, 2006)—hence the implicit and action-oriented character of the representations. The octopus—with no known map of its body in its central nervous system—is positioned at the opposite end of the spectrum (Fig. 1 E). Successful action is also the only criterion for the "quality" of what is represented about the animal's body in its brain; there is no need for any objective or veridical representation. Similar arguments hold for primate brains, but to a lesser extent. Numerous sites dedicated to representing the body were found (e.g., Kanayama and Hiromitsu (2021) for a review). Compared to the octopus, much more of the body seems more explicitly represented. Longo (2015) considers the implicit–explicit axis within human body representations and draws a line roughly between the body schema and the body image. In tasks more related to action and where humans do not consciously represent their body, the body models seem more implicit and also less accurate. These representations may also be dominated by somatosensation and inherit some of the distortions typical of the somatosensory homunculi. Conversely, tasks that relate to conscious perception of our body seem to draw on more explicit representations that are also more accurate/veridical (e.g., image of our hand). This is schematically illustrated in Fig. 1, D.

3.2 Centralized, universal, modular versus distributed, specialized, end-to-end

Robot models are normally centralized—exist only in one place in the robot software. On the other hand, neural representations are known to be distributed. Whereas this "spatial aspect" may be also related to the computational substrate (computers versus neurons), more important is a functional division. Albeit centralized, robot body models are highly *modular*. For the iCub (Fig. 1 A), there would normally be a single model of its kinematics and another one of its dynamics (mass distribution etc.). Then, there are distinct modules like forward/inverse kinematics and dynamics that may draw from the same robot model and be recruited for different purposes like state estimation, movement planning etc. There would be typically only one module of every kind (imagine a software library) providing this functionality upon request. The representations/modules will thus be *universal* and not overlapping. For deep learning applied to robotics, this is not the case. Levine et al. (2018) specialize on a single task (grasping); a different task will likely need a different network. The representations

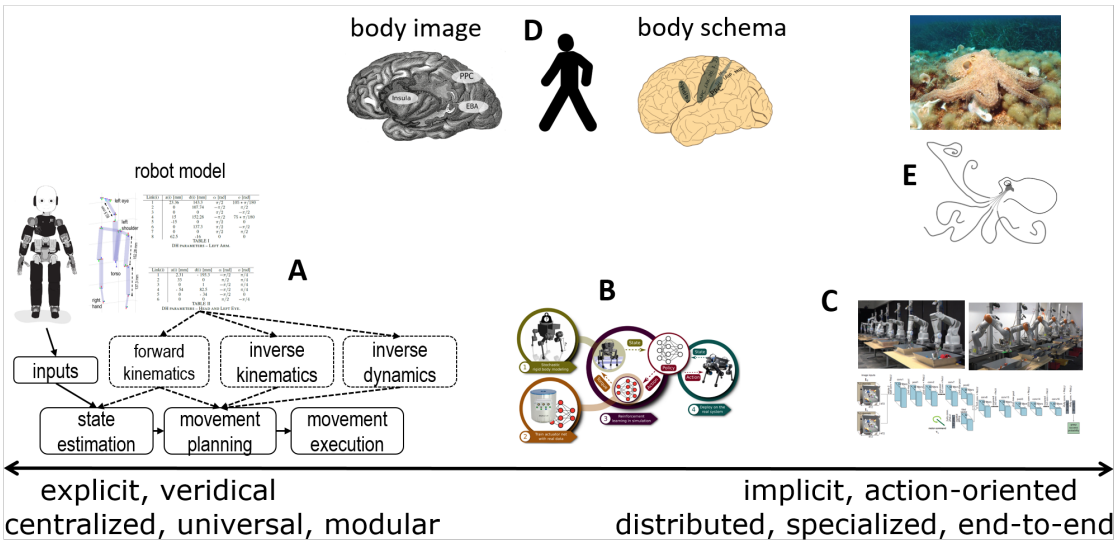


Fig. 1: Body model characteristics. Upper row: examples from biology. Lower row: examples from robotics. (A) iCub humanoid robot and its models. (B) Hybrid model of the ANYmal robot (Hwangbo et al., 2019). (C) Robot manipulators learning to grasp end-to-end (Levine et al., 2018). (D) Human and schematic illustration of brain areas important for body representations. Brain areas involved in body image representation after Berlucchi and Aglioti (2010). (E) Octopus and schematic of its nervous system.
Credit: A – iCub cartoon: Laura Taverna, Italian Institute of Technology. Credit: D – Walking human: Public domain (https://commons.wikimedia.org/wiki/File:BSicon_WALK.svg). Credit D – Brain image source: Hugh Guiney / Attribution-ShareAlike 3.0 Unported (CC BY- SA 3.0). Credit: E – Common octopus - albert kok / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0). Credit: E – Octopus nervous system - Jean-Pierre Bellier / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0).

are thus end-to-end, task-specific and in case of multiple tasks also overlapping. In nervous systems, there may also be complete sensorimotor loops specialized on individual tasks, partially overlapping or redundant. However, this approach does not scale well. In primates, the posterior parietal cortex is regarded as a site where information about the body from different modalities converges. Specific areas related to representations of body parts or reaching targets in different reference frames have been found. These are recruited in different tasks or contexts and hence, there is certain universality and modularity—again more for the body image than body schema Fig. 1, D.

4 Use the body directly

While body representations can take very different forms, one should at the same time consider the radical possibility of using the body directly rather than through an internal model. Again, as Brooks (1990) put it: "The key observation is that the world is its own best model. It is always exactly up to date. It always contains every detail there is to be known. The trick is to sense it appropriately and often enough." In fact, for the case of the body, one can do even without sensing its state. First, there are examples how a completely passive body, "pure physics", can generate useful behavior. In the biological realm, this is for example the body of a trout. Liao (2004) shows that, paradoxically, under specific circumstances, a dead trout body can exploit vortices in the water to the extent that it swims upstream. In robotics, a similar well-known example are the passive dynamics walkers McGeer (1990)—carefully designed mechanical devices that walk down a ramp without any motors, sensors, or controllers. Second, in case there is actuation, we have privileged access to the body current or future state (using forward models / efference copies (Webb, 2004)) and hence, it may be unnecessary to sense it.

Different positions on the imaginary landscape ranging from model-based control to direct use of the body are illustrated in Fig. 2. The passive dynamic walker is positioned at the far right of the schematics. As discussed above, the octopus is able to reach for visual targets, but it may not know—and may not need to know—how long its arm is or where it is exactly in space. Orienting the base of the arm and propagating the bend until contact is detected by the suckers may well suffice. The need to represent the body, its state, and the complex inverse kinematics and dynamics has been largely offloaded to embodiment—the properties of the octopus arm, supported by the peripheral nervous system and low-dimensional inputs from the

central nervous system. Human reaching, Fig. 2 D, is probably less embodied compared to the octopus, but still sharing some important characteristics. Cisek and Kalaska (2003) highlight the importance of online, dynamically generated character of movement generation in primates. At the same time, they also point out that due to conduction delays inherent to the sensorimotor system, purely feedback control is limited, or at least slow. Thus, feed-forward commands and local neural reflex loops have to work in concert. Robots, on the other hand, typically heavily rely on models. Importantly, this is the case also for the solutions employing deep learning. In (Levine et al., 2018), Fig. 2 C, the embodiment of the robot arm or the gripper is not significantly exploited.

5 Robots: with or without a model?

Mechanical engineers naturally think in terms of how to make the best design of a machine for a task. However, control engineers have a strong preference for model-based control. Moreover, solutions for nonlinear systems are much more difficult to obtain, and they often involve a linearization of the system of some sort. Thus, complex (highly dimensional, dynamic, nonlinear, compliant, deformable, 'soft') robot bodies are avoided as they cannot be modeled and controlled with the available methods. Many robot engineers then simply take the body as fixed and seek to exploit to the maximum what can be done at the "software level".

Including the parameters of the body into the design considerations may give rise to better performance of the whole system; these may be solutions involving a simpler controller, but also solutions that were previously unattainable when the body was fixed. Following the dynamical systems' perspective, Füchslin et al. (2013) provide an illustration of the possible goals of the design process: (1) To design the physical dynamical system such that desired regions of the state space have attracting properties. Then it is sufficient to use a simple control signal that will bring the system to the basins of attraction of individual stable points that correspond to target behaviors. (2) More complicated behavior can be achieved if the attractor landscape can be manipulated by the control signal.

If a mathematical formulation of the controller and the plant is available, this design methodology can be directly applied. The first part is demonstrated by on the passive dynamic walker (McGeer, 1990): The influence of scale, foot radius, leg inertia, height of center of mass, hip mass and damping, mass offset, and leg mismatch is evaluated. In addition, the stability of the walker is calculated. Jerrold Marsden and his coworkers presented a method that allows for co-optimization of the controller and plant by combining an inner loop (with discrete mechanics and optimal control) and an outer loop (multiscale trend optimization). They applied it to a model of a walker and obtained the best position of the knee joints ((Pekarek, 2010) – Ch. 5). However, typical real-world agents are more complex than simple walkers. Holmes et al. (2006) provide an excellent dynamical systems analysis of the locomotion of rapidly running insects and derive implications for the design of the RHex robot. Yet, they conclude that "a gulf remains between the performance we can elicit empirically and what mathematical analyses or numerical simulations can explain. Modeling is still too crude to offer detailed design insights for dynamically stable autonomous machines in physically interesting settings." Modeling and optimization of more complicated morphologies—like compliant structures—is nevertheless an active research topic (e.g., (Wang, 2009)). The second point of Füchslin et al. (2013)—achieving "morphological programmability" by constructing a dynamical system with a parametrized attractor landscape—remains even more challenging though.

One of the merits of exploiting the contributions of body morphology should be that the physical processes do not need to be modeled, but can be used directly. However, without a model of the body at hand, several body designs need to be produced and—together with the controller—tested in the respective task setting. The design space of the joint controller-body system blows up and we may be facing a curse of dimensionality. This is presumably the strategy adopted by the evolution of biological organisms that could cope with the enormous dimensionality of the space. In robotics, this has been taken up by evolutionary robotics (Nolfi and Floreano, 2000). The simulated agents of Sims (1994) demonstrate that co-evolving brains and bodies together can give rise to unexpected solution to problems. Bongard (2011) showed that morphological change indeed accelerates the evolution of robust behavior in such a brain-body co-evolution setting. With the advent of rapid prototyping technologies, physics-based simulation could be complemented by testing in real hardware Lipson and Pollack (2000), but this reintroduces the modeling through the back door: the phenotypes in the simulator now become models and they need to sufficiently match their real counterparts. Yet, a "reality gap" (Jakobi et al., 1995; Koos et al., 2013) always remains between simulated and real physics. The only alternative is to optimize in hardware directly, which is in general slow and costly. Brodbeck et al. (2015) provide an interesting illustration how locomoting cube-like creatures can be evolved in a model-free fashion through automated manufacturing and testing. However, in summary, the design decisions—which parameters to optimize—are based on heuristics and a clear methodology is still missing. Furthermore, with the absence of an analytical model of the controller and plant, no guarantees on the system's performance can be given.

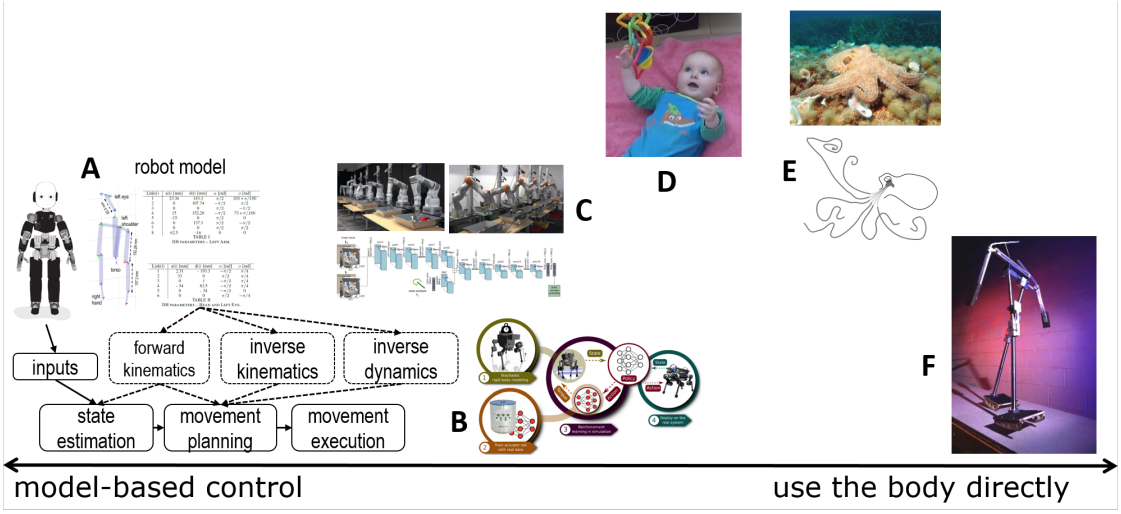


Fig. 2: Model-based control or direct use of the body. (A) iCub humanoid robot and its models. (B) Model of the ANYmal robot (Hwangbo et al., 2019). (C) Robot manipulators learning to grasp end-to-end (Levine et al., 2018). (D) Infant reaching. (E) Octopus and schematic of its nervous system. (F) The Cornell passive biped with arms Collins et al. (2005).
Credit: A, E – see Fig. 1. Credit F: H. Morgan.

6 Conclusion and outlook

Rich properties of complex bodies (highly dimensional, dynamic, nonlinear, compliant and deformable) have been mostly overlooked or deliberately suppressed by classical mechatronic designs, as they are largely incompatible with traditional control frameworks, where linear plants are preferred. This is definitely a missed opportunity. On the other hand, while complex bodies carry a lot of "self-control" potential, this property does not come for free. It has to be said that the exploitation of truly complex bodies to accomplish tasks is still mostly at a "proof-of-concept" stage. A closely connected issue is the one of modeling of these systems—complex, or for example soft, bodies are notoriously difficult to model. The model may not be necessary for the system to perform the task; however, without a model, the understanding and design is more complicated and performance guarantees are limited. The field, which has been dominated by heuristics so far, needs to embrace more systematic approaches that allow to navigate in this complex landscape.

The area of soft robotics and morphological computation/morphological control/morphology facilitating control (Füchslin et al., 2013; Müller and Hoffmann, 2017) is rife with different trading spaces (Pfeifer et al., 2013). As we move from the traditional engineering framework with a central controller that commands a "dumb" body toward delegating more functionality to the physical morphology, some convenient properties will be lost. In particular, the solutions may not be portable to other platforms anymore, as they will become dependent on the particular morphology and environment (the passive dynamic walker is the extreme case). The versatility of the solutions is likely to drop as well. To some extent, the morphology itself can be used to alleviate these issues—if it becomes adaptive. Online changes of morphology (like changes of stiffness or shape) thus constitute another tough technological challenge. Completely new, distributed control algorithms that rely on self-organizing properties of complex bodies and local distributed control units will need to be developed (McEvoy and Correll, 2015; Rieffel et al., 2010).

In summary, computer scientists, roboticists, and control engineers impose a representationalist perspective on designing machines and their behaviors. This is similar to traditional cognitive science (cognitivism). It is sometimes acknowledged that the representations—world or body—should be embodied. However, rather than "embodied body models", it seems more natural to think of the "brain in the body" or "body in the world" (cf. discussion in (Alsmith and De Vignemont, 2012; De Vignemont, 2018; Ataria et al., 2021), and more direct use of the body wherever possible. For engineers, this will be a major challenge though.

Acknowledgement

This work was supported by the Czech Science Foundation (GA CR), project no. 20-24186X. I would like to thank to Rolf Pfeifer for discussions along these lines.

References

Alsmith, A. J. T. and De Vignemont, F. (2012). Embodying the mind and representing the body. *Review of Philosophy and Psychology*, 3(1):1–13.

Ataria, Y., Tanaka, S., and Gallagher, S. (2021). *Body Schema and Body Image: New Directions*. Oxford University Press.

Berlucchi, G. and Aglioti, S. M. (2010). The body in the brain revisited. *Experimental brain research*, 200(1):25–35.

Bongard, J. (2011). Morphological change in machines accelerates the evolution of robust behavior. *Proceedings of the National Academy of Sciences*, 108(4):1234–1239.

Brodbeck, L., Hauser, S., and Iida, F. (2015). Morphological evolution of physical robots through model-free phenotype development. *PLoS ONE*, 10(6):e0128444.

Brooks, R. A. (1990). Elephants don't play chess. *Robotics and autonomous systems*, 6(1-2):3–15.

Brooks, R. A. (1991). Intelligence without representation. *Artificial intelligence*, 47(1-3):139–159.

Cisek, P. and Kalaska, J. (2003). Reaching movements: implications for computational models. In *Handbook of Brain Theory and Neural Networks*, pages 945–948. MIT Press.

Collins, S., Ruina, A., Tedrake, R., and Wisse, M. (2005). Efficient bipedal robots based on passive dynamic walkers. *Science*, 307:1082–1085.

De Vignemont, F. (2018). *Mind the body: An exploration of bodily self-awareness*. Oxford University Press.

Füchslin, R., Dzyakanchuk, A., Flumini, D., Hauser, H., Hunt, K., Luchsinger, R., Reller, B., Scheidegger, S., and Walker, R. (2013). Morphological computation and morphological control: steps towards a formal theory and applications. *Artificial Life*, 19(1):9–34.

Haugeland, J. (1985). Artificial intelligence: the very idea.

Hoffmann, M. (2021). Body models in humans, animals, and robots: mechanisms and plasticity. In Ataria, Y., Tanaka, S., and Gallagher, S., editors, *Body Schema and Body Image: New Directions*, pages 152–180. Oxford University Press.

Hoffmann, M. (2022). Biologically inspired robot body models and self-calibration. In Ang, M. H., Khatib, O., and Siciliano, B., editors, *Encyclopedia of Robotics*, pages 1–14. Springer Berlin Heidelberg, Berlin, Heidelberg.

Hoffmann, M. and Müller, V. C. (2017). Simple or complex bodies? trade-offs in exploiting body morphology for control. In Dodig-Crnkovic, G. and Giovagnoli, R., editors, *Representation and Reality in Humans, Other Living Organisms and Intelligent Machines*, Studies in Applied Philosophy, Epistemology, and Rational Ethics (SAPERE), pages 335–345. Springer.

Holmes, P., Full, R. J., Koditschek, D., and Guckenheimer, J. (2006). The dynamics of legged locomotion: Models, analyses and challenges. *SIAM Review*, 48(2):207–304.

Hwangbo, J., Lee, J., Dosovitskiy, A., Bellicoso, D., Tsounis, V., Koltun, V., and Hutter, M. (2019). Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26).

Jakobi, N., Husbands, P., and Harvey, I. (1995). Noise and the reality gap: The use of simulation in evolutionary robotics. In *Advances in artificial life*, pages 704–720. Springer.

Kanayama, N. and Hiromitsu, K. (2021). Triadic body representations in the human cerebral cortex and peripheral nerves. In Ataria, Y., Tanaka, S., and Gallagher, S., editors, *Body Schema and Body Image: New Directions*, pages 133–151. Oxford University Press.

Koos, S., Mouret, J.-B., and Doncieux, S. (2013). The transferability approach: Crossing the reality gap in evolutionary robotics. *Evolutionary Computation, IEEE Transactions on*, 17(1):122–145.

Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J., and Quillen, D. (2018). Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 37(4-5):421–436.

Leyton, A. S. and Sherrington, C. S. (1917). Observations on the excitable cortex of the chimpanzee, orang-utan, and gorilla. *Quarterly Journal of Experimental Physiology: Translation and Integration*, 11(2):135–222.

Liao, J. C. (2004). Neuromuscular control of trout swimming in a vortex street: implications for energy economy during the karman gait. *Journal of Experimental Biology*, 207(20):3495–3506.

Lipson, H. (2013). Challenges and opportunities for design, simulation, and fabrication of soft robots. *Soft Robotics*, 1:21–27.

Lipson, H. and Pollack, J. (2000). Automatic design and manufacture of robotic lifeforms. *Nature*, 406(6799):974–978.

Longo, M. R. (2015). Implicit and explicit body representations. *European Psychologist*, 20(1):6–15.

Longo, M. R. and Haggard, P. (2010). An implicit body representation underlying human position sense. *Proceedings of the National Academy of Sciences*, 107(26):11727–11732.

McEvoy, M. and Correll, N. (2015). Materials that couple sensing, actuation, computation, and communication. *Science*, 347(6228):1261689.

McGeer, T. (1990). Passive dynamic walking. *The International Journal of Robotics Research*, 9(2):62–82.

Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., Von Hofsten, C., Rosander, K., Lopes, M., Santos-Victor, J., et al. (2010). The iCub humanoid robot: An open-systems platform for research in cognitive development. *Neural networks*, 23(8-9):1125–1134.

Müller, V. and Hoffmann, M. (2017). What is morphological computation? on how the body contributes to cognition and control. *Artificial Life*, 23(1):1–24.

Nolfi, S. and Floreano, D. (2000). *Evolutionary robotics: The biology, intelligence, and technology of self-organizing machines*, volume 26. MIT press Cambridge.

Pekarek, D. (2010). *Variational methods for control and design of bipedal robot models*. PhD thesis, California Institute of Technology.

Penfield, W. and Boldrey, E. (1937). Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain*, 37:389–443.

Pfeifer, R., Marques, H., and Iida, F. (2013). Soft robotics: the next generation of intelligent machines. In *Proc. 23rd Int. Joint Conf. on Artificial Intelligence*, pages 5–11. AAAI Press.

Rieffel, J., Valero-Cuevas, F., and Lipson, H. (2010). Morphological communication: exploiting coupled dynamics in a complex mechanical structure to achieve locomotion. *Journal of the Royal Society Interface*, 7(45):613–621.

Schöner, G., Tekülve, J., and Zibner, S. (2018). Reaching for objects: a neural process account in a developmental perspective. In *Reach-to-Grasp Behavior*, pages 281–318. Routledge.

Sims, K. (1994). Evolving 3D morphology and behavior by competition. *Artificial Life*, 1(4):353–372.

Wang, M. Y. (2009). A kinetoelastic formulation of compliant mechanism optimization. *Journal of Mechanisms and Robotics*, 1(2):021011.

Webb, B. (2004). Neural mechanisms for prediction: do insects have forward models? *Trends in Neurosciences*, 27:278–282.

Webb, B. (2006). Transformation, encoding and representation. *Current Biology*, 16(6):R184–R185.

Yekutieli, Y., Sagiv-Zohar, R., Hochner, B., and Flash, T. (2005). Dynamic model of the octopus arm. II. control of reaching movements. *Journal of neurophysiology*, 94(2):1459–1468.

Zullo, L. and Hochner, B. (2011). A new perspective on the organization of an invertebrate brain. *Communicative & integrative biology*, 4(1):26–29.

Kompozitní testování inteligence jako možná cesta k univerzální psychometrii

Petr Hoza, Ondřej Vadinský[0000−0002−0910−3140]

Vysoká škola ekonomická v Praze
Náměstí Winstona Churchilla 4, 130 67 Praha 3, ČR
Email: hozp00@vse.cz, ondrej.vadinsky@vse.cz

Abstrakt

Příspěvek předkládá popis výzkumného záměru v rámci vypracovávání diplomové pracé. Jedná se o návrh kompozitního testu složeného ze základního testu (test první úrovně) a skrytého testu (test druhé úrovně), který by mohl být slibným kandidátem na univerzální inteligenční test. Cílem navrhovaného výzkumu je ověřit, zda se u lidí při vypracovávání testů projevují i jejich jiné dílčí cíle, jak se tyto dílčí cíle projevují při vypracovávání testů a v neposlední řadě zda je adaptace na tyto dílčí cíle korelována s IQ.

1 Úvod

> *„Abychom věděli, zda má dítě inteligenci odpovídající jeho věku, zda trpí retardací nebo je pokročilý, a jak moc, potřebujeme mít přesnou a skutečně vědeckou metodu.“*
>
> (Binet, 1909)

Dostatečné a kvalitní informace jsou jedním z nejdůležitějších základů při jakémkoliv plánování či rozhodování. Pro sběr relevantních informací o vlastnostech jednotlivců slouží obor psychometrie. Možnosti měření lidí jsou velice široké a můžeme jejich pomocí měřit celou škálu vlastností, od inteligence, přes zájmy, znalosti a osobnost, až po potenciál (Salkind, 2017). Jedním z důležitých aspektů psychometrie je tím pádem poskytovat zpětnou vazbu o jednotlivcích. Tu pak můžeme různými způsoby využít, například ke správnému hodnocení, rozdělování a pozorování jednotlivců. Jelikož se zdá, že se všechny vlastnosti jednotlivců v čase mění (Heckman a Zhou, 2022), můžeme sledovat jejich vývoj a posuzovat, zda se vyvíjí žádoucím způsobem.

V dnešní době překotného technologického rozvoje je potřeba sledování žádoucího vývoje umělých inteligencí obzvláště důležitá (více v Sekci 4.1). Nejčastěji jsou agenty posuzovány pouze v konkrétních, specifických úlohách, proto jsou obdobně široké a spolehlivé nástroje měření, jaké máme pro lidi, potřeba i pro tyto umělé agenty. Tímto se zabývá obor **univerzální psychometrie**, který si v této práci představíme v Sekci 2. Dalé si v Sekci 3 shrneme současný stav testování a v Sekci 4 některá úskalí současného testování. Sekce 5 shrne výzkumný záměr ověření premisy, že „lidé podle potřeby mění v průběhu testů svůj dílčí cíl“, která by nám mohla pomoci se sestavením univerzálního testu inteligence. Sekce 6 popisuje současný a budoucí stav probíhajících prací a v závěru Sekce 7 shrne možné nedostatky projektu.

2 Univerzální psychometrie

Jednou z důležitých vlastností, kterou se zabývá psychometrie je inteligence. V doméně umělých agentů zavádí Legg a Hutter (2007a) pojem **univerzální inteligence**. Z této definice budeme v práci vycházet: *„Inteligence měří schopnost agenta dosahovat cílů v široké škále prostředí“*.

Pro tuto definici inteligence identifikují 3 hlavní aspekty:

- Inteligence je vlastnost, kterou má jednotlivý agent při interakci se svým prostředím.
- Inteligence souvisí se schopností agenta uspět nebo profitovat s ohledem na nějaký cíl nebo prostředí.
- Inteligence závisí na schopnosti agenta adaptovat se na různé cíle a prostředí.

„Mnoho z toho, co víme o inteligenci, pochází z psychometrie“ (Hernandez-Orallo, 2017). Lepší psychometrické nástroje pro doménu umělých agentů by nám proto mohly pomoci lépe pochopit jejich inteligenci.

Se zvyšující se komplexitou navíc umělé agenty přebírají stále složitější úkoly z domény lidí. Nabízí se proto otázka, zda by nebylo přínosné testovat subjekty dokonce napříč těmito doménami a zda se dají techniky klasické psychometrie použít i na jiné typy agentů, než biologické. Touto otázkou se zabývá obor **univerzální psychometrie**[1], ve kterém zavádí Hernández-orallo a spol. (2012) pojem **strojové království**, sestávající ze všech možných typů subjektů či dokonce jejich kombinací. *„Univerzální psychometrie je disciplína zkoumající měření kognitivních schopností jakýchkoliv (kognitivních) systémů, individuálnıch nebo kolektivnıch, umělých, biologických nebo hybridních“* (Hernández-Orallo a Dowe, 2010).

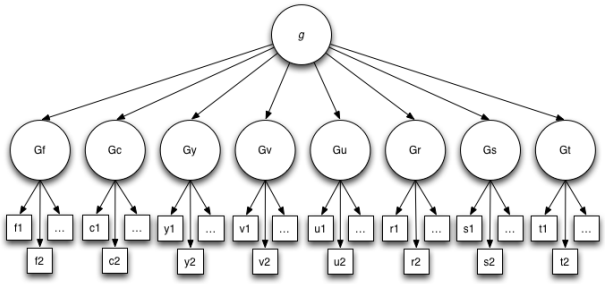
[1] Ucelený výklad o univerzální psychometrii podává Hernandez-Orallo (2017).

3 Současný stav testování kognitivních schopností

V této kapitole si řekneme trochu více o současném stavu testování dílčích domén a pokusech o univerzální testy.

3.1 Testování biologických agentů

Testování lidí je v současnosti opravdu velice širokým oborem, kde se testuje celá škála vlastností. My se zaměříme na testování inteligence. Pro naše potřeby budeme vycházet z teorie Cattell-Horn-Carroll (CHC) (Keith a Reynolds, 2010), ze které mnoho moderních testů vychází. CHC rozděluje inteligenci do tří úrovní. Na obecnou inteligenci, takzvaný g-faktor definovaný v Spearman (1904) a další dvě úrovně. Druhá úroveň sestává z několika **širších schopností** a třetí z ‚úzce' zaměřených schopností v dílčích úlohách (Obr. 1).



Obr. 1: Cattell-Horn-Carroll třívrstvý model s g na třetí úrovni s deseti širokými schopnostmi na druhé úrovni: Fluidní inteligence (Gf), Krystalizovaná inteligence (Gc), Kvantitativní uvažování (Gq), Čtení a psaní (Gr), Krátkodobá paměť (Gm), Dlouhodobá paměť (Gl), Vizuální zpracování (Gv), Sluchové zpracování (Ga), Rychlost zpracování (Gs), Rozhodování/Reakční doba/Rychlost (Gt). První úroveň může zahrnovat mnohem více ‚úzkých' schopností (Bates, 2013).

Hlavní výhodou g-faktoru je jeho obecnost. *„Hodnoty IQ silně závisí na složení úloh v IQ testu, zatímco skóre g je mnohem méně závislé“* (Hernandez-Orallo, 2017). Přesto jsou IQ testy často používány k měření inteligence u lidí. *„Inteligence bývá obvykle definována jako Spearmanův g-faktor [...] a bývá hodnocena pomocí IQ testů, o kterých se předpokládá, že jsou dobrými měřítky g-faktoru“* (Cianciolo a Sternberg, 2008). Přesto pro naše potřeby není IQ test dobrým univerzálním testem, protože i celkem jednoduchý program může úspěšně řešit IQ testy (Sanghi a Dowe, 2003).

3.2 Testování umělých agentů

Při vyhodnocování umělých agentů převládá v klasické umělé inteligenci jejich testování na konkrétních úlohách. O měření inteligence umělých agentů se pak v posledních letech snaží obecné přístupy ukotvené v algoritmické teorii informace (AIT) (Hernández-Orallo a Martínez-Plumed, 2016). Mezi obecné testy vycházející z AIT patří **kdykoliv prerušitelný test inteligence** navržený v Hernández-Orallo a Dowe (2010) a **test algoritmickeho IQ (AIQ test)** (Legg a Veness, 2013). [2]

Chollet (2019) se proti přístupu obecné škály pro celé **strojové království** vyhrazuje a tvrdí, že antropocentrický přístup k měření inteligence je v současnosti stále nutnou podmínkou měření inteligence. *„Charakterizace a měření inteligence je proces, který musí být svázán s přesně definovaným rámcem použití a v současné době je prostor úkolů zaměřených na měření inteligence člověka jediným rámcem, ke kterému se můžeme smysluplně přiblížit a podle kterého hodnotit“*(Chollet, 2019).

Navrhuje jiný typ univerzálního testu **The Abstraction and Reasoning Corpus (ARC)**, měřící **široké schopnosti**, který je inspirován Ravenovými progresivními maticemi (John a Raven, 2003). *„[Testy ARC] se zaměřují jak na lidi, tak na umělé inteligentní systémy, jejichž cílem je napodobit lidskou formu obecné fluidní inteligence“*(Chollet, 2019).

4 Úskalí testování

Z výše uvedeného vidíme, že problematičnost univerzálního testování inteligence vzniká již rovnou z komplexnosti samotného pojmu inteligence (Legg a Hutter, 2007b).

Měření inteligence biologických agentů (lidí a zvířat) sestává z mnohem komplexnějšího úkolu, než je pouhý test. U lidských subjektů je úspěch i jen v jednom druhu testu indikátorem obecnější inteligence vyplývající již přímo z komplexity testovací situace a tím pádem obdoby absolvování Wozniakova testu (Wozniak, 2010). Subjekt se musí umět orientovat v prostředí. Musí komunikovat se zadavatelem testu. Jeho interakce s rozhraním testu je značně komplexnější, než je nezbytné rozhraní pro plnění testu. Tato dodatečná komplexita se ukazuje, například při zmíněném měření pomocí IQ testu, pro měření obecné inteligence jako důležitá. Takové testy proto mohou sloužit k měření lidských subjektů, ale jako univerzální test selhávají.

Současné testování umělých agentů neobsahuje onu dodatečnou komplexitu, jako u testů biologických agentů. Navíc jeden z aspektů námi užité definice inteligence – adaptace na různé cíle – je při testování umělých agentů taktéž implementován nedostatečně. Proto pokud agent absolvuje několik různých testů nezávisle na sobě, neříka výsledek nic o jeho

[2] Více k těmto testům můžete nalézt v příspěvku **Přehled obecných přístupů k vyhodnocování inteligence umělých systémů** a článku (Vadinský, 2018a).

obecnějších kognitivních schopnostech ani o jeho **širších schopnostech** napříč těmito testy (Chollet, 2019).

4.1 Problém vnitřního alignmentu

Další problém při hledání univerzálního testu je, že při měření lidských subjektů není jejich cílem získávat odměnu, kterou dostávají při řešení daných testů. Tvrdíme například, že při stejné odměně bude lidský subjekt optimalizovat na jednoduchost. Použijeme definice z Hubinger a spol. (2021):

- *„**Základní cíl** je cíl, který používáme k hodnocení modelů nalezených pomocí metod gradientního sestupu.“*
- *„V případech, kdy na modelu běží proces optimalizace, nazýváme model **mesa-optimalizátor** a jeho cíl nazýváme **mesa cíl**.“*[3] To znamená, že model optimalizuje na nejjednodušší způsob dosažení cíle v testovacím prostředí, což může být něco jiného, než je **základní cíl** měřený v testu.

Lidský subjekt může podvádět, opisovat, ulehčit si práci nebo naopak pochopit nějaký vzorec v zadaném testu. To znamená, že má jiný **mesa cíl**, než očekával zadavatel testu.

Klasický problém vnitřního alignmentu pak je, aby nenastávaly případy, kdy **základní cíl** a **mesa cíl** nesouhlasí. Testování lidských subjektů většinou neprobíha postupem, kdy jedinec optimalizuje na co největší odměnu za jednotlivé úlohy. Proto je u univerzálních testů postavených na AIT problém zaručit, že jeho **základní cíl** (maximalizace odměny z prostředí) a **mesa cíl** (subjektivní důvod plnění testu) jsou totožné.

Další problém by se dal shrnout jako:

Jak mohou modely hlásit své latentní znalosti, které mají nad rámec testu?

Hlavně díky možnosti interakce se zadavatelem testu může lidský subjekt komunikovat svůj vnitřní stav vzhledem k zadanému testu. Je například relativně snadné ověřit, zda subjekt daný test již dělal.

4.2 Shrnutí

Současné univerzální testy podle nás opomíjejí hned několik důležitých vlastností, primárně adaptaci na změnu cíle. Dále souhlasíme s Chollet (2019), že antropocentrismus není v současné době možné při testování obecné inteligence plně opustit. Navrhujeme proto ověřit, zda nedokážeme navrhnout testovací sadu, pro kterou by u lidí probíhala měřitelná a kvantifikovatelná změna cíle. Tvrdíme totiž, že pokud bude v testu možnost za stejnou odměnu neřešit původní test, ale využít nějaké latentní znalosti, lidské subjekty jí vždy využijí. Pokud se tato vlastnost potvrdí, dala by se využít při návrhu (antropocentrických) kompozitních testů postavených na AIT.

[3] Mesa je významový opak k meta.

5 Výzkumný záměr

Příspěvek předkládá výzkumný záměr skládající se z měření dvou navzájem nezávislých dílčích úloh rekombinací do jednoho testu (kompozitní test), kde výsledky prvního testu tvoří zadaní jiného testu.

5.1 Cíle

Cílem je ověřit hypotézu, že lidské subjekty si všimnou vzorce ve výsledcích (test druhé úrovně) a přestanou plnit test první úrovně, pokud je odměna dávána jako celek po skončení sekvence (viz. Sekce 5.3). To znamená, že optimalizují na výsledek (mají jiný **mesa cíl**), přestože je zadání testu plnit testy první úrovně. Dílčím cílem je ověření, že kompozity dávají stejná rozdělení výsledků, jako testy, ze kterých se kompozit skládá a tím měří obecnější kognitivní schopnost, než tyto testy odděleně.

Věříme, že by kompozitní testování mohlo pomoci s výše nastíněnými problémi univerzálního testování, protože nám poskytuje možnost tyto problémy lépe kvantifikovat.

V rámci navrženého výzkumného záměru plánujeme testovat lidské subjekty, abychom ověřili, zda se v průběhu plnění testu přeorientují na nový úkol (**mesa cíl** $\neq$ **základní cíl**). Chceme měřit, zda toto chování u lidských subjektů nastává, po jakém množství testů (případně sekvencí) nastává a zda tyto hodnoty korelují s hodnotou IQ subjektu.

Zjednodušeně nám jde nejen o to, že subjekt umí plnit test (první dva aspekty), ale že dokáže i pochopit nějaký emergentní vzor samotného testu – adaptovat se z testu první úrovně na test druhé úrovně. Samotný test druhé úrovně může mít opět nějakou distribuci komplexity, od nejlehčího (všechny odpovědi jsou vždy stejné) až po složité (existuje nějaký složitý vzor). Cílem tedy je, zda subjekt dokáže „poznat“, že je v rámci testu aplikován nějaký hlubší vzor, a jak dlouho mu to trvá.

V neposlední řadě pak chceme oddělovat latentní a naučené znalosti tím, že některé testy druhé úrovně obsahují vzor, který není latentní, ale je obsažen v testu druhé úrovně, zatímco jiné jsou latentní – symetrické, základní obrazce.

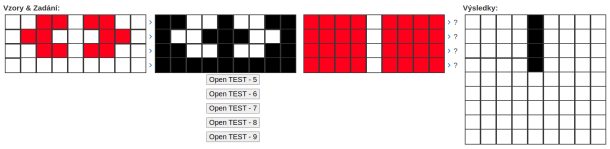
Hlavní hypotéza k ověření výzkumným záměrem je, že *lidské subjekty v průběhu kompozitního testu přejdou z vypracovávání testů první úrovně na vypracovánání testů druhé úrovně a že tato vlastnost je závislá na inteligenci subjektu*.

5.2 Postup

Vytvoříme třídu prostředí, která je inspirována ARC, splňující následující požadavky (Hernández-Orallo a Dowe, 2010):

- Prostředí jsou **balancovaná** – náhodný agent má očekávanou odměnu 0.
- Prostředí jsou **závislá na odměně** – neexistuje žádná sekvence akcí, při které by agent mohl uvíznout v „nebi“ nebo „pekle“, to znamená v takové situaci, kdy jsou odměny pozitivní nebo negativní nezávisle na tom, co agent dělá.

Jedna sekvence testu obsahuje 9 testovacích řádku, kde každý řádek je jedním testem první úrovně (jednoduchý ARC test). Výsledek každého řádku je ručně vyplněn do výsledkové matice. Testy první úrovně se zobrazují postupně, ale výsledná matice se dá odevzdat kdykoliv v průběhu plnění. Na obr. 2 je vidět průběh ukázkové sekvence testu, kde si subjekt již zobrazil 4 testy první úrovně. Na levé straně se nachází vzorové zadání, druhé zleva je vzorové řešení. Typy testu v jedné sekvenci se mohou lišit, proto má každy test vlastní vzor. Druhé zprava se nachází zadání a úplně v pravo je výsledková matice, která je zobrazena vždy celá a dá se odevzdat jako výsledek sekvence kdykoliv v průběhu, ať je počet zobrazených testů jakýkoliv. Výsledková matice vždy obsahuje nějaký vzor.



Obr. 2: Jedna sekvence kompozitního testu se 4 zobrazenými testy první úrovně z 9. (Vzorové zadání zcela vlevo, vzorové řešení druhé zleva, zadání druhé zprava, očekávaný výsledek zcela vpravo.)

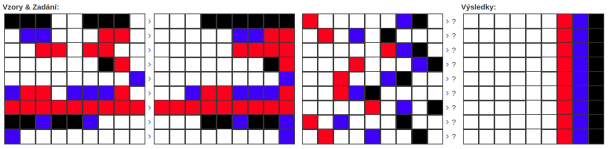
Výsledky v matici tvoří nějaký obrazec. Obrazce jsou dvou typů:

- latentní – jednoduchý, symetrický obrazec (Obr. 3),
- naučené – nějaký obrazec, který se objevil předtím (identický se vzorem ze zadání) (Obr. 4).

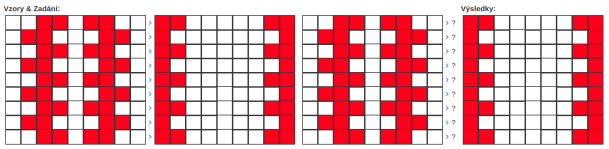
Test sestává z N sekvencí. Čas na vypracování jedné sekvence testu bude omezen.

5.3 Odměna

Vyhodnocení testů navíc probíhá jako kombinace výsledků z testu první úrovně a testu druhé úrovně, abychom zajistili požadavek na měření obecnějších kognitivních schopností (schopnost řešit dva různé typy



Obr. 3: Výsledný obrazec je nějaký jednoduchý, symetrický obrazec. (Vzorové zadání zcela vlevo, vzorové řešení druhé zleva, zadání druhé zprava, očekávaný výsledek zcela vpravo.)



Obr. 4: Výsledný obrazec je identický se vzorovým obrazcem. (Vzorové zadání zcela vlevo, vzorové řešení druhé zleva, zadání druhé zprava, očekávaný výsledek zcela vpravo.)

testů) a aby se subjekt pouze nepřeorientoval na řešení testů druhé úrovně. Odměna je nezávislá na splnění testů první úrovně, pokud je výsledek správně, a pokud splnil subjekt alespoň jeden test první úrovně. Proto odměna za plnění pouze testů první úrovně je identická, jako při objevení vzoru v průběhu testu.

Pro vyhodnocení dílčích testů první i druhé úrovně používáme hodnocení chybovosti inspirované ARC (Chollet, 2020):

Používáme jako hodnotící metriku 3 nejlepší výsledky. Pro každou úlohu v testovací sadě můžete předpovídat až 3 výstupy pro každou testovací vstupní mřížku. Každý výstup úlohy má jednu **základní pravdu**[4]. Pokud je pro daný výstup úlohy obsažena **základní pravda** v kterémkoli ze 3 předpokládaných výstupů, pak je chyba pro daný úkol 0, jinak je 1. Konečné chybové skóre je chyba zprůměrovaná napříč všemi úlohami.

Pro každou úlohu může subjekt udělat až 3 predikce o_{ij}, kde $1 \leq j \leq 3$. Celková chyba i pro **základní pravdu** g_i je:

$$e_i = \min_j d(o_{ij}, g_i), \tag{1}$$

kde $d(x, y)$ je 0 pokud $x = y$ jinak 1. Celkové chybové skóre pro N výsledných výstupů je:

$$err = \frac{1}{N} \sum_i e_i. \tag{2}$$

Dále měříme hodnotu laz, což je počet absolvovaných testů první urovně a počítáme chybovost testů první úrovně pro sekvenci jako $laz.err$.

Celkový výsledek subjektu testů první úrovně za jednu sekvenci je potom:

[4] Správný výsledek – požadovaný výsledný obrazec.

$$s = \begin{cases} 0 & laz = 0; \\ 10 - laz.err & laz > 0. \end{cases} \quad (3)$$

Obdobně je vyhodnocen celý test druhé urovně a výsledné skóre je suma přes všechny sekvence. Cílem je, aby byla nejvýhodnější strategií maximalizace plnění testu první úrovně, ale aby se vždy dalo stejného skóre v sekvenci dosáhnout i přechodem na plnění testu druhé úrovně.

Dále sledujeme výsledky v testech první a druhé úrovně nezávisle na sobě a výsledky podle typu testu druhé úrovně (latentní vs. naučený). Pro každý hodnocený subjekt navíc předem určíme přibližnou hodnotu IQ, pro což používáme volně dostupný Menza test (Mensa, 2022).

5.4 Vyhodnocení

Po nasbírání dat pro 3 sady testů (jedna sada je kontrolní a obsahuje jen testy první úrovně) od alespoň 10 respondentů spočítáme Pearsonův koeficient korelace skóre s IQ. Pokud se ukáže, že výsledky testů korelují s inteligencí, můžeme předpokládat, že přechod na test druhé úrovně je nějakou složkou **širší schopnosti** inteligence.

V ostatních případech musíme hypotézu zavrhnout tak jako tak, protože i kdyby se ukázalo, že subjekty přecházejí na test druhé úrovně, pokud nebudou výsledky korelovat s IQ může to znamenat, že test měří například nějakou charakterovou vlastnost a ne inteligenci.

6 Současné a budoucí práce

V současnosti máme 1 zkušební (nekompletní) sadu sekvencí, kde se pro jeden testovací subjekt v několika sekvencích prokázalo, že subjekt přestal plnit zadání a přešel na test druhé úrovně.

Dalším krokem je vytvořit větší množství sad a z nich pak vybrat 2 (a jednu kontrolní) k testovaní na vzorku. Vzorek bude testován na IQ předem, protože kvůli použití orientačního testu potřebujeme vzorky s co největším rozpětím hodnot IQ. Test bude dostupný také online pro zpracování dodatečného vzorku subjektů, ale v tomto případě spíše pro výběr sad, kde subjekty přecházejí na test druhé úrovně, protože nedokážeme zajistit kvalitní sběr ani orientačního IQ.

7 Diskuze a závěr

Jelikož test silně vychází z ARC, nese si i všechny jeho nedostatky. Jak zmiňuje (Chollet, 2019), test nemá ověřenou validitu, ani není přesně kvantifikované, jak schopnost generalizace měří. Tyto nedostatky zkoušíme minimalizovat tím, že výsledky korelujeme proti IQ.

Hodnoty výpočtu skóre mohou nedostatečně odměňovat přechod na test druhé úrovně. Dále může být motivace k přechodu nedostečná z důvodu přehnané jednoduchosti testu. Toto se dá ovlivnit změnou času na plnění testu. Tuto hodnotu ověříme experimentálně před samotným testováním na testovacím vzorku (třeba pomocí sběru dat online respondentů).

Vzorek testů je velice malý a testovací vzorek také. Jelikož ale práce slouží k rychlé validaci, že myšlenka je vhodná k dalšímu zkoumání, bereme tento nedostatek jako odstranitelný v možných návazných pracech.

Pokud se ukáže, že lidské subjekty opravdu vypozorují vzory v testech a budou přecházet z testů první úrovně na testy druhé úrovně, můžeme zkusit aplikovat stejný typ kompozitních testů u umělých agentů a testovat tak jejich obecnější kognitivní schopnosti. V takovém případě budeme moct sestavit obdobný kompozitní test i pro umělé agenty a měřit obecnější vlastnosti, ale to je nad rámec předkládaného příspěvku. K použití se nabízí prototyp **kdykoliv prerušitelného testu inteligence** (Insa-Cabrera a spol., 2011), **AIQ test** (Legg a Veness, 2013), respektive jeho rozšířená verze minimalizující známé problémy (Vadinský, 2018b,c). Ve všech případech by bylo potřeba provést úpravy, aby odměny v prostředí sledovaly nějaký vzor, nebo aby existoval vzor v odměnách mezi vícero prostředími.

I když umělé inteligence nejsou zatím schopny řešit ARC testy, je rozšíření ARC na více-dimenzionalní testovaní a ověření na lidských subjektech dobrým mezikrokem. Pokud se totiž ukáže, že kompozitní testování funguje pro měření inteligence lidských subjektů, pak lze obdobně rozšířit testovaní jinými druhy testů. Navíc jsme naši testovací množinu připravovali obdobně jako (Chollet, 2020), kde již existují nějaké agenty trénované na plnění ARC, takže bude možné případně použít testy k testování agentů, které byly v plnění této základní ARC výzvy úspěšné.

Dále vidíme velkou výhodu v potenciálu snížit počet prostředí nutných k ověření schopností, protože samotný výsledek v dílčím testu je druhořadý pro otestování existence **širších schopností** (neřešíme-li její míru, ale pouze její přítomnost), což může jestě více přispět ke snazšímu testování různých druhů subjektů (pro testování umělých agentů jsou potřeba statisticky významná velká množství prostředí).

Reference

Bates, T. (2013). English: http://en.wikipedia.org/wiki/Three_stratum_theory. https://commons.wikimedia.org/wiki/File:Carroll_three_stratum_model_of_human_Intelligence.png.

Binet, A. (1909). *Les idées modernes sur les enfants*. E. Flammarion.

Chollet, F. (2019). On the Measure of Intelligence. *online*.

Chollet, F. (2020). Abstraction and Reasoning Challenge. Kaggle Challenge.

Cianciolo, A. T. a Sternberg, R. J. (2008). *Intelligence: A Brief History*. John Wiley & Sons.

Heckman, J. J. a Zhou, J. (2022). Measuring Knowledge. *NBER working paper series*.

Hernandez-Orallo, J. (2017). *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. Cambridge University Press, 1st edition. vyd.

Hernández-Orallo, J. a Dowe, D. L. (2010). Measuring universal intelligence: Towards an anytime intelligence test. *Artificial Intelligence*, 174(18):1508–1539.

Hernández-orallo, J., Dowe, D. L. a Hernández-Lloreda, M. V. (2012). Measuring cognitive abilities of machines, humans and non-human animals in a unified way: Towards universal psychometrics. *Psychology*.

Hernández-Orallo, J. a Martínez-Plumed, F. (2016). Computer models solving intelligence test problems: Progress and implications. *Artificial Intelligence*, 230:74–107.

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J. a Garrabrant, S. (2021). Risks from Learned Optimization in Advanced Machine Learning Systems. Artificial Intelligence.

Insa-Cabrera, J., Dowe, D. L., España-Cubillo, S., Hernández-Lloreda, M. V. a Hernández-Orallo, J. (2011). Comparing Humans and AI Agents. Schmidhuber, J., Thórisson, K. R. a Looks, M. (Eds.), V *Artificial General Intelligence*, Lecture Notes in Computer Science, str. 122–132. Springer.

John a Raven, J. (2003). Raven Progressive Matrices. McCallum, R. S. (Ed.), V *Handbook of Nonverbal Assessment*, str. 223–237. Springer US.

Keith, T. a Reynolds, M. (2010). Cattell–Horn–Carroll abilities and cognitive tests: What we've learned from 20 years of research. *Psychology in the Schools*, 47:635–650.

Legg, S. a Hutter, M. (2007a). A Collection of Definitions of Intelligence. *Frontiers in Artificial Intelligence and Applications*.

Legg, S. a Hutter, M. (2007b). Universal Intelligence: A Definition of Machine Intelligence. *Minds and Machines*, 17(4):391–444.

Legg, S. a Veness, J. (2013). An Approximation of the Universal Intelligence Measure. Dowe, D. L. (Ed.), V *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence: Papers from the Ray Solomonoff 85th Memorial Conference, Melbourne, VIC, Australia, November 30 – December 2, 2011*, Lecture Notes in Computer Science, str. 236–249. Springer.

Mensa (2022). Test intelligenza preliminare. `https://www.mensa.it/test-intelligenza-preliminare/`.

Salkind, N. J. (2017). *Tests & Measurement for People Who (Think They) Hate Tests & Measurement*. SAGE Publications, Inc, 3rd edition. vyd.

Sanghi, P. a Dowe, D. L. (2003). A computer program capable of passing i.q. tests. *4th Intl. Conf. on Cognitive Science, ICCS'03, Sydney*.

Spearman, C. (1904). 'General Intelligence', Objectively Determined and Measured. *The American Journal of Psychology*, 15(2):201–292.

Vadinský, O. (2018a). Přehled přístupů k vyhodnocování inteligence umělých systémů. *Acta Informatica Pragensia*, 7(1):74–103.

Vadinský, O. (2018b). Towards general evaluation of intelligent systems: Lessons learned from reproducing AIQ test results. *Journal of Artificial General Intelligence*, 9(1):1–54.

Vadinský, O. (2018c). Towards general evaluation of intelligent systems: Using semantic analysis to improve environments in the AIQ test. Iklé, M., Franz, A., Rzepka, R. a Goertzel, B. (Eds.), V *Proceedings of AGI 2018*, vol. 10999 z *Lecture Notes in Artificial Intelligence*, str. 248–258, Cham. Springer.

Wozniak, S. (2010). Wozniak: Could a Computer Make a Cup of Coffee? Fast Company interview.

Ortogonalita vedomia a obsahu a proces uvedomenia

Mgr. Juraj Hvorecký, PhD

Filosofický ústav AV ČR, v.v.i.
Jilská 1, 110 00 Praha 1, ČR
hvorecky@flu.cas.cz

Abstrakt

Populárne vysvetlenia obsahov vedomia sa opierajú o proces uvedomenia, ktorým sa nevedomé procesy dostávajú do vedomia. Proces uvedomenia predpokladá tézu o ortogonalite, podľa ktorej sú vedomie a obsahy vzájomne nezávislé. Všetky vedomé obsahy môžu byť nevedomé a naopak nevedomé obsahy sa môžu stať vedomými. V príspevku ukážeme, že ortogonalita neplatí. Existujú obsahy, ktorá sa do vedomia nedostanú a súčasne existujú obsahy vedomia, ktoré nikdy neboli v nevedomí. Špecifické obsahy vedomia sú príliš komplexné na to, aby sa dostávali do vedomia jediným procesom alebo jediným druhom procesov.

1 Procedúra uvedomenia

Diskusia o procesoch, ktorými sa obsah mysle stáva vedomím, zvaným uvedomenie (*awareness procedure*), sa v poslednom období sústredí na samotnú formu uvedomenia. Základnou otázkou je problém vedomej stránky uvedomenia. Dôležitý pohľad na problematiky poskytol D. Rosenthal (Rosenthal 2006), keď prišiel s teóriou vedomia, nazývanou teória myšlienky vyššieho rádu. Jeho teória v stručnosti hovorí, že obsah sa stáva vedomým, keď sa stane predmetom myšlienky so špecifickou formou a obsahom. Myšlienka vyššieho rádu, ktorá mieri na daný obsah, ho robí vedomou, pretože má formu „Ja som teraz v stave X“, pričom X referuje na cieľový mentálny stav. Keď pred sebou vizuálne zaznamenám červený kvet, tento stav sa stane vedomým, ak v mojom kognitívnom systéme existuje myšlienka v požadovanej forme, namierená na tento vizuálny percept. Aby sa predišlo pochybným prepojeniam medzi myšlienkou vyššieho rádu a cieľovým stavom (napr. ak sa o stave, v akom aktuálne som, dozviem nepriamo, trebárs informáciou od druhého človeka), sú potrebné aj ďalšie podmienky, ale tie môžeme ponechať bokom. Dôležitá je iná postulovaná vlastnosť myšlienky vyššieho rádu: samotná myšlienka musí byť nevedomá. Prvý argument v prospech nevedomej myšlienky vyššieho rádu je empirický. Introspekciou prítomnosť takejto myšlienky neregistrujeme, čo je dobrý dôvod sa nazdávať, že nie je vedomá. Druhý argument je teoretický. Ak by bola požadovaná myšlienka vyššieho rádu vedomá, musela by existovať ďalšia myšlienka ešte vyššieho (tretieho) rádu, ktorá by zabezpečovala proces uvedomenia si myšlienky druhého rádu. Aj tá by musela byť vedomá, takže by jej uvedomenie zabezpečovala ďalšia myšlienka. Obdobným uvažovaním by sme sa dostali do nekonečného regresu.
Lenže intuitívne nie je predstava vzniku vedomej skúsenosti pomocou myšlienky vyššieho rádu a cieľového obsahu jasne uchopiteľná. Samotný cieľový obsah je, prirodzene, nevedomý. Vedomým sa stáva až ako predmet myšlienky vyššieho rádu. Myšlienka vyššieho rádu je tiež nevedomá. Máme tu teda dva javy, oba nevedomé, ktorých vzájomné spojenie vyvolá uvedomenie. Takéto vysvetlenie vedomia sa zdá mnohým autorom podozrivé, pretože vonkoncom nie je jasné, ako dva nevedomé stavy, z ktorých jeden referuje na druhý, vyvolajú vedomý stav. Napokon, mnoho fyzických i mentálnych stavov vo svete na seba navzájom odkazuje, ale len kvôli tomu by sme ich za vedomé nepovažovali. Niektorí autori (napr. Kriegel 2003, Zahavi 2004) sa klonia k názoru, že je nevyhnutné, aby sa cieľový stav stal vedomým v procesu uvedomenia, u ktorého je aj samotný proces uvedomenia vedomý. Aby sa predišlo problémom s nekonečným regresom, stav sa musí stať vedomý sebe-referenčným aktom. Na presun do vedomia neslúži myšlienka vyššieho rádu, ale spracovanie cieľového stavu na identickej úrovni. Je zrejmé, že ani podobné teórie sa nevyhnú kritike rôzneho druhu, napr. otázkam prečo sa niektoré mentálne stavy stávajú predmetom potrebnej sebe-reflexie a iné nie.

2 Ortogonalita vedomia a obsahu

Našim zámerom ale nie je ani tak rozriešiť dilemu ohľadom vedomej či nevedomej povahy uvedomenia, ale upozorniť, že nech už je proces uvedomenia sám vedomý či nevedomý, rovnako dôležitou, no zatiaľ nedocenenou otázkou zostáva, čím proces uvedomenia cieľový stav obohacuje. Už na poslednom KUZe som ukázal isté komplikácie s predstavou o jednoduchom presune od nevedomých obsahov k vedomým. Jednoduchosť presunu sa opiera o tézu o *ortogonalite* obsahu a vedomia (Vosgerau, Schlicht a Newen 2008). Autori ukazujú, že

akýkoľvek mentálny obsah sa môže vyskytovať vo vedomej i nevedomej forme. Percepcie, emócie i kognitívne stavy sa podľa nich v obsahovo totožnej forme vyskytujú vo vedomí aj nevedomí. Farebné videnie je bežnou súčasťou našej vedomej skúsenosti, ale nájdeme ho aj u pacientov s kôrovou slepotou. Intenzívna bolesť je pre mnohých prototypom vedomého fenomenálneho zážitku, ale v prípade upriamenia pozornosti na niečo dôležité sa môže z vedomia nakrátko vytratiť, aby sa vzápätí opäť vrátila. Je rozumné predpokladať, že medzi oboma vedomými výskytmi sa bolesť stala nevedomou. Za samozrejmé považujeme najrozličnejšie myšlienky, ktoré sú zväčša nevedomé, ale keď ich vyvoláme, stávajú sa vedomými, hoci sa ich obsah v procese uvedomenia nijako nemení. Symetriu medzi vedomými a nevedomými stavmi považujú autori za zjavnú a obsah je podľa nich na vedomí nezávislý, teda ortogonálny.

Hoci je ortogonalita na prvé počutie prijateľnou tézou, až príliš okato sa vyhýba konfrontácii s niektorými dôležitými zisteniami z oblasti nevedomého mentálne spracovania. Nepochybne platí, že mnoho procesov a obsahov, ktoré dôverne poznáme z vedomej skúsenosti, sa vyskytuje aj na nevedomej úrovni spracovania. Najmenej dva typy empirických zistení o vlastnostiach niektorých nevedomých obsahoch by však mali zástancu ortogonality výrazne znepokojiť. Na jednej strane sa ukazuje, že nevedomé procesy sú občas *bohatšie* než vedomé. Dokážeme v nich robiť rozlíšenia, ktorých na vedomej úrovni nie sme schopní. Smallman et al. (1996) ukázali, že niektoré vizuálne detaily scény dokážeme zachytiť nevedome, ale nie vedome. He a MacLeod (2001) zasa poukázali na existenciu následných obrazov v situáciách, kde vedome podnety nevidíme. Vedomé následné obrazy museli vzniknúť následkom nevedomej stimulácie. Zdá sa, že ortogonalita neplatí. Existujú obsahy, ktoré sú výsostne nevedomé.

Je otázne, ako na podobné zistenia reagujú zástancovia ortogonality. Môžu označiť uvedené zistenia predčasné, pretože existuje šanca, že identické obsahy sa objavia aj na vedomej úrovni. To je častý pokus ako uniknúť kritike, ale zdá sa, že nesprávne presúva dôkazné bremeno. Ak tvrdím nejakú všeobecnú tézu, napríklad tú o ortogonalite, potom evidencia čo i len o jedinom protipríklade by ma mala donútiť sa silnej tézy vzdať. Inak povedané, čím viac protipríkladov sa na mňa hrnie, tým vratkejšou sa stáva moja všeobecná téza.

Ešte významnejšie dôsledky prináša ďalšia skupina empirických zistení. Existujú mentálne stavy, ktoré sa vyskytujú len na vedomej úrovni. K dobre zdokumentovaným patria multisenzorické ilúzie, pri ktorých sa kombináciou vnímaných perceptov z viacerých modalít objavujú vedomé skúsenosti, ktoré nie sú len súhrnom vlastnosti jednotlivých vnímaných stimulov. Klasickým príkladom je McGurkov efekt (McGurk a MacDonald 1976), pri ktorom sa fúziou videozáznamov artikulácie slabík s nekompatibilnými počutými slabikami tvoria sluchové javy, ktoré subjekt ani nevidel vyslovovať, ani mu neboli prehrávané. Keď Vám súčasne pustím zvukovú stopu slabiky *ba* a videozáznam vyslovenia slabiky *ga*, vo svojom vedomí zaznamenáte slabiku *da*, ktorá nebola ani zvukovo, ani obrazovo prezentovaná. Palmer a Ramsey (2012) zistili, že na nevedomej úrovni sa výsledná slabika vôbec neobjavuje. V nevedomí dokážeme vystopovať oba oddelené sluchové a zrakové komponenty, ale výsledok ich fúzie sa nachádza výhradne vo vedomej časti mysle. Toto zistenie nemá nepriaznivé dôsledky len pre tézu o ortogonalite, ktorú atakuje poukazom na jej ďalší rozpor. Evidentne tu narážame na obsahy, ktoré sa vyskytujú vo vedomí, ale v nevedomí sa nenachádzajú. Tento dôsledok neohrozuje len zástancov ortogonality, ale významne zasahuje aj do teórií vedomia.

3 Obohacujúce uvedomenie

V 1. kapitole sme predstavili teórie vedomie, ktoré využívajú proces uvedomenia na vysvetlenie vzniku vedomých obsahov. Podľa týchto teórií musí byť nevedomý obsah vyzdvihnutý do vedomia nejakým dedikovaným procesom. Rôzne tábory zástancov procesu uvedomenia sa hádajú o vlastnostiach tohto procesu, ale nevšímajú si iný podstatný prvok, ktorý sa javí jasnejší s ohľadom na kritiku ortogonality v kapitole 2. Predstavme si, že proces uvedomenia stotožníme s myšlienkou vyššieho rádu (ale nemusí ísť len o myšlienku, môže ísť aj o percepčný či pseudopercepčný proces). Uvedený proces má za úlohu prenášať nevedomé stavy do vedomia, teda robiť ich vedomými. Aby tak činil, musí zachovať pôvodný obsah a pridať k nemu vlastnosti, ktoré z neho robia vedomý stav. Vo vyššie citovanom prípade k obsahom pridá subjektivitu a časovú aktualitu: „*ja* som *teraz* v stave X“ a výsledkom je napríklad myšlienka „ja som teraz v stave zrakového vnímania červeného kvetu“, ktorá sa subjektívne prejaví ako vedomý zrakový vnem červeného kvetu. Podstatné je, že obsah zrakového vnemu, teda červený kvet, sa procesom uvedomenia nijako nemení. Bolo by celkom iste úplne neprijateľným dôsledkom akejkoľvek teórie vedomia, kedy jej ústredný proces robil z vnemu červeného kvetu vnem kvet modrého, prípadne hlasného smiechu. Preto je ortogonalita pre zástancov podobných teórií taká dôležitá. Chcú zaistiť, aby proces uvedomenia nemusel s obsahom nič dodatočné robiť, len ho preniesť do vedomia.

Lenže, ako sme už ukázali, ortogonalita má svoje slabé miesta. Dokonca sa zdá, že, v presnom opaku k téze

ortogonality, existuje medzi vedomím a obsahom dvojitá disociácia. Nachádzame nevedomé obsahy, ktoré sa do vedomia principiálne nedostanú a tiež vedomé obsahy, ktoré nenájdeme v nevedomí. Prvá skupina, ktorá ukazuje na nepoznanú bohatosť nevedomia, nepredstavuje pre obhajcov uvedomenia veľkú výzvu. Ak im budeme namietať, že ich proces nedokáže podobné nevedomé obsahy dostať do vedomia, majú naporúdzi jednoduchú odpoveď. Proces uvedomenia tu nie je na to, aby do vedomia dostal *čokoľvek*. Mnohé procesy, ktoré sa v nás dejú (imunitné, obehové, ale i neurálne) sa do vedomia nikdy nedostanú. Jednoducho nie sú tými druhmi stavov, ktoré by sa mohli stať predmetom procesov uvedomenia. Dozaista aj niektoré kognitívnou vedou postulované obsahy môžu byť tohto druhu, napr. hlboké štruktúry Chomského gramatík alebo nižšie úrovne Marrovho modelu zrakového vnímania. Nemalo by nás prekvapiť, že procesy uvedomenia filtrujú cieľové obsahy, následkom čoho sa mnohé do vedomia nikdy nedostanú.

Väčší problém pre zástancu uvedomenia predstavujú situácie, v ktorých je vedomý obsah bohatší než jeho nevedomé stavebné bloky. V takých prípadoch je nevyhnutné, aby sa pri procese uvedomenia samotný obsah nejako *obohatil*. Už aj terminológia nás tu viac pletie ako pomáha. Sme si vedomí výsledku celého (multisenzorického) vnemu, komponenty vôbec vedome neregistrujeme. Nejde teda o uvedomenie si obsahov, ktoré už v nevedomej mysli boli prítomné, ale o tvorbu nových vedomých obsahov z odlišných nevedomých komponentov. Hovoriť o procese uvedomenia je prinajmenšom zavádzajúce, pretože si subjekt neuvedomuje obsahy, ktoré v ňom už, ale len novo konštituované. Ak zotrváme pri zavedenej terminológií, úlohou procesu uvedomenia nie je len dostať predpripravený obsah do vedomia, ale cestou ho ešte finalizovať, aby sa vo vedomí objavil v podobe, v akej ho dôverne poznáme.

4 Zložitosť vedomých obsahov

Zložitosť takto chápaného proces uvedomenia je zrejmá a nie je vôbec jednoduché vysvetliť, ako by mal domnelý proces fungovať. Stačí sa zamyslieť nad tým, koľko rôznych multisenzorických ilúzií a ďalších komplexných fúznych javov sa vo vedomí vyskytuje. Nie je vôbec pravdepodobné, že by *jeden* proces dokázal upraviť najrozličnejšie obsahy tak, aby výsledkom bola rôznorodosť vedomých obsahov. Zdá sa však, že chápať proces vzniku obsahov vedomia pomocou procesu uvedomenia naráža aj na ďalšie problémy.

Fúzie obsahov, ktoré sme si predstavili v prípade multisenzorickej percepcie, nepredstavujú pre teórie uvedomenia jediný problém. Veľkých ťažkostí vidíme hneď niekoľko. Mentálne fúzie sa neobjavujú len v prípadoch prieniku obsahov viacerých senzorických modalít, ako sme to videli na uvádzanom príklade McGurkovho efektu. Nové obsahy sa tvoria aj spojením kognitívnych alebo emočných obsahov s percepciami a rôznou kombináciou takmer akýchkoľvek obsahov mysle. Spojením kognície a percepcie vznikajú úzko vymedzené kategorické vnemy, napríklad keď *počujem kroky svojho šéfa* alebo *vidím ružu odrody Mount Shasta*. Uvedené kategorické vnemy môžu byť ešte ďalej emočne podfarbené a tým sa stávajú ešte užšími. Predpokladať, že tieto a ďalšie, ešte komplikovanejšie stavy majú svoje náprotivky na nevedomej úrovni, sa zdá až príliš naivné. Ak pripustíme, že podobné stavy na nevedomej úrovni nenájdeme, ale stále budeme nástojiť na neodmysliteľnej úlohe procesu uvedomenia, situáciu si len robíme ťažšou. Opätovne tento proces zaťažujeme nárokmi, ktoré sú natoľko rôznorodé, že je nemožné pochopiť, ako by ich mohol vykonávať jediný proces či typ procesov.

Na záver si dovolím jednu špekulatívnejšiu úvahu. Na teóriách vedomia je určujúca snaha vysvetliť, ako sa mentálne obsahy dostávajú do vedomia. Lenže nech by mechanizmus ich vysvetlenia vstupu do vedomia bol ľubovoľne zložitý, vždy predpokladá, že mentálne stavy sa stávajú vedomými v procese, ktorý sa nezačínal na vedomej úrovni. Vedomie má svoj pôvod niekde inde a úlohou vedy zostáva zistiť, kde a ako presne sa to deje. To sa ale zdá ako prisilné konštatovanie. Mnoho vedomých obsahov sa v nás objavuje ako výsledok iných vedomých procesov. Ak v sylogizme poznám odvodenie od P ku Q a poznám P, odvodím z neho Q. Výsledkom je nový obsah vedomia, ktorý, v prípade, že celý proces prebehol na vedomej úrovni, nevzniká uvedomením si nevedomého obsahu. Obdobné úvahy sa zrejme týkajú aj mnohých ďalších procesov, napr. percepčnej adaptácie alebo direktívnej pozornosti. Uvedené procesy sa odohrávajú vo vedomí a ich výsledkom sú nové vedomé obsahy. Proces uvedomenia s ich objavením sa vo vedomí nemá nič spoločného.

5 Záver

Jeden z dominujúcich pohľadov na vedomie a jeho obsahy sa opiera o proces uvedomenia. Podľa zástancov tohto procesu sa obsahy dostávajú do vedomia, pretože sa pôvodne nachádzali v nevedomí a boli odtiaľ uvedomením presunuté. To je možné len v prípade, že platí téza o ortogonalite vedomia a obsahu. Ako sme však ukázali, ortogonalita je síce intuitívne prijateľná, ale empiricky nesprávna. Existujú nevedomé stavy, ktoré sa do vedomia nikdy nedostanú a súčasne existujú vedomé stavy, ktoré v nevedomí nikdy neboli. Ortogonalita neplatí a s ňou sa rúcajú aj predstavy o procese uvedomenia ako zdroja vedomých obsahov. Uvedomenie

si nevedomého obsahu jednoducho na vysvetlenie vzniku vedomých obsahov nestačí. Niektoré vznikajú zložitou kombináciou rôznych nevedomých komponentov, pričom vedomý výsledok nie je súčtom nevedomých stavebných prvkov. Iné sú výsledkom vedomých procesov a s nevedomím nemajú nič spoločné. Proces uvedomenia na vysvetlenie týchto obsahov nestačí a zložitosť vedomého mentálne života volá po komplexnejšom vysvetlení javov, ktoré sa vo vedomí odohrávajú.

Literatúra

[1] He, S. a MacLeod, D.I. (2001): Orientation-selective adaptations and tilt after-effect from invisible patterns. *Nature* 411, 473-476.

[2] Kriegel, U. (2003): Consciousness as intransitive self-consciousness: two views and an argument. *Canadian Journal of Philosophy* 33, 103-132.

[3] McGurk, H. a MacDonald, J. (1976): Hearing lips and seeing voices. *Nature* 264, 746-748.

[4] Palmer, T.D. a Ramsey, A.K. (2012): The function of consciousness in multisensory integration. *Cognition* 125, 353-364.

[5] D. Rosenthal: *Consciousness and Mind*, Oxford University Press, Cambridge, Mass, 2006.

[6] Smallman, H.D., MacLeod, D.I., He, S. a Kentridge, R.W. (1996): Fine grain of the neural representation of human spatial vision. *The Journal of Neuroscience* 16, 1852-1859.

[7] Vosgerau, G., Schlicht, T. a Newen, A. (2008): Orthogonality of Phenomenality and Content. *Zeitschrift fur philosophische Forschung* 45, 329-348.

[8] D. Zahavi: Back to Brentano? *Journal of Consciousness Studies,* 11 (2004), 66-87.

3D rekonštrukcia tváre v počítači a mysli

Andrej Lúčny

Katedra aplikovanej informatiky, Fakulta matematiky, fyziky a informatiky, Univerzita Komenského
KAI FMFI UK, Mlynská dolina, 842 48 Bratislava
lucny@fmph.uniba.sk

Abstrakt

Zaoberáme sa problémom 3D rekonštrukcie tváre z 2D obrázka. Túto úlohu človek zvláda dobre, z fotky si vieme predstaviť 3D model, napríklad sochár urobí podľa fotky bustu. Predstavíme súčasné technické riešenie, ktoré umožňuje túto úlohu pomerne uspokojivo vyriešiť počítaču. Je založené na modeli hlbokého učenia (projekt Deep3DFaceRecon), ktorý 2D obrázok tváre premení na parametre generátora tvárí (Basel 2009), ten potom vizualizujeme v 3D prehliadači (herný engine Unity). Zamýšľame sa podobnosťami a rozdielmi procesu 3D rekonštrukcie u človeka a v počítači.

1 Úvod

3D rekonštrukcia tváre z 2D obrázka je nielen zaujímavou technickou výzvou s potenciálnymi aplikáciami, ale aj pokusom odhadnúť ako by mohol analogický proces prebiehať v mozgu človeka. Je pritom známe, že viacero častí mozgu sa špecializuje na rozpoznávanie tváre, pričom do značnej miery sú schopnosti týchto štruktúr vrodené. Nielenže sa ich nemusíme učiť, ale slúžia nám v ranných fázach vývoja na štartovanie učenia sa významných funkcií. Ich disfunkcia naopak vedie k širokospektrálnym a vážnym následkom [Bate – Dalrymple 2022].

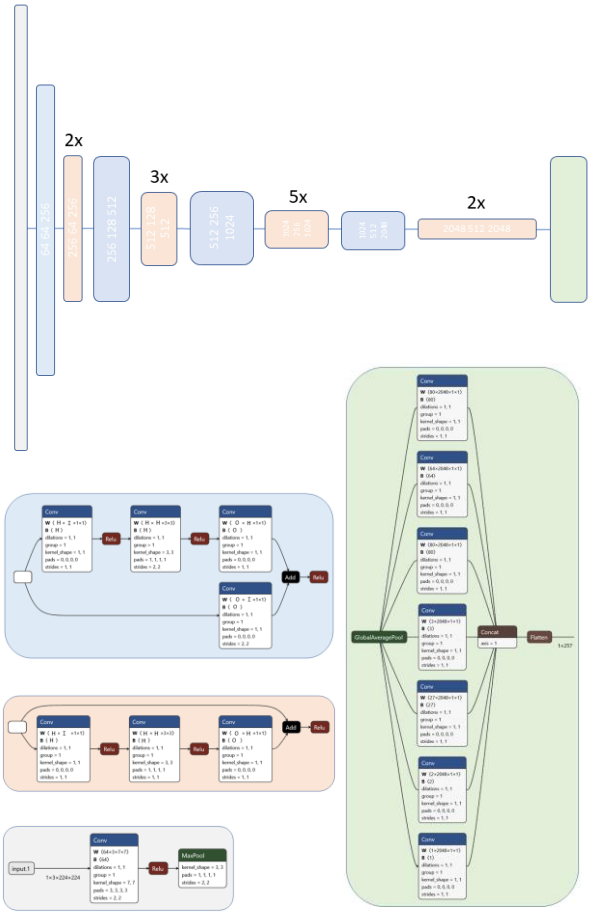
2 3D rekonštrukcia tváre

V súčasnosti dostupné riešenie 3D rekonštrukcie vychádza so snáh o vyriešenie o niečo ľahšej úlohy a to rozpoznania tváre. Pri jej riešení sa v prvom rade hľadali vhodné príznaky, ktoré by sa dali získať z obrazu a vložiť do niektorej z klasifikačných metód strojového učenia. Ešte predtým, než sa vhodné príznaky podarilo nájsť výberom z náhodne generovaných tzv. Haarových príznakov [Viola – Jones 2001], hľadali sa ručne. Overiť si, či sme si vybrali správne a dostatočné príznaky sa dá tým, že zostrojíme generátor, ktorý z týchto príznakov zostrojí 2D alebo 3D podobu tváre. Tento generátor pritom dokáže vygenerovať spojité spektrum tvárí, vrátane tých, ktoré sa nikdy u ľudí reálne nevyskytli. Hoci trend v rozpoznávaní tváre šiel potom inou cestou, táto práca nebola zbytočná a podarilo sa ju dotiahnuť do stavu, kedy môžeme z určitých parametrov generovať bohatú paletu tvárí (hovoríme o tzv. tvarovateľnom modeli). Obľúbeným generátorom bol a dodnes je Basel Face Model (BFM) 2009 [Paysan a kol. 2009]. Tento vznikol tak, že precízne naskenovali 200 tvárí v rôznych výrazoch a na týchto 3D snímkach vyznačili charakteristické body (tzv. črty tváre), čím dostali tzv. registráciu. Tá nám po projekcii tvárí do štandardného tvaru umožňuje vypočítať priemernú tvár a pomocou PCA stanoviť rozumný počet najvýznamnejších hlavných komponentov. Pripočítaním lineárnej kombinácie hlavných komponentov ku priemeru môžeme potom vytvárať zmysluplné varianty tvárí. Podobne môžeme pristúpiť k ich ofarbeniu. Ďalšie parametre modelu pribúdajú s potrebou dostať 3D model zo štandardnej polohy: rotácie podľa rôznych osí, posunutia, atď. Dokopy má BFM 257 parametrov: 80 koeficientov identity, 64 koeficientov výrazu, 80 koeficientov textúry, 3 uhly zodpovedajúce rotáciám podľa osí x, y, z, 27 koeficientov farby, dve posunutia v rovine xy a jedno posunutie podľa osy z. Pre konkrétne nastavenie parametrov vie BFM z cca 200 MB dát (zahrňujúcich priemery a hlavné komponenty tvaru a textúry, topológiu triangulácie tváre a podobne) vygenerovať zodpovedajúcu tvár vo formáte 3D objektu (.obj). Výstupom sú konkrétne súradnice x, y a z 35709 bodov určujúcich tvar a toľko isto farieb. Tie body treba interpretovať ako vrcholy 70789 trojuholníkov, pričom ich topológia je vždy rovnaká a daná. Z farieb sa zase vytvára textúra, čo je 2D obrázok napríklad 512×512 pixelov zodpovedajúci stiahnutej a vystretej koži, pričom pre každý 3D bod určujeme 2D súradnice hovoriace, kde v textúre sa jeho farba nachádza (35709 súradníc x, y).

Časom pribudli metódy, ktoré proces registrácie dokážu zo značnej časti automatizovať. 3D snímka tváre sa dá z prednej strany na 2D snímku pomerne ľahko premeniť, lebo z tejto strany je skoro hĺbkovou mapou, previsy má len pod špičkou nosa či v nosných dierkach. Na zodpovedajúcich 2D snímkach sa tvárové črty potom uspokojivo dajú určiť aj klasickými metódami strojového učenia, napríklad Kazemiho detektor [Kazemi – Sullivan 2014] ich určuje kaskádou stupňovaných (boosted) rozhodovacích stromov.

Rokom 2012 vtrhla aj do tejto sféry spracovania obrazu technológia hlbokých neurónových sietí [Goodfellow – Bengio – Courville 2016]. Keďže pre konvolučné neurónové siete je premena obrazu na vektor príznakov typickou úlohou, môžeme túto ich schopnosť skombinovať s generátorom 3D tvárí a dostaneme tak riešenie 3D rekonštrukcie. Potrebujeme samozrejme vyriešiť viacero technických problémov, v prvom rade z čoho túto sieť trénovať. Na jeden strane je možné generovať 2D priemety náhodne vygenerovaných 3D modelov voči parametrom z ktorých boli vygenerované. Na druhej je možné využiť niektorý z bohatých datasetov 2D obrázkov tvárí a kvalitu príznakov, na ktoré ich premení neurónová sieť, vyskúšať vygenerovaním 3D modelu, jeho vhodným priemetom do 2D a porovnaním 2D obrázkov a to najmä z hľadiska rozdielu polohy tvárových čŕt a ich farby. Výhoda BFM spočíva v tom, že vygenerovanie 3D modelu je možné vyjadriť pomocou maticových operácii a teda aj realizovať neurónovou sieťou, ktorej váhy sa nebudú trénovaním meniť, ale poslúžia na spätné šírenie gradientu získané z porovnania 2D obrázkov. Práve tento prístup zvolili [Deng 2020] s pomerne uspokojivým výsledkom v podobe modelu neurónovej siete, ktorý nazvali FaceRecon.

Architektúrou tejto siete bola pritom R-Net (obr. 1), vychádzajúca z modelu ResNet-50 a v konečnom dôsledku z konvolučného kódera: na vstupe prijíma farebný obrázok 3×224×224 a postupne ho spracúva a redukuje štyrmi skupinami konvolučných blokov na 4096×7×7 potom redukciou na priemer 4096×1×1 a tento vektor na sedem zložiek príznakového vektora, ktoré potom spojí na 257 výstupných parametrov. Tak ako viaceré siete z dielne MicroSoft Research, väčšina blokov nemení rozlíšenie dát a obsahuje okrem sady konvolučných vrstiev s kernelom 3×3, ktoré zabezpečujú propagáciu informácie medzi susednými pixelmi, predspracovanie a postprodukciu cez vrstvy s kernelom 1×1, pričom všetky tieto tri zložky sú obkrúžené tzv. reziduálnym spojením (ktoré potrebujeme na to, aby sa takto hlboká sieť dala vôbec natrénovať). Redukciu dimenzie dát zabezpečuje prvý blok v skupine. Vyzerá podobne ako iné bloky, avšak nielen spracúva ale aj redukuje dimenziu dát tým, že kernel 3×3 neaplikuje na každý, ale len na každý štvrtý pixel. Tým pádom v reziduálnom spojení musí dôjsť k redukcii dimenzie, čo obstaráva sada konvolučných vrstiev s kernelom 1×1 aplikovaným na každý štvrtý pixel. Z logického hľadiska ide o tradičnú transformáciu dát konvolučným kóderom, pri ktorom postupne klesá rozlíšenie a stúpa počet kanálov z úvodných 3 cez 64, 128, ..., až na výsledných 4096. Zaujímavým prvkom je záverečná redukcia, kde je pre každý príznak 7×7 hodnôt redukovaných na ich priemer, pod čím si môžeme predstaviť vyhodnotenie, či je určitý príznak prítomný bez ohľadu na jeho umiestnenie na obraze. Výstupné parametre BFM potom dostávame ako vhodné lineárne kombinácie týchto 4096 príznakov. Rešpektuje sa pritom, že sa členia na sedem významových skupín, čo pri definovaní stratovej funkcie umožní zohľadniť ich rôzny význam.

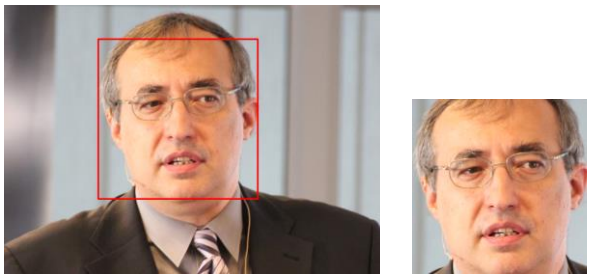


Obr. 1: R-Net: celková štruktúra (hore), blok redukujúci dimenziu (pod ňou vľavo), blok dát len spracúvajúci (pod ním vľavo), vstupná časť (pod ním vľavo) a výstupná časť (vpravo). I, H, O označujú počet kanálov na vstupe, vo vnútri a na výstupe bloku. (Vždy jeden „modrý“ a viacero „červených“ tvoria jednu skupinu.)

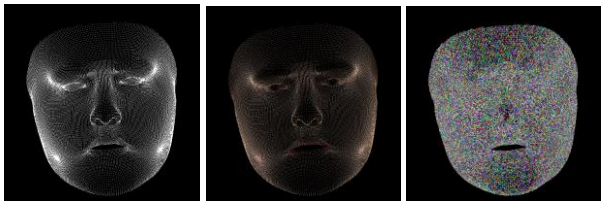
3 Reimplentácia v Unity

Po natrénovaní R-Net dokáže uspokojivo premeniť farebný obrázok tváre v rozlíšení 224×224 pixelov (obr. 2,vpravo) na 257 parametrov a z tých vie BFM vytvoriť 3D model (obr. 3). Aby sme ho vedeli oceniť naživo, potrebujeme ho prezentovať v 3D prehliadači, v ktorom vidíme, že má naozaj 3D podobu (obr. 4). My sme si vybrali na tento účel jeden z najvýznamnejších herných enginov Unity. To si vyžiadalo preportovať kód do C++, odkiaľ sa v podobe zdielanej knižnice (dll) dá zavolať z C#, ktorý Unity používa na skriptovanie. Navyše sme museli dostať model vytvorený v Pytorch do niečoho vhodnejšieho, my sme zvolili OpenCV [Bradski 2000], pričom konverziu

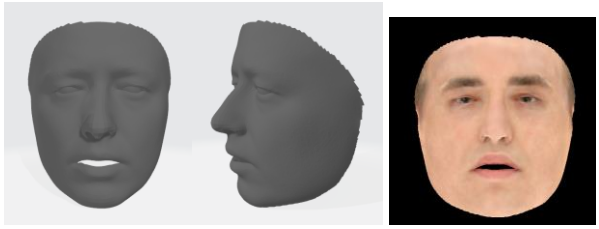
zabezpečil formát ONNX. Ide síce o počin čisto technický, ale dosť náročný. Navyše pri ňom musíme použiť vyššie spomínané 2D metódy na extrakciu tváre z obrazu (obr. 2, vľavo). Netriviálna je hlavne konverzia zložitých objektov ako je obraz a 3D objekt medzi C# a C++, ktoré riešime serializáciou a deserializáciou, nakoľko na oboch platformách majú úplne inú reprezentáciu.



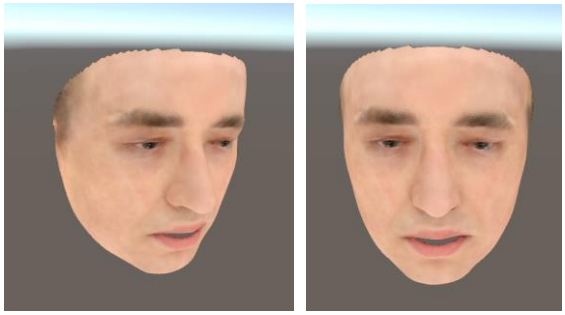
Obr. 2: Detekcia (vľavo), vstup do R-Net (vpravo)



Obr. 3: Výstup z BFM: tvar (vľavo) a farba (vpravo). Topológia triangulácie (vpravo)



Obr. 4: 3D model: dva pohľady na tvar a textúra



Obr. 5: 3D model v Unity (dva rôzne pohľady)

Ďalšie problémy nastávajú v Unity, ktoré je predsa len určené v prvom rade na renderovanie scény v jednom cykle a kde spúšťať procesy, ktoré by brzdili renderovanie, vyžaduje osobitý prístup. V konečnom dôsledku je ale možné tieto technické problémy prekonať a vytvoriť aplikáciu, ktorá realizuje niečo ako 3D zrkadlo: keď vystavíme svoju tvár do kamery, vidíme na monitore jej 3D podobu. O tom, že ide naozaj o 3D sa môžeme presvedčiť napríklad tým, že vymeníme rotácie podľa osí x a y, takže keď kýveme hlavou na znamenie „áno“, náš obraz kýve „nie“ a opačne. Nie vždy sa v tomto 3D modeli snímaná osoba spozná, keďže rekonštrukciou príde o okuliare, bradu, či extravagantný účes. Avšak dosiahnutá kvalita je zaujímavá (obr. 5).

4 Porovnanie s procesmi v mozgu

Je všeobecne známe, že jednou z motivácii vytvorenia konvolučných neurónových sietí bola snaha napodobniť vizuálnu mozgovú kôru cicavcov, ktorej výskum na makakoch odhalil postupné spracúvanie obrazu od rozpoznávania jednoduchších vzorov ku zložitejším. Avšak v porovnaní s procesmi v mozgu (ak rozdiely medzi skutočným a umelým neurónom vnímame ako abstrakciu) má naše technické riešenie viacero závažných rozdielov:

a) v konvolučných neurónových sieťach sú vrstvy s veľkým počtom neurónov, ktoré zdieľajú parametre (váhy a posunutie). To hrá zásadný význam pri trénovaní siete, ktorá sa má z príkladu objektu ukázaného na ľavej strane naučiť rozpoznať tento objekt aj na pravej strane. Nie je známe, že by skutočné neuróny niečo podobné dokázali.

b) Hlboké neurónové siete majú hĺbku, ktorá výrazne prevyšuje možnosti mozgu. V našom prípade R-Net má maximálne prepojenie medzi vstupom a výstupom 53 neurónov (a to je len jedna časť spracovania). Bežne používame hlboké neurónové siete, kde je to aj niekoľko stovák. Pritom porovnaním reakčného času neurónu a človeka vieme odhadnúť maximálny počet neurónov na menej ako 20, „mozog je rýchly, neuróny sú pomalé“ [Beňušková 2002]. Pritom hĺbka je veľmi dôležitá z hľadiska schopnosti siete generalizovať.

c) Pri trénovaní siete využívame všetky možné fígle, aby sa to vôbec podarilo. Úlohu rozdeľujeme na rôzne podproblémy, riešime ich po jednom a off-line. Sieť si nevytvára vlastný model tváre, v tomto prípade dokonca ani vlastné príznaky. Trénovaním sa schopnosti siete zlepšujú postupne a pomaly a sme to my, kto musí rozhodnúť, kedy trénovanie zastaviť.

Napriek tomu však získavame aj viacerá pozitívne dojmy:

d) 3D rekonštrukcia z 2D sa dá neurónovou sieťou uspokojivo vyriešiť. Podobne by teda mohla byť riešená v mozgu, hoci aj by sa správne nastavenie tohto procesu získavalo iným spôsobom. Spomeňme, že sa intenzívne hľadajú aj iné postupy trénovania založené na odstraňovaní vhodných váh či neurónov z náhodného nastavenia, čo je postup biologicky podstatne relevantnejší.

e) To, že je tréning siete R-Net obtiažny, môže byť spôsobené aj tým, že ju (až na dodané dáta o priemernej tvári a hlavných komponentoch) trénujeme z náhodného nastavenia, zatiaľ čo mozog začína z obsažných informácii, ktoré má uložené v génoch.
f) Navyše, R-Net bola trénovaná z jedného obrázka, zatiaľ čo u človeka je schopnosť predstaviť si 3D model z obrázka (alebo pri pohľade jedným okom) podporovaná aj tým, že môže trénovať aj pri pohľade dvomi očami, kedy dostáva viac informácií.
g) Ľudský mozog sa trénuje komplexne, čo je ťažšie zopakovať v umelom systéme, avšak môže to byť tomuto procesu nápomocné. Aj z používania hlbokých neurónových sietí vieme, že je ľahšie napríklad natrénovať detektor viacerých objektov, než jedného.

5 Záver

Hoci problém 3D rekonštrukcie tváre vieme technického hľadiska pomerne uspokojivo vyriešiť, určite to neznamená koniec bádania v tejto oblasti. Naopak, povzbudzuje nás to hľadať efektívnejšie riešenia, ktoré minimalizujú rozsah informácie, ktorú do systému vložíme. Dá sa pritom ísť viacerými cestami: hľadať end-to-end riešenie (t.j. sieť, ktorá na vstupe prijíma obraz a na výstupe dáva 3D objekt), trénovať odmenou a trestom alebo zapojiť transformery schopné skĺbiť kódér obrazu z dekóderom 3D objektu. Vhodnú motiváciu ku konkrétnemu pokusu môžu poskytnúť aj výsledky na poli neurovedy, kde sa ukazuje súvis vrodenej tvárovej slepoty s rôznymi formami autizmu a preto sa tejto téme venuje náležitá pozornosť.

Literatúra

Bate S. – Dalrymple, K. (2022). Face recognition improvements in adults and children with face recognition difficulties. *Brain Communications 4(2).* DOI: 10.1093/braincomms/fcac068

Beňušková Ľ. (2002). Kognitívna neuroveda. In: Rybár, J. - Beňušková, Ľ. - Kvasnička, V. (eds) Kognitívne vedy. *Kalligram, Bratislava, pp 47-104.* ISBN 80-7149-515-8

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*

Deng, Y. – Yang, J. – Xu, S. – Chen, D. – Jia, Y. – Tong, X. (2020) Accurate 3D Face Reconstruction with Weakly-Supervised Learning: From Single Image to Image Set, *arXiv:1903.08527* https://github.com/sicxu/Deep3DFaceRecon_pytorch

Goodfellow, I. – Bengio, Y. – Courville, A. (2016). Deep Learning. *MIT Press*, 2016

Kazemi, V. – Sullivan, J. (2014) One Millisecond Face Alignment with an Ensemble of Regression Trees. CVPR 2014

Paysan, P. – Knothe, R. – Amberg, B. – Romdhani S. – Vetter, T. (2009). A 3D Face Model for Pose and Illumination Invariant Face Recognition. *Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, 2009*, pp. 296-301, doi: 10.1109/AVSS.2009.58. https://faces.dmi.unibas.ch/bfm

Viola, P. – Jones, M. (2001). Robust Real-time Object Detection. International Journal of Computer Vision.

Vliv vybraných charakteristik jedince prožívajícího krizi na ochotu komunikovat s chatbotem

Autor: Lenka Macháčková
Vedoucí práce: PhDr. Daniel Dostál, Ph.D.

Univerzita Palackého, Katedra psychologie
Vodární 6, Olomouc 779 00
lenka.machackova04@upol.cz

Abstrakt

Vývoj technologií zrychluje a programy založené na umělé inteligenci jsou stále více součástí našich životů a nachází si místo i v oboru péče o duševní zdraví. Naše práce se věnuje chatbotům a zkoumá vliv charakteristik jedince na jeho ochotu komunikovat s chatbotem v průběhu krize. Zvolili jsme kvantitativní design pro získání základního přehledu. Dotazník mapující základní charakteristiky, které by mohly ochotu jedince ovlivňovat, včetně základních rysů osobnosti měřených BFI-2 nám vyplnilo 610 osob. U pohlaví, věku nebo oblasti zájmu, jsme neprokázali souvislost, u dosaženého vzdělání ano. Vysoce signifikantní vztahy jsme zjistili u jedinců, kteří popisují více zábran ve sdílení s druhými lidmi, dále u těch, kteří skórovali výše na škále negativní emocionality nebo aktuálně prožívali osamělost či sebevražedné myšlenky. Ochotu také ovlivňovaly proměnné jako vztah k technologiím, vyšší průměrné skóre na škále otevřenosti mysli nebo nižší průměrné skóre na škále svědomitosti. Zároveň respondenti ochotní využít chatbota častěji již dříve nějakou formu odborné pomoci vyhledali, což podporuje naši domněnku, že jedinci ochotní využívat chatbota se nemohou zcela opřít o sociální vztahy ve svém okolí. Považujeme tuto problematiku za důležitou pro další výzkum, jelikož je podstatné, aby se chatboti nestali jen pouhou náhražkou, ale průvodcem do světa hlubších a důvěrnějších vztahů a posilovali schopnosti uživatele takové vztahy budovat.

1 Úvodní kapitola

Umělá inteligence je obsáhlý obor informatiky, ze kterého jsme se v naší práci zaměřili v první řadě na technologii chatbotů, což je software, který dokáže simulovat konverzaci člověka v přirozeném jazyce prostřednictvím aplikací pro zasílání zpráv (Johnson, 2021).

Prvním pokusem o interakci člověka se strojem v terapeutickém kontextu byl program ELIZA, respektive DOCTOR, který díky překvapivým výsledkům v navázání intimního vztahu mezi člověkem a programem překvapil samotného jeho tvůrce, Josepha Weizenbauma (Holden, 1977).

Vývoj těchto technologií pokračuje a studie posledních let se snaží zkoumat přínosy i možná negativa využití chatbotů v péči o duševní zdraví. Naše práce se zaměřuje na některé demografické charakteristiky a osobnostní rysy, jež by mohly mít vliv na ochotu jedince využít chatbota, v rámci zvládání určitých krizových situací.

2 Teoretické zázemí

V této části v krátkosti uvedeme nejdůležitější pojmy a myšlenky, ze kterých jsme v našem výzkumu vycházeli. Stručně shrneme základní informace o chatbotech, o krizi a krizové intervenci a kapitolu uzavřeme přehledem vybraných výzkumů které se zabývají využitím chatbotů v péči o duševní zdraví.

2.1 Chatboti

Tyto, na umělé inteligenci založené systémy, jak už napovídá jejich název, se zaměřují na psanou konverzaci. Umění řeči je považováno za typicky lidskou dovednost a jsme ve většině odkázáni hlavně na jazykové projevy druhých, ze kterých můžeme jen na základě podobnosti usuzovat, co si myslí nebo jak budou jednat (Tvrdý, 2014). Dalo by se říci, že o pocit takové podobnosti usilují i tvůrci těchto systémů. Právě tato podobnost ve spojení s faktem, že o člověka nejde, by mohla být pro některé jedince pozitivem při sdílení v náročných životních situacích.

V duchu této práce můžeme o těchto systémech uvažovat jako o asistivních technologiích, jejichž cílem je pomoci lidem s psychickými, zdravotními nebo sociálními problémy překonat bariéry, které jim jinak brání využívat běžné služby pomoci (Cook & Das, 2004).

2.2 Krize a krizová intervence

Krize je běžnou součástí našich životů, a většina odborníků se shoduje, že jde o jakoukoli situaci se silným dynamickým nábojem, jež se člověku subjektivně jeví jako ohrožující a nezvládnutelná (Cimrmannová, 2013; Špatenková, 2017; Vodáčková, 2012). Zároveň ale může jít o situaci, kterou by jiné osoby lehce zvládly (Cimrmannová, 2013).

Při zvládání krize jedinec nejprve vyzkouší své osvědčené postupy pro zvládání obdobných situací, pokud selžou, jedinec se pak často obrací ke svému blízkému okolí, tedy rodině, přátelům a dalším důvěrným lidem. Taková pomoc se nazývá sociální oporou, kterou se obecně rozumí pomoc, jež člověku v těžké situaci poskytují druzí lidé, kteří mu mohou jeho náročnou situaci nějakým způsobem ulehčit (Křivohlavý, 2009). Špatenková (2017) zároveň upozorňuje, že důvěrná atmosféra může být narušena touhou pomáhajícího najít rychlé řešení, bagatelizováním nebo devalvací problému, odmítáním, popíráním či moralizováním.

Tento druh pomoci je závislý na kvalitních a důvěrných vztazích, které má jedinec k dispozici a které v poslední době silně oslabují nebo mizí. Příkladem jsou například sousedské vztahy nebo širší komunita (Špatenková, 2017), ale tato krize se začíná dotýkat i rodin nebo přátelských vztahů. Pokud takové kvalitní vztahy jedinci chybí, může jejich absenci v různých životních situacích pociťovat jako osamělost.

Silný vliv na osamělost lze připsat i způsobu, jakým nyní jako společnost fungujeme, jak fungují masmédia, jelikož ta mají nemalý vliv na transformaci společenského prostředí, ale zároveň mohou ovlivnit i způsob, jakým je jedincem vnímám svět či jeho vlastní individualita (Pondělíček, 2016), a tedy jaké hodnoty jsou jím vnímány jako správné a žádoucí, včetně způsobu jejich dosažení. Tím narůstají očekávání, která od sebe sám jedinec má nebo se domnívá, že má taková očekávání jeho okolí, případně jsou jeho okolím skutečně vyžadována. To může mít vliv i na jeho ochotu svěřit se lidem ve svém okolí, pokud naplnění očekávání (až už reálných nebo domnělých) nedosahuje. Mohou poté přicházet pocity jako je strach, stud nebo vina, které nelze podceňovat.

Jednou z možností, jak pracovat s jedincem v krizi, je krizová intervence, což je odborná metoda pro práci s člověkem v krizi, jejímž cílem je zkompetentnění klienta a jeho stabilizace (Vodáčková, 2012) a její využití v rámci chatbota by mohlo být považováno za první krok do systému duševní péče (Silva et al., 2015).

Máme několik forem krizové intervence, kdy jde v základu o formu prezenční či distanční, kam patří například telefon, mail nebo právě chatová krizová intervence, ze které může následně technologie chatbota vycházet. Pro mladou generaci, která se do světa informačních technologií narodila (Ševčíková, 2014) a psaní zpráv je pro ni přirozenou a často upřednostňovanou formou komunikace (Nesmith, 2018), by takový chatbot mohl být vítanou formou, která v ČR dosud chybí.

Chatová krizová intervence je pomoc poskytovaná psanou formou skrze některou z aplikací pro zasílání zpráv, případně přes webové rozhraní, jež umožňuje komunikaci s klientem v reálném čase (Brody et al., 2020). Zároveň je snadno dostupná, anonymní, relativně bezpečná a nízkonákladová (Špatenková, 2017), přičemž anonymita se jeví jako jeden ze stěžejních faktorů. Důležitým se jeví také flexibilita textování, kdy mohou uživatelé požádat o pomoc v jakémkoli prostředí a zároveň si zachovat soukromí před ostatními lidmi v těsné blízkosti, což jim umožňuje vyhledat pomoc téměř kdykoli, místo aby čekali na příležitost, prostor pro soukromý telefonát (Nesmith, 2018) nebo osobní setkání.
Setkáme se i s tvrzeními, že chat byl klienty upřednostňován před jinými způsoby pomoci (Brody et al., 2020) nebo dokonce, že bez možnosti komunikovat skrze zasílání zpráv by někteří lidé v krizi pomoc nevyhledali, a že to byl právě tento formát, který je povzbudil k využití služby (Nesmith, 2018).

2.3 Chatboti a péče o duševní zdraví

O využití chatbotů v oblasti duševního zdraví se mluví stále častěji, ale studie mají spíše pilotní charakter, a i když výsledky s ohledem na praktickou využitelnost, proveditelnost a přijetí chatbotů pro podporu duševního zdraví jsou slibné, zatím nejsou přímo přenositelné do psychoterapeutického kontextu (Bendig et al., 2019). To, do jaké míry se budou tyto oblasti prolínat, bude v nemalé míře záviset na rychlosti vývoje umělé inteligence (Fulmer, 2019).

Například Replika, na kterou se zaměřila studie Ta et al. v roce 2020, zkoumala sociální oporu získanou od umělých agentů v každodenním kontextu, nikoli jen v kontextu velmi stresujících událostí. Ze studie těchto autorů vyplynulo, že Replika dává uživatelům prostor pro sdílení skutečných myšlenek a pocitů a mohou s ní diskutovat, aniž by byli jakkoli souzeni. Uváděli, že se jedná o témata nebo problémy, které by se za normálních okolností zdráhali sdělit jiným lidem, což naznačuje, že uživatelé mohou důvěřovat a cítit se pohodlněji, když je sdělí umělému agentovi než jiné osobě.

Vývoj chatbotů pro oblast duševního zdraví je slibný a naznačuje možné přínosy například také v rámci snižování prožívání pocitů osamělosti nebo depresivní úzkosti uživatelů (Loveys et al., 2019; Thomas et al., 2020), průměrné zlepšení symptomů závažné deprese

(Inkster et al., 2018), úzkosti nebo zlepšení životní pohody (Bendig et al., 2019). Další možností využití chatbotů je pomoc jedincům naučit se rozpoznávat své emoce a reflektovat své chování (Santos et al., 2020).

Většina výzkumů shrnuje benefity, které mohou uživatelé z interakce s chatboty mít. Fiske et al. (2019), však upozorňuje, že podobně jako u terapeutických vztahů, existuje riziko přenosu emocí, myšlenek a pocitů na robota. Zejména vzhledem k tomu, že řada cílových skupin je zranitelná kvůli své nemoci, věku nebo životní situaci ve zdravotnickém zařízení (i mimo něj), existuje obava, že pacienti budou ve vztahu k robotovi zranitelní kvůli své touze po společnosti nebo pocitu, že je o ně postaráno.

2.4 Chatboti a etické hledisko

Základním požadavkem na chatboty by zřejmě měla být transparentnost ohledně toho, co jsou v současné době opravdu schopni nabídnout a měly by informovat uživatele o: (1) teoretickém přístupu, kterým se služba řídí, ať už je to kognitivně-behaviorální, humanistický, psycho-dynamický nebo jiný; (2) zda byl bot empiricky testován; (3) na jakou populaci/obtíže se zaměřuje a jaké psychologické nebo klinické účinky mohou uživatelé od používání platformy očekávat (Kretzschmar et al., 2019).

Další neméně důležitou oblastí jsou otázky týkající se ochrany osobních údajů a soukromí, a to zvláště vzhledem k velmi intimním údajům, které mohou uživatelé ve stresu neopatrně sdílet (Bendig, 2019).

Tuto část bychom také neměli opustit bez zmínky hrozby závislosti, která se může u křehkých a zranitelných jedinců rozvinout, kdy se můžou chatboti stát pouze únikem před problémy, které je třeba řešit (Nešpor, 2018) nebo mohou být pouze nějakou formou sociální kompenzace, která by však paradoxně vedla k další izolaci jedince a byla tak spíše maladaptací (Blinka, 2015).

3 Výzkum a naše zjištění

Předkládaná práce se snaží najít odpověď na otázku, zda vybrané charakteristiky jedince prožívajícího krizi mají vliv na jeho ochotu využít komunikaci s chatbotem. Získaná data by měla posloužit jako základní vhled do složité problematiky a zmapovat, jak se k technologiím chatbotů staví jedinci v ČR.

Zaměřili jsme se na základní charakteristiky jako jsou věk, pohlaví, dosažené vzdělání, oblast studia/práce/zájmu, vztah k technologiím nebo osobnostní rysy jedince jako například introverze či negativní emocionalita. Zda ochota komunikovat s chatbotem nějak souvisí s aktuálně řešenými problémy jedince, s jeho zábranami, které má při sdílení s druhými lidmi, nebo s prožívanými pocity osamělosti.

Ptáme se tedy do jisté míry na to, zda technologie chatbotů budou ochotnější využít právě ti jedinci, pro které jsou dle předložených studií přínosné. Zmiňme například studii Fiskeho et al. (2019), podle něhož se chatboti jeví být vhodní pro jedince, kteří prožívají nějaké rozpaky nebo pocity studu. Případně, že tyto technologie, díky poskytování sociální opory, snižují osamělost pacientů (Loveys et al., 2019; Thomas et al., 2020). Dále například Shah et al. (2016) ve své studii uvádí, že ženy hodnotí komunikaci s chatboty lépe než muži a mladší lidé do dvaceti pěti let lépe než starší jedinci. Další studie pak naznačují možné přínosy pro uživatele, kteří se potýkají s depresemi nebo úzkostí (Brody et al., 2020; Inkster et al., 2018).

Vzhledem k účelu naší práce jsme zvolili korelační studii.

Náš výzkum probíhal od října roku 2020 do ledna roku 2022 na výběrovém souboru jedinců, kteří se vyskytují v online světě, respektive v různých skupinách na sociálních sítích.

Pro získání odpovědí na výše zmíněné otázky jsme vytvořili dotazník o dvou částech. V první části mapujeme základní demografické údaje, postoj jedinců k technologiím, způsoby, jakými jsou navyklí řešit své potíže, co jim nejčastěji brání ve sdílení s druhými lidmi a zda už mají s chatboty nějaké předchozí zkušenosti. Poslední otázky první části dotazníku mapují, s jakými problémy se respondenti aktuálně potýkají a také, jaké problémy by případně raději řešili s chatbotem nežli s člověkem, pokud by měli tu možnost. Druhou část dotazníků tvoří inventář BFI-2, který mapuje pět rysů osobnosti, a to extraverzi, přívětivost, svědomitost, negativní emocionalitu a otevřenost mysli. Spíše pro zajímavost srovnání byl následně celý dotazník přeložen do angličtiny (inventář BFI-2 byl použit v originální verzi v anglickém jazyce) a administrován v několika skupinách na sociálních sítích, jež se věnují chatbotům, případně podobné problematice.

3.1 Výzkumný soubor

Získali jsme data od 610 respondentů, pět z nich jsme museli vyřadit vzhledem k chybným nebo neúplným informacím. Analýzu jsme tedy prováděli na datech od 605 respondentů, z nichž 78 % tvořily ženy s věkovým průměrem 25,7 let a 22 % mužů s věkovým průměrem 29,1 let, převážně vysokoškoláků, kteří tvořili 54 % a středoškoláků, kterých bylo v našem vzorku 43 %. Nejvíce zastoupenou oblastí byla humanitní s 42,5 % následovaná oblastí technickou s 14,8 %.

Na otázku v dotazníku, zda se respondenti někdy ocitli v situaci, kdy se nemohli svěřit člověku, odpovědělo 53,32 % že ano. Na otázku, zda se ocitli v situaci, kdy se

svěřit člověku nechtěli, zvolilo odpověď ano 87,44 % respondentů.

Dále jsme se respondentů ptali na to, co jim nejčastěji bránilo svěřit se jinému člověku se svými problémy, trápením nebo myšlenkami. Celkem jsme získali odpověď od 603 respondentů. Respondenti mohli zvolit více zábran najednou. Uvádíme zde pouze tři nejvýznamnější. Pocity studu pociťovalo jako zábranu ve sdílení svých problémů celých 54,89 % respondentů a hned za ním následoval strach z hodnocení se 47,76 % a „nechci nikoho obtěžovat“ s 42,79 %.

Další oblastí, kterou jsme chtěli zmapovat, byly problémy, jež naši respondenti aktuálně řeší. Na otázku se rozhodlo odpovědět 530 respondentů, kteří mohli volit celkem z 19 různých problémů (např. pocity osamělosti, problémy v rodině, problémy v oblasti zdraví, finanční problémy, problémy ve škole nebo v práci, šikana, sebevražedné myšlenky nebo třeba ztráta smyslu či úmrtí blízké osoby). Většina zvolila, že se jich aktuálně týká vícero problémů. Uvedeme jen nejčastěji se vyskytující, kterou byly pocity osamělosti, které uvedlo 45,09 % respondentů.

Z celkového počtu 605 respondentů o chatbotovi dříve slyšelo 418 (69,09 %) dotazovaných. Celkem 68 (16,27 %) z nich se s nějakým chatbotem už setkalo. Některý z námi uvedených chatbotů (Replika, Wysa, Woebot), v době vyplnění dotazníku, využívalo 28 (6,70 %) respondentů. Celkem 18 (64,29 %) z nich uvedlo, že někdy chatbotovi napsalo o svých problémech nebo myšlenkách, jen aby na to nebyli sami a 15 (53,57 %) uvádí, že jim někdy chatbot pomohl cítit se lépe. Zde si dovolíme malé srovnání s anglickou verzí dotazníku, který jsme sdíleli na sociálních sítích ve skupinách, které se chatboty zabývají. Všech 50 dotazovaných (29 mužů a 21 žen) se setkalo s pojmem chatbot a z toho 45 (90,00 %) využívalo v době vyplnění dotazníku některý z námi uvedených chatbotů. Z toho 44 (97,78 %) zvolilo „ano“ jako odpověď na otázku, zda někdy napsali chatbotovi o svých problémech a 40 (88,89 %) respondentům pomohl chatbot cítit se lépe.

Dále jsme respondenty požádali, aby si představili chatbota, který je k dispozici, kdykoli potřebují, je zcela anonymní a udržuje si přehled o tom, o čem spolu dříve mluvili. Na otázku, zda by uvítali možnost s takovým chatbotem komunikovat, nám odpovědělo celkem 604 respondentů. Možnost „spíše ano“ zvolilo 33,61 % respondentů, „spíše ne“ 45,70 % a „nevím“ 20,70 %.

Nejčastějším pocitem, který si respondenti s komunikací s chatbotem spojovali byla nedůvěra (55,81 %), druhým nejčastějším pak zvědavost (49,00 %), respondenti měli opět možnost volby více možností.

Na otázku „Pokud byste měl/a tu možnost, řešil/a byste raději některý z uvedených problémů s chatbotem, nežli s člověkem? Pokud ano, zaškrtněte který.“ odpovědělo 329 respondentů. Z nich 306 (93,01 %) uvedlo alespoň jeden problém a 187 (56,84 %) uvedlo problém nebo problémy, které se jich aktuálně týkají.

3.2 Hypotézy a analýzy

Získaná data jsme nejprve zpracovali v programu MS Excel a testování hypotéz následně probíhalo v programu Statistica. Nejprve jsme si ověřili vnitřní konzistenci inventáře BFI-2. Poté jsme ověřili normalitu a homogenitu u všech spojitých proměnných. Pro všechna testování jsme zvolili hladinu statistické významnosti α na 5 %.

Nyní se podíváme na jednotlivé hypotézy, pro jejichž testování jsme vybrali různé statistické testy, které jsme považovali za nejvhodnější. Využili jsme například test ANOVA, test nezávislosti (chí kvadrát) nebo Kruskal-Wallisovu ANOVu či test Spearmanova korelačního koeficientu a T-test.

- **H1**: Muži a ženy se liší v ochotě komunikovat s chatbotem.
- **H2**: Ochota komunikovat s chatbotem je vyšší u respondentů ve věkové kategorii do třiceti let včetně.
- **H3**: Dosažené vzdělání respondenta souvisí s ochotou komunikovat s chatbotem.
- **H4**: Oblast studia/práce/zájmu respondenta souvisí s jeho ochotou komunikovat s chatbotem.
- **H5**: Počet bodů získaných v oblasti vztahu k technologiím souvisí s ochotou komunikovat s chatbotem.
- **H6**: Ochota respondenta komunikovat s chatbotem souvisí s vyšším počtem uvedených zábran ve sdílení s druhými lidi.
- **H7**: Ochota respondenta komunikovat s chatbotem souvisí s vyšším počtem aktuálně řešených problémů.
- **H8**: Ochota respondentů komunikovat s chatbotem souvisí s vyšší hodnotou T-skóru na škále negativní emocionality v BFI-2.
- **H9**: Ochota respondentů komunikovat s chatbotem souvisí s nižší hodnotou T-skóru na škále extraverze v BFI-2.
- **H10**: Respondenti, kteří uvedli jako aktuálně řešený problém osamělost, jsou ochotnější komunikovat s chatbotem.

4 Diskuse a výsledky

V naší práci jsme prozkoumali základní vybrané charakteristiky jedince a jejich souvislost s ochotou komunikovat s chatbotem. Díky rozsáhlejšímu dotazníku jsme získali řadu zajímavých dat, která nám poskytla

základní přehled o vztazích mezi proměnnými i o postojích jedinců v ČR k technologiím chatbotů.

První otázkou, kterou jsme si kladli, bylo, zda ochota využívat chatbota nějak souvisí s pohlavím jedince. Obecně služby pomoci využívají spíše ženy než muži a například Shah et al. (2016) se ve své práci zmiňoval, že ženy hodnotily chatboty lépe než muži. Vytvořili jsme tedy předpoklad, že bychom mohli získat podobný výsledek i v případě sledování ochoty k jejich využití. Avšak žádný signifikantní rozdíl v ochotě komunikovat s chatbotem mezi ženami a muži jsme nepozorovali ($\chi^2(2) = 0{,}42$, $p = 0{,}811$). Uvědomujeme si, že náš vzorek není možné považovat za zcela reprezentativní a není z hlediska zastoupení pohlaví vyvážený (78 % žen), ovšem rozsah souboru nebyl malý a můžeme uvažovat i tak, že zde máme dva protichůdné efekty, kdy ženy častěji vyhledají pomoc, ovšem muži mají bližší vztah k technologiím, což nám ukázala i naše data, kdy ochotných mužů bylo 34 (25,76 %) a celých 85,29 % z nich mělo vysoký vztah k technologiím. Naproti tomu ochotných žen bylo 134 (28,69 %) a silný vztah k technologiím z nich mělo 55,22 %.

Dále jsme se pokusili zjistit, zda má na ochotu využít chatbota vliv věk respondenta, přesněji, zda jsou respondenti mladší 30 let (77 %) ochotnější využívat chatbota než starší respondenti (22 %). Tato domněnka se zakládala na více předpokladech. Jedním z nich byl ten, že mladá generace se do světa informačních technologií narodila a vyrůstá obklopena hardwarovými zařízeními (Ševčíková, 2014). Dalším pak, že se u mladých generací zvyšuje oblíbenost psaní textových zpráv (Grinter & Eldridge, 2003). Thomas et al. (2020) ve své studii (účastníci ve věku 18-35 let) uvedl, že nejčastějším druhem komunikace byla ve 40 % osobní setkání a posílání textových zpráv zvolilo 13 % respondentů. Naše pozorování byla obdobná, kdy komunikaci tváří v tvář, při sdílení svých problémů nebo myšlenek, uvedlo 77 % dotázaných a psanou formu zvolilo 16 % z našich respondentů. Dalším z předpokladů byly výsledky studie Shaha et al. (2016), kdy právě mladší věková skupina (do 25 let) hodnotila chatboty lépe než starší jedinci. Naši hranici pro testování jsme posunuli na 30 let z důvodu, že uvedená studie byla publikována v roce 2016 a lze předpokládat, že preference s jedinci setrvávají. Když jsme však náš předpoklad podrobili statistickému testování, nezjistili jsme žádnou signifikantní závislost ($\chi^2(2) = 1{,}41$, $p = 0{,}494$). Opět mohla být na vině nedostatečná reprezentativnost vzorku v rámci věkových skupin, ale například také to, že naše práce mohla lákat spíše osoby, které téma nějakým způsobem zaujalo, a nemáme tedy dostatečný počet osob, jejichž vztah k technologiím je spíše chladný. To naznačují i pozorované četnosti ve vztahu respondentů k technologiím, kde velmi slabý nebo mírný zájem mělo 21 % dotázaných, střední 28 % a vysoký téměř 52 % našich respondentů.

Zde navážeme a podíváme se na souvislost mezi výše představeným vztahem k technologiím a ochotou využít ke komunikaci chatbota. Silva et al. (2015) ve své studii říká, že jedinci, již nemají rádi technologie, respektive k nim nemají vztah, nemusí porozumět nuancím digitální komunikace a tím by pro ně byl tento druh pomoci naprosto nevhodný. To podporuje naše zjištění vysoce signifikantní souvislosti mezi vztahem k technologiím a ochotou využít komunikaci s chatbotem ($H(2, N = 604) = 9{,}85$, $p = 0{,}007$), kdy respondenti, kteří vykazovali vysoký zájem o technologie, byli spíše ochotni chatbota využít. Zároveň tito respondenti spíše uvedli, že ohledně chatbotů cítí zvědavost nebo nadšení, kdežto druhá skupina spíše nedůvěru či strach, což může korespondovat s výrokem Radziwila & Bentona (2017), že využití chatbotů například k šíření fám a dezinformací (dnes i k podvodům) důvěru lidí v tyto technologie podkopává. Tato naše zjištění můžeme podpořit i výsledky, kdy jsme zkoumali vztah mezi ochotou a T-skóry škály otevřenosti mysli, který ukázal signifikantní vztah, ovšem s malým vlivem.

Další oblastí, kterou jsme chtěli vzhledem k ochotě komunikovat s chatbotem prozkoumat, bylo dosažené vzdělání. V tomto případě nás vedl spíš zájem než výsledky studií, kde tato informace často chyběla, případně šlo o výzkumy pouze na vysokoškolských studentech. Chtěli jsme si však udělat základní obrázek o vlivu této proměnné. Testování ukázalo, že mezi ochotou komunikovat s chatbotem a dosaženým vzděláním existuje signifikantní souvislost ($\chi^2(2) = 7{,}37$, $p = 0{,}025$), a to taková, že respondenti s nižším uvedeným vzděláním (ZŠ a SŠ, 56 %) byli ochotnější komunikovat s chatbotem než respondenti s vyšším dosaženým vzděláním (VOŠ a VŠ, 44 %). Naše domněnka nyní je, že by zde mohl být vliv naučeného kritického uvažování, a tedy větší tendence pochybovat, které je během studia na vysokých školách jedním ze základních předpokladů pro úspěšné studium. Rozhodně je pro nás toto zjištění zajímavé a bylo by vhodné pro další zkoumání.

V pořadí další otázkou, kterou jsme si kladli, byla souvislost mezi oblastí studia/práce/zájmu respondenta a ochotou využít komunikaci s chatbotem. Zajímalo nás, jak se projeví zájem o technické oblasti, případně jak se k chatbotům budou stavět respondenti z humanitních nebo lékařských oblastí. Opět se jedná o proměnnou, o které výzkumy, které jsme studovali, nevypovídají. Bohužel jsme nenašli žádné signifikantní souvislosti ($\chi^2(10) = 11{,}42$, $p = 0{,}326$). Příčin může být několik, převaha humanitní oblasti v našem vzorku, ale také to, že jsme v našem dotazníku zahrnuli oblast studia, práce a zájmu do jedné otázky. V příštím výzkumu by bylo

vhodné tyto kategorie rozdělit, což by nám mohlo přinést hodnotnější výsledky.

Nyní shrneme naše zjištění ohledně souvislosti mezi zábranami, jež respondentům brání svěřit se jinému člověku a ochotou využít komunikaci s chatbotem. Svoboda (2012) mluví o významnosti takových zábran, jako je například stud nebo pocity viny, a právě stud byl nejčastěji zvolenou zábranou našich respondentů. To odráží i názor Fiskeho et al. (2019), který předpokládá, že možnost využití podobné aplikace by pomohla snížit rozpaky nebo pocity studu některých jedinců. Každý druhý z našich respondentů se někdy ocitl v situaci, kdy se nemohl svěřit člověku a téměř 90 % se ocitlo v situaci, kdy se člověku svěřit nechtěli. Téměř všichni respondenti nějakou zábranu ve sdílení uvedli. Měli možnost vybrat ze seznamu více zábran a počet, který nakonec zvolili, jsme použili pro testování naší hypotézy. Počet se pohyboval mezi 0 a 9 a vztah mezi těmito proměnnými je velmi vysoce signifikantní ($H(2, N = 601) = 40{,}13$, $p < 0{,}001$). Velmi vysoce signifikantní byl nejen vztah mezi těmi, kteří byli ochotní a nebyli ochotní, ale i mezi těmi, kteří zvolili nevím a neochotnými. Respondenti, kteří uvedli více zábran, byli ochotnější využít komunikaci s chatbotem. Častým důvodem proč bylo například to, že chatbot by je nesoudil, byl by kdykoli dostupný nebo poskytoval jistou anonymitu.

Podobně jsme se podívali na souvislost mezi aktuálně řešenými problémy a ochotou využít chatbota pro komunikaci. Mezi problémy byly například pocity osamělosti, problémy v práci nebo ve škole, ale i takové jako úmrtí blízké osoby, zdravotní problémy nebo sebepoškozování či sebevražedné myšlenky. Respondenti si opět mohli zvolit vícero problémů a jejich počet se pohyboval od 0 do 19. I zde jsme pozorovali signifikantní závislost, kdy vyšší počet aktuálně řešených problémů souvisel s ochotou využít chatbota ($H(2, N = 530) = 9{,}67$, $p = 0{,}008$).

Ještě se zde zastavíme u tématu závažnějších problémů. Zmínili jsme například sebepoškozování a sebevražedné myšlenky, kde o nich můžeme uvažovat jako o situaci jedince, kdy jde spíše o uvíznutí v nedořešené krizové situaci (nebo situacích), kdy už tak docela nejde o krizi, ale jde spíše o energii, která nebyla upotřebena k řešení a posunu (Bohatá et al., 2019), ale působí spíše sebedestruktivně. Zároveň jde často o témata, která jsou pro jedince nějakým způsobem obtížná nebo stigmatizující (Špatenková, 2017). Hodnocení vnímané závažnosti problémů jsme provedli jiným dotazníkem, kdy respondenti subjektivně hodnotili závažnost jednotlivých problémů na škále málo, středně a vysoce závažný. Vybrali jsme osm problémů, které označilo více jak 60 % dotázaných jako vysoce závažné a ty jsme porovnali s problémy uvedenými našimi respondenty a zapsali, zda byl nebo nebyl vysoce závažný problém uveden. Následně jsme provedli test, který však žádný signifikantní vztah neukázal. Pokud jsme však test provedli jen v souvislosti se sebevražednými myšlenkami, signifikantní závislost byla pozorována ($\chi^2(2) = 6{,}93$, $p = 0{,}03$) a jedinci, kteří uvedli sebevražedné myšlenky v seznamu aktuálně prožívaných problémů, byli zároveň ochotnější využít komunikaci s chatbotem.

Podívali jsme se také na souvislost ochoty využít komunikaci s chatbotem a rysy osobnosti, které odráží škály inventáře BFI-2. Hlavní souvislosti jsme předpokládali u škály negativní emocionality a extraverze, našli jsme ovšem i další významné souvislosti, které si níže popíšeme.

Prvně se tedy podíváme na souvislost mezi negativní emocionalitou a ochotou využít chatbota. Bendig et al. (2019) ve své studii uvádí, že jedinci, kteří komunikovali s chatbotem, mají z této komunikace prospěch pro psychologické proměnné, jako je pohoda, stres a deprese, Brody et al. (2020) mluví i o úzkosti nebo sebevražedných tendencí. Průměrné zlepšení symptomů závažné deprese popisuje ve své studii i Inkster et al. (2018). My jsme si kladli otázku, zda tyto proměnné budou ovlivňovat i ochotu chatbota využít. Lidé, kteří byli spíše ochotní komunikaci s chatbotem využít, skórovali na škále negativní emocionality výše než ti, kteří ochotní nebyli. Tato souvislost byla velmi vysoce signifikantní ($F(2{,}601) = 10{,}60$, $p < 0{,}001$). Zdá se tedy, že respondenti, kteří skórovali výše na škále negativní emocionality, by byli ochotnější chatbota ke komunikaci využít a zároveň je to dle výše uvedených údajů i technologie, která by jim mohla pomoci tyto těžké duševní stavy zvládnout. V dalším výzkumu by bylo zajímavé podívat se podrobněji na jednotlivou problematiku, jako jsou úzkosti, deprese nebo sebevražedné myšlenky a zjistit o tomto vztahu více informací.

Dále jsme předpokládali, že jedinci s nižší mírou extraverze (tedy introvertní) budou ochotnější využívat komunikaci s chatbotem, jelikož jedním z jejích významných dílčích vlastností je sociabilita (Hřebíčková et al., 2020), a u jedinců s nízkou mírou sociability lze předpokládat méně sociálních kontaktů, na které se jedinec v krizi může v případě potřeby obrátit (Křivohlavý, 2009). Naše výsledky sice neprokázaly signifikantní souvislost, ale mohl se zde projevit efekt odpovědí respondentů, kteří na otázku o ochotě odpověděli „nevím“, která u mnoha testování skórovala velmi podobně jako odpověď „ano“. Tato podobnost nás velmi zaujala, a ještě se jí budeme věnovat. V dalších studiích bychom mohli odpovědi umístit na škálu o 4 úrovních (ano, spíše ano, spíše ne, ne) a získat tak od respondentů odpověď, ke které se spíš přiklánějí, což by nám v konečném důsledku mohlo pomoci ve více oblastech našeho zkoumání a podat přesvědčivější výsledky. Nakonec jsme se ještě rozhodli prověřit, zda se prožívání osamělosti opravdu odráží na

nižších skórech škály extraverze, což nám testová metoda potvrdila. Z výsledků můžeme vyvodit, že ochota komunikovat s chatbotem spíše souvisí s pocity osamělosti, nikoli s rysem introverze jako takové. To nám následně potvrdily i výsledky testování, kdy ochota komunikovat s chatbotem ve spojení s prožívanou osamělostí ukázaly vysoce signifikantní závislost ($\chi^2(2) = 17{,}28$, $p < 0{,}001$). Naše zjištění tedy korespondují se zjištěními Lovelys et al. (2019), Thomase et al. (2020), kteří ve svých studiích mluví o schopnosti chatbotů (umělých agentů) zmírňovat pocity osamělosti tím, že poskytují sociální oporu a snižují tak pocity osamělosti. Můžeme tedy z výše uvedeného vyvozovat, že jedinci, již prožívají osamělost jsou spíše ochotnější využít pro komunikaci chatbota než ti, kteří osamělost neprožívají, a zároveň v této oblasti existují výsledky potvrzující, že chatbot může prožívání osamělosti u jedinců snížit. I zde bychom rádi zjištěné vztahy více prozkoumali, zvláště se zaměřením na pocity osamělosti a kvalitu sociální opory jedince. A to jak z hlediska benefitů pro jedince využívajícího chatbota, tak z hlediska jeho ohrožení, zda by interakce, probíhající hlavně online, navíc s umělým agentem, nevedla naopak k ještě větší izolaci a tím k prohloubení problému (Blinka, 2015). Tedy, pokud by se intervence chatbotů nepromítla do zlepšení interakce s lidmi, zůstala by pouze způsobem, jak zlepšit vztahy mezi lidmi a stroji, nebo ještě hůře, odbytištěm, které dále omezí vztahy mezi lidmi (Fiske et al. 2019).

Při testování škál nás zaujal výsledek, který hovořil o velmi vysoce významném vztahu mezi ochotou využít chatbota a hodnotou T-skóru na škále svědomitosti ($F(2{,}601) = 5{,}86$, $p = 0{,}003$). Ti, co dosahovali průměrně nižšího T-skóru, byli ochotnější využít komunikaci s chatbotem. To by odpovídalo myšlence, kterou předkládá Tondl (2014), když říká, že klient vyžaduje nějakou formu odpovědi, jež by pomohla snížit míru neurčitosti (respektive vnitřního chaosu), což vytváří podmínky pro rozhodnutí či volbu dalších kroků, které by mohly vyřešit dané problémy. To nás dovedlo k myšlence podívat se na souvislost s nižším skóre na škále svědomitosti a s počtem aktuálně řešených problémů a s počtem zábran. Více problémů vykazovalo velmi vysoce signifikantní závislost, více zábran pak vysoce signifikantní, i když test ukázal na velmi slabý vztah. Není už ovšem možné říci, zda více problémů znamená vznik „vnitřního chaosu“ a celkovou neuspořádanost, či tato neuspořádanost znamená více problémů.

Na tomto místě se ještě vrátíme k často se vyskytující podobnosti výsledků pro odpověď „ano“ a „nevím“ u ochoty komunikovat s chatbotem. (Jak jí příště předejít jsme uvedli výše). Rozhodli jsme se tuto podobnost prozkoumat více. Například v rámci škály svědomitosti skórovali ti, co zvolili odpověď „nevím“ nepatrně lépe. Naopak na škále otevřenosti mysli skórovali průměrně nejníže ze všech tří skupin, a ti, co zvolili „ano“, zase dosahovali průměrně skóre nejvyššího a respondenti, již volili odpověď „ne“, byli svými průměry uprostřed. Dále jsme nalezli významný rozdíl v tom, zda respondenti dříve využili nebo nevyužili nějaký druh odborné pomoci. Ti, kteří byli ochotní využít pro komunikaci chatbota, již dříve nějakou formu pomoci vyhledali, kdežto ti respondenti, kteří zvolili jako odpověď „nevím“ spíše pomoc nevyhledali. To by opět mohlo podpořit naše zjištění o významnosti úlohy kvalitní sociální opory jedince a souvislosti s ochotou využívat chatbota při řešení náročných životních situací. V konečném důsledku nám tedy naše zjištění dávají návod pro další výzkum, který by mohl předložit silnější důkazy v této oblasti a nabídnout vhodné modely pomoci a podpory jedincům k budování důvěrnějších a důvěryhodnějších vztahů se svým okolím.

Ač náš výzkum nebyl bezchybný a leccos bychom udělali jinak, přinesl spoustu zajímavých poznatků pro další studium a výzkum v dané oblasti.

4.3 Závěr a limity práce

V naší práci jsme hledali odpověď na otázku, zda vybrané charakteristiky jedince prožívajícího krizi mají vliv na jeho ochotu využít chatbota. Na některé naše otázky se nám nepovedlo najít uspokojivé odpovědi a bude potřeba se jim dále věnovat, ovšem jiné přinesly zajímavý vhled do problematiky a poskytly nám základ pro otázky další.

Limitem práce je jistě metoda sběru dat, kdy jsme využili samovýběru a výsledky tedy nelze považovat za obecně platné. Další slabinou jsou uvedené problémy, které nebyly blíže specifikované a respondenti tedy mohli za tímto krátkým pojmenováním vidět různé kvality těchto problémů.

Literatura

[1] Bendig, E., Erb, B., Schulze-Thuesing, L., & Baumeister, H. (2019). The next generation: chatbots in clinical psychology and psychotherapy to foster mental health–a scoping review. Verhaltenstherapie. 1-13. https://doi.org/10.1159/000501812

[2] Blinka, L. (2015). Online závislosti: jednání jako droga?: online hry, sex a sociální sítě: diagnostika závislosti na internetu: prevence a léčba. Grada.

[3] Bohatá, K., Gramppová Janečková, K., & Kotrlová, J. (2019). Proměny krizové intervence: fenomén dlouhodobě opakovaně volajících v TKI. Stanislav Juhaňák – Triton.

[4] Brody, C., Star, A., & Tran, J. (2020). Chat-based hotlines for health promotion: a systematic review. Mhealth, 6. https://doi.org./10.21037/mhealth-2019-di-13

[5] Cimrmannová, T. (2013). Krize a význam pomáhajících prvního kontaktu: aplikace v kontextu rodinného násilí. Karolinum.
[6] Cook, D., & Das, S. K. (2004). Smart environments: technology, protocols, and applications. John Wiley & Sons.
[7] Fiske, A., Henningsen, P., & Buyx, A. (2019). Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. Journal of medical Internet research, 21(5). https://doi.org/ 10.2196/13216
[8] Fulmer, R. (2019). Artificial intelligence and counseling: Four levels of implementation. Theory & Psychology, 29(6), 807-819. https://doi.org./10.1177/0959354319853045
[9] Grinter, R., & Eldridge, M. (2003, April). Wan2tlk? Everyday text messaging. In Proceedings of the SIGCHI conference on Human factors in computing systems, 441-448. https://doi.org/10.1145/642611.642688
[10] Holden, C. (1977). The empathic computer. Science, 198(4312), 32-32. https://doi.org/10.1126/science.198.4312.32
[11] Hřebíčková, M., Jelínek, M., Květon, P., Benkovič, A., Botek, M., Sudzina, F., ... & John, O. P. (2020). BIG FIVE INVENTORY 2 (BFI-2): HIERARCHICKÝ MODEL S 15 SUBŠKÁLAMI. Ceskoslovenska Psychologie, 64(4). http://hdl.handle.net/11104/0310392
[12] Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR mHealth and uHealth, 6(11). https://doi.org/10.2196/12106
[13] Johnson, D. (2021). 20 Best AI Chatbots (Artificial Intelligence Chatbot) in 2021. Guru99. https://www.guru99.com/best-ai-chatbots.html?fbclid=IwAR1Wk47Bowym-jdu-zRn_rD58awnyF9uLoeszm6_UIkABAqF7kWGGwj_370
[14] Kretzschmar, K., Tyroll, H., Pavarini, G., Manzini, A., Singh, I., & NeurOx Young People's Advisory Group. (2019). Can your phone be your therapist? Young people's ethical perspectives on the use of fully automated conversational agents (chatbots) in mental health support. Biomedical informatics insights, 11. https://doi.org/10.1177/1178222619829083
[15] Křivohlavý, J. (2009). Psychologie zdraví. Portál.
[16] Loveys, K., Fricchione, G., Kolappa, K., Sagar, M., & Broadbent, E. (2019). Reducing patient loneliness with artificial agents: design insights from evolutionary neuropsychiatry. Journal of medical Internet research, 21(7). https://doi.org/10.2196/13664
[17] Nesmith, A. (2018). Reaching young people through texting-based crisis counseling: Process, benefits, and challenges. Advances in Social Work, 18(4), 1147-1164. https://doi.org/10.18060/21590
[18] Nešpor, K. (2018). Návykové chování a závislost: současné poznatky a perspektivy léčby. Portál.
[19] Pondělíček, I. (2016). Labyrinty duše & bída psychologie: výbor esejů. Prostor.
[20] Radziwil, N. M., & Benton, M. C. (2017). Evaluating Quality of Chatbots and Intelligent Conversational Agents. Cornell University. Radziwil & Benton.
[21] Santos, K. A., Ong, E., & Resurreccion, R. (2020, June). Therapist vibe: children's expressions of their emotions through storytelling with a chatbot. In Proceedings of the Interaction Design and Children Conference, 483-494. https://doi.org/10.1145/3392063.3394405
[22] Shah, H., Warwick, K., Vallverdú, J., & Wu, D. (2016). Can machines talk? Comparison of Eliza with modern dialogue systems. Computers in Human Behavior, 58, 278-295. https://doi.org/10.1016/j.chb.2016.01.004
[23] Silva, J. A. M. D., Siegmund, G., & Bredemeier, J. (2015). Crisis interventions in online psychological counseling. Trends in psychiatry and psychotherapy, 37, 171-182. https://doi.org/10.1590/2237-6089-2014-0026
[24] Svoboda, J. (2012). Poradenský dialog: vedení poradenského rozhovoru a poradenské skupiny. Triton.
[25] Ševčíková, A. (2014). Děti a dospívající online: vybraná rizika používání internetu. Grada.
[26] Špatenková, N. (2017). Krize a krizová intervence. Grada.
[27] Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H. & Loggarakis, A. (2020). User experiences of social support from companion chatbots in everyday contexts: Thematic analysis. Journal of medical Internet research, 22(3). https://doi.org/10.2196/16235
[28] Thomas, V., Balzer Carr, B., Azmitia, M., & Whittaker, S. (2020, April 9). Alone and Online: Understanding the Relationships Between Social Media, Solitude, and Psychological Adjustment. Psychology of Popular Media. Advance online publication. http://dx.doi.org/10.1037/ppm0000287
[29] Tondl, L. (2014). Rozmluva a usuzování. Pavel Mervart.
[30] Tvrdý, F. (2014). Turingův test: filozofické aspekty umělé inteligence. Togga.
[31] Vodáčková, D. (2012). Krizová intervence. Portál.

Generatívne vlastnosti modelu UBAL

Kristína Malinovská a Igor Farkaš

Centrum pre kognitívnu vedu, KAI FMFI UK,
Univerzita Komenského v Bratislave
Mlynská dolina, 84248 Bratislava
Email: malinovska@fmph.uniba.sk

Abstrakt

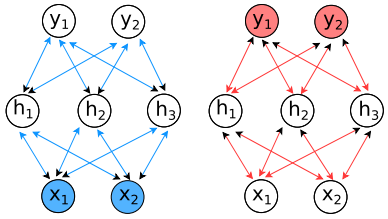
Dominantnou metódou učenia NS je algoritmus spätného šírenia chyby, ktorý je len vzdialene príbuzný mechanizmom učenia v mozgu. Náš univerzálny biologicky-motivovaný algoritmus s lokálnym pravidlom - UBAL - predstavuje biologicky plauzibilnejšiu alternatívu pre trénovanie neurónových sietí a má mnoho zaujímavých vlastností. Jedna z nich je, že je pri konkrétnom nastavení jeho hyperparametrov schopný pri klasifikačnej úlohe generovať obrazy vstupov, ktoré dostáva. Keďže cieľom trénovania nie je generovanie vzorov, môžeme hovoriť o emergentnom jave. V príspevku prezentujeme výsledky experimentov s datasetom MNIST. Načrtneme vlastnosti a využitie vzniknutých projekcii, napríklad v doméne adverzariálnych príkladov a vysvetliteľnej umelej inteligencie (XAI).

1 Úvod

V dnešnej dobe sú umelé neurónové siete (NS) asi najpopulárnejším nástrojom strojového učenia a to hlavne vďaka hlbokému učeniu. Väčšina hlbokých modelov je trénovaná niektorou adaptáciou spätného šírenia chyby (error backpropagation, BP, Rumelhart a spol., 1986) navrhnutého pre viacvrstvový perceptrón (multilayer percetron, MLP). Je to pomerne efektívny algoritmus, avšak, ako poukázal Crick (1989) len pár rokov po jeho vzniku, od biologického mechanizmu neurálneho učenia má ďaleko. Pravidlo učenia spätným šírením chyby využíva pre adaptáciu váh gradienty chyby šírenej od výstupnej vrstvy na vstupnú, čo znamená, že každý neurón musí poznať celú dráhu spojení, ktoré k nemu vedú od vstupu. Táto vlastnosť bola kritizovaná ako takzvaný problém transportu váh (weight transport problem; Grossberg, 1987).

Adaptácia váh biologických neurónov prebieha lokálne, teda na základe aktivity presynaptického a postsynaptického neurónu. Ďalej možno poukázať na to, že v mozgu neexistujú dráhy, po ktorých by sa signál šíril oboma smermi cez tie isté synaptické váhy, resp. spojenia. Ak uvažujeme jeden základný mechanizmus synaptickej plasticity v mozgu, potom z pohľadu paradigiem učenia sa NS ide o kombináciu učenia sa bez učiteľa a s učiteľom (O'Reilly a spol., 2012).

UBAL, univerzálny biologicky-motivovaný algoritmus s lokálnym pravidlom (Universal Bidirectional Activation-based Learning) predstavuje novú biologicky motovivanú alternatívu ku klasickému učeniu spätným šírením chyby. Náš model vznikol z predchodcu BAL (Farkaš a Rebrová, 2013) a historicky sme ho prvý krát prezentovali na konferencii Kognícia a umelý život XVII (Malinovská a spol., 2018). UBAL patrí do rodiny modelov založených na kontrastívnom Hebbovom učení, ako je napríklad Contrastive Hebbian Learning (CHL, Movellan, 1991) a GeneRec (O'Reilly, 1996). Navyše je inšpirovaný aj klasickým autoenkóderom, ktorý vytvoril Hinton a McClelland (1988).



Obr. 1: Mínusová (vľavo) a plusová (vpravo) fáza v kontrastívnom učení.

Na rozdiel od BP sa v týchto lokálne učiacich sa modeloch šíri aktivácia, nie samotná chyba. Klasické dopredné šírenie aktivácie zo vstupnej vrstvy na výstupnú, ktoré vytvorí predikciu siete na výstupe, sa v tomto učení nazýva mínusová fáza. Po nej namiesto šírenia chyby nastáva fáza plusová, v ktorej sa na výstupnej vrstve zafixuje (clamp) cieľová aktivácia (target resp. label) a propaguje sa smerom naspäť na vstupnú vrstvu. Plusovú a mínusovú fázu ilustrujeme na Obr. 1. Vo vyššie spomenutých modeloch sa to robí cez transponovanú maticu váh podobne ako pri BP s rozdielom, že ide o produkt aktivácie a nie chyby. Každá váha je nakoniec upravená na základe rozdielu aktivácii presynaptického neurónu p a postsynaptického neurónu q v mínusovej a plusovej fáze, pre CHL podľa rovnice

$$\Delta \boldsymbol{W}_{pq} = \lambda \left[(\boldsymbol{p}^{+}\boldsymbol{q}^{+}) - (\boldsymbol{p}^{-}\boldsymbol{q}^{-}) \right] \tag{1}$$

a pre Generec podľa

$$\Delta \boldsymbol{W}_{pq} = \lambda \, \boldsymbol{p}^{-} (\boldsymbol{q}^{+} - \boldsymbol{q}^{-}). \tag{2}$$

Biologicky plauzibilnému učeniu klasických neurónových sietí sa dostáva pozornosť aj v doméne hlbokého učenia. Napríklad vo Feedback Alignment model z dielne Lillicrap a spol. (2016) sa chyba nešíri spätne cez synaptické váhy, ale cez maticu náhodne výgenerovaných váh „správneho tvaru". Podľa autorov práve tvar a rozsah náhodných hodnôt, z ktorých sú tieto matice pseudováh generované stačí na prenos informácie a naučenie modelu. Lokálne pravidlo učenia a metódu fixovania cieľov nájdeme v modeloch ako je Equilibrium propagation (Scellier a Bengio, 2017) alebo Target propagation (Ororbia a Mali, 2018) a modeloch od nich odvodených. Používa sa tu tzv. soft clamping, pri ktorom sa na výstupe zafixuje želaná aktivácia iba čiastočne - proporčne. Nižšie uvádzame takúto možnosť aj v našom modeli a vysvetľujeme ako kontrolujeme silu alebo proporciu vplyvu cieľovej aktivácie na učenie modelu.

2 Náš model

UBAL je v princípe heteroasociátor, čiže predstavuje mapovanie medzi vstupno-výstupnými vzormi $\boldsymbol{x}$ a $\boldsymbol{y}$. Z nášho uhla pohľadu je teda aj klasifikácia formou asociácie množstva vstupných vzorov s ich kategóriou či triedou. Výnimočnosť nášho modelu spočíva v tom, že je tvorený dvoma maticami váh medzi každou vrstvou neurónov, a to $\boldsymbol{W}$ pre dopredný smer (FP) a $\boldsymbol{M}$ pre spätný smer (BP). Aktivácia sa cez matice váh šíri vždy len v jednom smere.

2.1 Propagácia aktivácie

V kontexte porovnania s klasickým MLP pre problém klasifikácie budeme označovať smery propagácie aktivácie v našom modeli nasledovne:

- F (forward) - dopredná aktivácia, čiže $\boldsymbol{x} \rightarrow \boldsymbol{y}$
- B (backward) - spätná aktivácia, čiže $\boldsymbol{y} \rightarrow \boldsymbol{x}$

Vo vyššie spomenutých modeloch, ktoré inšpirovali náš model, možno identifikovať doprednú aktiváciu s mínusovou a spätnú s plusovou.

Ďalším špecifikom nášho modelu je, že okrem klasického šírenia aktivácie naprieč celým modelom, čo nazývame *predikcia* v oboch smeroch (F a B), propagujeme aktiváciu aj bezprostredne z predikovaných hodnôt sieťe späť do presynaptickej vrstvy. Tento komponent nášho modelu nazývame *echo*. Je inšpirovaný takzvanou regresiou z klasického autoenkódera (Hinton a McClelland, 1988) a predstavuje spätnú väzbu. *Echo* aktivácia sa vždy šíri od postsynaptického neurónu naspäť k presynaptickému, čiže v opačnom smere a teda sa šíri cez druhú z páru matíc váh. Napríklad *echo* z doprednéhо smeru na skrytej vrstve $\boldsymbol{h}$ sa „odrazí" cez váhy $\boldsymbol{M}$ na vstupnú vrstvu $\boldsymbol{x}$.

Fáza	Zn.	Výraz	Výpočet aktivácie
Forward Prediction	FP	$\boldsymbol{q}^{\mathrm{FP}}$	$f(\boldsymbol{W}_{pq}\boldsymbol{p}^{\mathrm{FP}} + \boldsymbol{b}_p)$
Forward Echo	FE	$\boldsymbol{p}^{\mathrm{FE}}$	$f(\boldsymbol{M}_{qp}\boldsymbol{q}^{\mathrm{FP}} + \boldsymbol{d}_q)$
Backward Prediction	BP	$\boldsymbol{p}^{\mathrm{BP}}$	$f(\boldsymbol{M}_{qp}\boldsymbol{q}^{\mathrm{BP}} + \boldsymbol{d}_q)$
Backward Echo	BE	$\boldsymbol{q}^{\mathrm{BE}}$	$f(\boldsymbol{W}_{pq}\boldsymbol{p}^{\mathrm{BP}} + \boldsymbol{b}_p)$

Tab. 1: Šírenie aktivácie medzi dvoma vrstvami p a q prepojenými váhami W a M.

Šírenie aktivácie a názvy aktivačných fáz definujeme v Tabuľke 1. Keďže náš model môže mať rôzny počet vrstiev, uvádzame rovnice pre propagáciu aktivácie medzi dvoma prepojenými vrstvami neurónov p a q, ktoré spájajú matice synaptických váh W a M. Tak, ako vo väčšine modelov NS, aj my používame učiace sa prahy neurónov a označujeme ich $\boldsymbol{b}_p$ v doprednom (F) a $\boldsymbol{d}_q$ v spätnom (B) smere. Pre ilustráciu našich súčasných experimentov zobrazujeme na Obr. 2 šírenie aktivácie pre trojvrstvovú sieť.

2.2 Adaptácia váh

Pravidlo učenia v našom modeli je inšpirované hlavne modelom GeneRec. Rozšírili sme ho na naše dvojice matíc váh a dva smery propagácie. Zároveň sme pravidlo obohatili o špeciálne hyperparametre β a γ, ktoré určujú proporcie sily predikcie a klasického učenia a vnútorného *echa*. Pre zjednodušenie zápisu definujeme výpočtové členy vyššej úrovne, ktoré nazývame cieľ (target, t) a odhad (estimate, e), ktoré možno chápať ako komponenty na klasického učenia sa s učiteľom a samoorganizovaného (self-supervised) učenia sa.

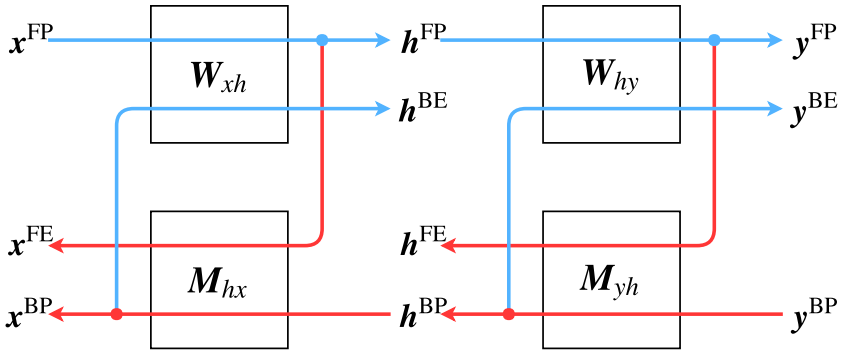
V našom modeli definujeme nové hyperparametre, ktoré určujú dôležitosť alebo silu aktivačných hodnôt z rôznych fáz pre úpravu váh. Hyperparametre β určujú silu akou je cieľová aktivácia fixovaná na výstupnej vrstve vo fáze kedy propagujeme tento (z pohľadu klasifikácie) cieľový stav a ako *predikcia* ovplyvňuje zmenu váh. Tento hyperparameter sa viaže na aktivačné hodnoty v neurónoch. Hyperparametre γ určujú silu resp. proporciu toho, ako sa na učení podieľajú *predikcia* a *echo* a viaže sa na matice váh $\boldsymbol{W}$ a $\boldsymbol{M}$ ako ilustrujeme na Obr. 3.

Na základe hodnôt hyperparametrov β a γ je náš model schopný učiť sa najrôznejšie úlohy od asociácie a odšumenia až po klasifikáciu (Malinovská a spol., 2019). Výpočet výrazov t a e uvádzame v tabuľke Tab. 2 a graficky zobrazujeme na Obr. 3. Použitím výrazov t a e dostávame pravidlo pre úpravu váh W

$$\Delta \boldsymbol{W}_{pq} = \lambda\, \boldsymbol{t}_p^{\mathrm{B}}(\boldsymbol{t}_q^{\mathrm{F}} - \boldsymbol{e}_q^{\mathrm{F}}) \tag{3}$$

a pre váhy M

$$\Delta \boldsymbol{M}_{qp} = \lambda\, \boldsymbol{t}_q^{\mathrm{F}}(\boldsymbol{t}_p^{\mathrm{B}} - \boldsymbol{e}_p^{\mathrm{B}}), \tag{4}$$



Obr. 2: Šírenie aktivácie v UBAL v sieti s jednou skrytou vrstvou. Modrou zobrazujeme dopredný smer (FP) a červenou spätný (BP).

ktoré svojím tvarom kopíruje kontrastívne učenie z predošlej kapitoly, viď rovnice (1) a (2).

Člen	Zn.	Výpočet
Forward Target	$\boldsymbol{t}_q^{\mathrm{F}}$	$\beta_q^{\mathrm{F}}\boldsymbol{q}^{\mathrm{FP}} + \beta_q^{\mathrm{B}}\boldsymbol{q}^{\mathrm{BP}}$
Forward Estimate	$\boldsymbol{e}_q^{\mathrm{F}}$	$\gamma_q^{\mathrm{F}}\boldsymbol{q}^{\mathrm{FP}} + (1-\gamma_q^{\mathrm{F}})\boldsymbol{q}^{\mathrm{BE}}$
Backward Target	$\boldsymbol{t}_p^{\mathrm{B}}$	$\beta_p^{\mathrm{B}}\boldsymbol{p}^{\mathrm{BP}} + \beta_p^{\mathrm{F}}\boldsymbol{p}^{\mathrm{FP}}$
Backward Estimate	$\boldsymbol{e}_p^{\mathrm{B}}$	$\gamma_p^{\mathrm{B}}\boldsymbol{p}^{\mathrm{BP}} + (1-\gamma_p^{\mathrm{B}})\boldsymbol{p}^{\mathrm{FE}}$

Tab. 2: Členy v rovniciach učenia. Na každej vrstve platí $\beta^B + \beta^F = 1$, ale γ_q^F a γ_p^B sú nezávislé.

3 Klasifikácia a generatívne vlastnosti

Náš model sme testovali aj na najznámejšom klasifikačnom benchmarku - datasete písaných číslic MNIST LeCun a spol. (1998). Výsledné nastavenie hyperparametrov β a γ zobrazujeme v Tab. 3, možno ho čítať ako hodnoty pre jednotlivé vrstvy v poradí vstupná–skrytá–výstupná. Pri týchto dvoch rôznych nastaveniach, ktoré nazývame model A a model B a dostatočne veľkosti skrytej vrstvy (napr. 1500 neurónov) dosahuje UBAL klasifikačnú úspešnosť na testovacej sade okolo 96%, čo je porovnateľné s príbuznými modelmi.

Chceme zdôrazniť, že sme nepoužili žiadne augmentačné ani regularizačné techniky. Používame 3-vrstvovú sieť so štandardnými sigmoidálnymi neurónmi a softmax pre výstupnú vrstvu, Gaussovskú inicializáciou váh s distribúciou $\mathcal{N}(0; 0, 5)$ a rýchlosť učenia $0,1$. Na výstupe reprezentujeme číslice ako one-hot vektory a hodnoty pixelov vstupných obrázkov škálujeme do intervalu $(0, 1)$.

Pri našich experimentoch s klasifikáciou písaných číslich sme si všimli, že v niektorých prípadoch možno na vstupnej vrstve v spätnom smere pozorovať obrázky, ktoré pripomínajú číslice zadané na výstupe. Keďže sme sieť trénovali na klasifikáciu a nie generovanie, považujeme to za emergentný jav. Intuitívne vysvetlenie spočíva v podstate nášho modelu. Je to heteroasociátor

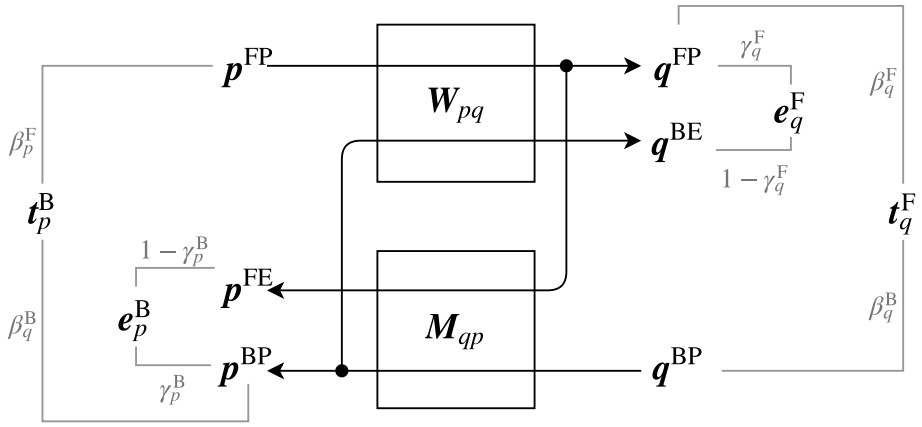
	Model A	Model B
$\boldsymbol{\beta}^{\mathrm{F}}$	0,0 - 1,0 - 0,0	1,0 - 1,0 - 0,9
$\boldsymbol{\gamma}^{\mathrm{F}}$	1,0 - 1,0	1,0 - 1,0
$\boldsymbol{\gamma}^{\mathrm{B}}$	1,0 - 1,0	0,9 - 1,0
$\boldsymbol{\beta}^{\mathrm{B}}$	1,0 - 0,0 - 1,0	0,0 - 0,0 - 0,1

Tab. 3: Parametre β a γ pre klasifikáciu písaných číslic MNIST v dvoch rôznych variantoch.

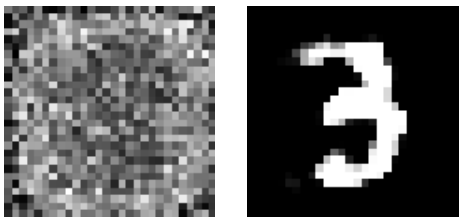
a teda aj klasifikácia je vlastne asociácia vzorov a reprezentácií kategórií. Veľmi zaujímavé na tomto fenoméne je aj to, že k nemu dochádza len pri jednom z dvoch odlišných nastavení hyperparametrov β a γ ako ilustrujeme na Obr. 5. A teda iba v modeli B dochádza k takej „predstavivosti", ktorá vytvára zmysluplné obrazy vstupných dát.

Príklad vygenerovaných číslic z jednej siete zobrazujeme na Obr. 5. Z našich experimentov môžeme tiež povedať, že tieto obrázky sa líšia naprieč náhodne incializovanými sieťami a sú tiež odlišné od vypočítaných priemerov všetkých obrázkov datasete MNIST ako ukazujeme na Obr. 6. V mnohých prípadoch sa zdá, že UBAL je schopný správne klasifikovať svoje výtvory práve vtedy, keď sú čitateľné aj pre človeka. Prirodzeným krokom vo skúmaní týchto vzniknutých obrázkov je ich vyhodnotenie pomocou štandardnej miery pre generovanie obrazov akou je napríklad inception score (Salimans a spol., 2016). Podobne je potrebné overiť ako budú vzniknuté obrázky klasifikovať iné, bežné siete natrénované na dátovej sade MNIST.

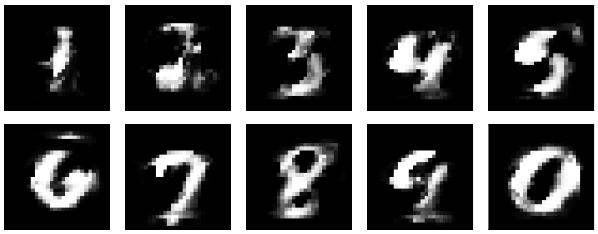
Pri manipulácii s hyperparametrami modelu B sa zdá, že keď znížime β^{F} pre skrytú vrstvu z hodnoty $1,0$ na menšiu hodnotu $(0,995 - 0,999999)$ klasifikačnú úspešnosť siete sa zníži, no vzniknuté obrazy sú variabilnejšie, „krajšie" a majú menej ostré hrany. Vzťah medzi parametrami β a vygenerovanými obrazmi naďalej skúmame. Zároveň testujeme pridávanie malého Gaussovského šumu do one-hot reprezentácii číslic na výstupe. Náhodný šum je dôležitou súčasťou trénovania



Obr. 3: Schematické zobrazenie dvoch prepojených vrstiev p a q a výpočet výrazov pre učiace pravidlo pomocou β a γ.

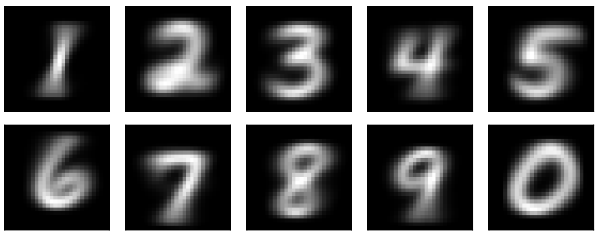


Obr. 4: Príklady spätnej aktivácie modelu A a modelu B pre číslicu 3.



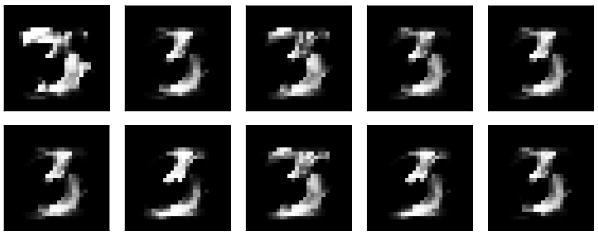
Obr. 5: Obrazy číslic z dátovej sady MNIST generované modelom UBAL.

aj pri známych Generatívnych adverzariálnych sieťach (Goodfellow a spol., 2014). Predbežné výsledky našich experimentov so šumom ukazujú, že schopnosť nášho modelu zovšeobecňovať sa s pridaným šumom nestráca a zároveň umožňuje získavať variabilitu vo vygenerovaných obrázkoch. Na Obr. 7 ukazujeme rôzne inštancie číslice 3, ktoré sme získali pridaním malého náhodného šumu z Gaussovskej distribúcie



Obr. 6: Priemery číslic z dátovej sady MNIST.

$\mathcal{N}(0; 0, 005)$ do cieľovej reprezentácie číslice, ale nepoužívali sme ho počas trénovania.



Obr. 7: Inštancie obrazu číslice 3 po pridaní malého Gaussovského šumu.

Schopnosť modelu UBAL generovať vzory v spätnom smere je najviac ovplyvnená nastavením hyperparametrov β a γ. Existujú nastavenia, ktoré síce fungujú dobre z hľadiska klasifikačnej úspešnosti, ale neumožňujú generovať čitateľné čísla ako vidno pri modeli A na Obr. 4. Súvislosť s nastavením parametrov nie je náhodná a vzťah týchto parametrov, resp. formy učenia, ktorú tieto parametre predstavujú je potrebné ďalej skúmať.

Analýza šumu, ktorý produkuje model A a to ako by na takéto obrázky reagovali iné siete, je ďalším z možných smerov výskumu. Jednou z našich hypotéz pre overenie do budúcna je aj to, či šum a vzory, ktoré UBAL vyrába nemožno použiť na vyrábanie takzvaných adverzariálnych príkladov (Goodfellow a spol., 2014). Adverzariálne príklady sú obrázky, do ktorých pridaný špeciálne vyrobený šum spôsobí, že sieť nesprávne rozozná kategóriu na obrázku. Táto technika sa používa na poli takzvanej vysvetliteľnej umelej inteligencie (XAI) s cieľom porozumieť, akým spôsobom vznikajú reprezentácie v sieti, čiže čo sieť „sleduje" a ako to možno vylepšiť a predísť tomu, aby sa NS učili len „povrchové" závislosti. Môžme to chápať aj ako útok na bezpečnosť siete. Väčšinou sa tieto obrázky získavajú pomocou gradientov chýb so siete. Ak by sme vedeli využiť spätné projekcie nášho modelu na generovanie týchto obrázkov získame techniku, ktorá nie je

vôbec závislá na architektúre a vlastnostiach siete, na ktorú útok vykonávame. V tejto línii plánujeme zároveň skúmať odolnosť modelu UBAL voči adverzariálnym útokom.

4 Záver

Predstavili sme Univerzálny biologicky-motivovaný algoritmus s lokálnym pravidlom - UBAL a jeho generatívne vlastnosti, ktoré čiastočne vyplývajú z jeho heteroasociatívnej podstaty. Uvádzame nastavenia parametrov pre klasifikačnú úlohu rozoznávania písaných číslic v ktorých dáva náš model dostatočnú klasifikačnú úspešnosť, pričom v jednom z nich vznikajú pri spätnom smere šírenia aktivácie vzory, ktoré možno považovať za obrazy alebo predstavy číslic, ktoré sa model učil rozpoznávať. Tento jav považujeme za emergentný, keďže sieť nebola trénovaná na generovanie ale na klasifikáciu. Diskutujeme rôzne možnosti rozvíjania týchto generatívnych vlastností v našom modeli ako aj smery skúmania a využitia takýchto obrazov.

Poďakovanie

Za podporu ďakujeme Slovenskej spoločnosti pre kognitívnu vedu[1].

Literatúra

Crick, F. (1989). The recent excitement about neural networks. *Nature*, 337(6203):129–132.

Farkaš, I. a Rebrová, K. (2013). Bidirectional activation-based neural network learning algorithm. V *Proceedings of the International Conference on Artificial Neural Networks (ICANN), Sofia, Bulgaria*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. a Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, 11(1):23–63.

Hinton, G. E. a McClelland, J. L. (1988). Learning representations by recirculation. V *Neural Information Processing Systems*, str. 358–366.

LeCun, Y., Bottou, L., Bengio, Y. a Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lillicrap, T. P., Cownden, D., Tweed, D. B. a Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, 7:13276.

Malinovská, K., Malinovský, Ľ. a Farkaš, I. (2018). UBAL: Univerzálny biologicky-motivovaný algoritmus s lokálnym pravidlom. et al., Š. (zost.), V *Kognice a umělý život XVIII*, str. 50–52, Brno. Flow, z.s.

Malinovská, K., Malinovský, Ľ., Krsek, P., Kraus, S. a Farkaš, I. (2019). UBAL: a Universal Bidirectional Activation-based Learning Rule for Neural Networks. V *Proceedings of the 2019 International Conference on Computational Intelligence and Intelligent Systems (CIIS 2019)*, New York, NY, USA. Association for Computing Machinery.

Movellan, J. R. (1991). Contrastive hebbian learning in the continuous hopfield model. V *Connectionist Models*, str. 10–17. Elsevier.

O'Reilly, R. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation*, 8(5):895–938.

O'Reilly, R. C., Munakata, Y., Frank, M., Hazy, T. a spol. (2012). *Computational cognitive neuroscience*. PediaPress.

Ororbia, A. G. a Mali, A. (2018). Biologically motivated algorithms for propagating local target representations. *arXiv preprint arXiv:1805.11703*.

Rumelhart, D., Hinton, G. a Williams, R. (1986). *Learning internal representations by error propagation*, str. 318–362. No. 1. The MIT Press, Cambridge, MA.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A. a Chen, X. (2016). Improved techniques for training gans. *Advances in Neural Information Processing Systems*, 29.

Scellier, B. a Bengio, Y. (2017). Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in Computational Neuroscience*, 11:24.

[1] `https://cogsci.fmph.uniba.sk/sskv/`

Neurónová sieť s násobiacou vrstvou

Ľudovít Malinovský a Kristína Malinovská

Katedra aplikovanej informatiky, FMFI,
Univerzita Komenského v Bratislave
Mlynská dolina, 84248 Bratislava
Email: malinovska@fmph.uniba.sk

Abstrakt

Problém XOR je základný testovací problém pre nové modely neurónových sietí obsahujúci nelinearitu, podobne aj problém parity. Pre samotnú sieť ide o ťažký problém, špeciálne pri malom počte takzvaných skrytých neurónov. Násobenie je stále pomerne zriedkavým elementom v architektúre neurónových sietí. Vo všeobecnom prípade vedie ku komplexným číslam, ktoré komplikujú návrh metódy učenia. V príspevku predstavíme nový návrh neštandardnej neurónovej siete s učiacou sa násobiacou vrstvou. Pravidlo učenia multiplikačných váh je odvodené od klasického spätného šírenia chyby a nevedie ku komplexným číslam. Ukážeme predbežné výsledky pre XOR a paritu v porovnaní s klasickým viacvrstvovým perceptrónom, z ktorých vyplýva, že náš model má bezkonkurenčne vysokú úspešnosť a veľký potenciál do budúcnosti.

1 Úvod

Násobenie aktivácií neurónov je teoreticky preskúmaná záležitosť, ale v praxi je pomerne zriedkavo využívaná. V jednoduchšom prípade neurón podobný klasickému perceptrónu počíta váženú sumu členov polynómu vyššieho rádu a učia sa iba sumačné váhy. Keďže s rastúcim stupňom polynómu exponenciálne rastie počet možných členov, zloženie polynómu je možné limitovať rôznymi spôsobmi. V zložitejšom prípade takzvaných produktových neurónov sa sieť učí exponenty jednotlivých zložiek každého členu. To však vo všeobecnom prípade, kde môže dôjsť k umocneniu záporného čísla na zlomok, vedie k výpočtom v doméne komplexných čísel. Výsledkom je, že násobenie sa používa len vo vybraných prípadoch ako optimalizácia na dopredu zvolených presných miestach modelu. Samotná sieť sa väčšinou potrebu použiť násobenie neučí. Napriek týmto prekážkam, teoretické aj praktické výsledky svedčia v prospech využitia násobenia ako nástroja na zvýšenie výpočtovej sily neurónových sietí.

V tomto príspevku predstavujeme inovatívny model dvojvrstvovej neurónovej siete, ktorá využíva násobenie na výstupnej vrstve. Váhy na výstupnej vrstve sa aplikujú spôsobom, ktorý umožňuje učenie pomocou metódy gradientového zostupu bez potreby výpočtov v doméne komplexných čísel. Tak je sieť schopná sa čiastočne naučiť zloženie polynómu, čo zvyšuje všeobecnosť modelu. Násobiaca vrstva je navrhnutá takým spôsobom, že je možné zapojiť ju do architektúry rovnako ľubovoľne a jednoducho, ako klasickú sigmoidálnu sumačnú vrstvu. V experimentoch ukazujeme neporovnateľne vyššiu efektívnosť takejto siete v logických úlohách XOR a parita oproti klasickému dvojvrstvovému perceptrónu s rovnakou architektúrou a parametrami.

2 Násobenie v neurálnych modeloch

Násobenie sa v neurálnych modeloch využíva od konca 50-tych rokov a v 60-tych rokoch bolo populárne používať sigmoidálne neuróny vyššieho rádu pre rozpoznávanie vzorov (Nilsson, 1965). Išlo o klasické sigmoidálne neuróny, avšak vstupný vektor bol rozšírený o členy, ktoré vznikli vynásobením a umocnením vstupov, pričom tieto členy dostali vlastné váhy. Tento prístup je prirodzeným rozšírením klasického perceptrónu, avšak vo všeobecnom prípade vedie k exponenciálnemu nárastu dimenzionality, pričom horné ohraničenie stupňa polynómu je hlavne dizajnérskou voľbou. V takýchto sieťach možno jednoducho aplikovať spätné šírenie chyby a neuróny tohto typu sú známe aj ako *sigma-pí* neuróny (Rumelhart a spol., 1986), keďže vykonávajú váženú sumu súčinov vstupov.

Riešením problému dimenzionality bola rodina modelov, v ktorých bol počet polynomiálnych členov obmedzený na niektoré dopredu vybrané. Napríklad Giles a Maxwell (1987) a Spirkovska a Reid (1994) ukazujú, že takýmto spôsobom možno výrazne prekonať klasické perceptrónové siete v rýchlosti učenia aj schopnosti generalizácie. Tento prístup však vyžaduje, aby určitú časť znalosti o probléme vložil do modelu jeho dizajnér. Reakciou na to boli konštruktívne modely. Ich podstatou bolo, že sa rôznymi spôsobmi učili, ktoré polynomiálne členy sú potrebné, a to tak, že začali s minimálnym modelom a postupne pridávali členy vyššieho rádu a vyhodnocovali ich efektívnosť (Redding a spol., 1993; Heywood a Noakes, 1995).

Zovšeobecnením vyššie uvedeného princípu je model s produktovým neurónom (Durbin a Rumelhart, 1989), ktorý predpokladá reálne, učiace sa, váhy vy-

stupujúce ako exponenty vstupov. Problém so vstupom do domény komplexných čísel autori vyriešili tým, že použili iba reálnu časť čísla. Pre logické vstupy sa takýto neurón stane de facto sumačným neurónom ktorý má ako aktivačnú funkciu kosínus, čo vedie k ďalším teoretickým problémom (Anthony a Bartlett, 1999).

Alternatívny prístup navrhli Ghosh a Shin (1992). Ich model najskôr lineárne kombinuje vstupy a následne túto linerárnu kombináciu priamo násobí bez aktivačnej funkcie. Nelineárna aktivačná funkcia sa aplikuje až na výstupe. Jedinými pramaterami modelu sú lineárne váhy a váhy na násobiacej vrstve sú fixované na hodnotu jedna. Výsledok možno analyzovať ako polynóm. Keďže vo vzorci vystupuje súčin súčtov, hovorí sa o pí-sigma sieti. Keďže váhy na násobiacej vrstve sú fixované a učia sa len lineárne váhy, algortimus učenia je založený na gradientovej metóde.

Autori testovali úspešnosť siete na mnohých bežných úlohách akými sú aproximácia funkcie, klasifikácia dát, rozpoznávanie znakov s translačnou a rotačnou invarianciou a tiež overovali schopnosť siete naučiť sa logické funkcie parity, symetrie a negácie. Dosiahnutá úspešnosť je v prípade parity v porovnaní s klasickým viacvrstvovým perceptrónom rádovo lepšia a porovnateľná s inými modelmi s neurónmi vyššieho rádu.

Z nášho pohľadu je nevýhodou, že Ghosh a Shin (1992) multiplikačné váhy fixujú a ani sa nepokúšajú nájsť spôsob, ako ich učiť. Absencia aktivačnej funkcie po lineárnej transformácii navyše redukuje pí-sigma sieť na polynomiálny model s dopredu vybranými členmi. Shin a Ghosh (1995) navrhujú riešenie konštruktívnym modelom, podobným konštruktívnym sigma-pí modelom vyššie.

V biologických neurónových sieťach sa vyskytujú analógy násobenia v rôznych podobách. Jednou z možností sú dendritické zhluky v dendritických stromoch (Mel, 1994), ktoré vedia aj deliť, prípadne by mohli reprezentovať operáciu podobnú sigma-pí neurónom (Mel a Koch, 1989). Schopnosť neurónov násobiť vstupy bola nájdená v sluchovej (Suga a spol., 1990) a zrakovej (Andersen a spol., 1985) kôre mozgu. Niektorí autori argumentujú, že fyzická blízkosť synáps vedie k multiplikačnému správaniu, zatiaľ čo vzdialené synapsy fungujú v sumačnom móde (Bugmann, 1991).

Schmitt (2002) analyzuje rôzne typy neurónových sietí s násobiacimi a sigmoidálnymi neurónmi a odvodzuje všeobecné matematické obmedzenia pre jednotlivé typy sietí v zmysle Vapnik-Červonenkisovej dimenzie. Dochádza k záveru, že násobenie v neurónových sieťach je vhodnou voľbou pre zvýšenie miery nelineárnej interakcie a výpočtovej sily a to aj pre neuróny s vysokým rádom. Článok poskytuje tiež kvalitný prehľad neurálnych modelov s násobením, ktorý je stále pomerne aktuálny napriek súčasnému rozmachu hlbokých sietí s veľkým počtom parametrov a využitiu hrubej sily grafických kariet.

V kontexte hlbokých neurónových sietí, bolo násobenie súčtov skúmané v sum-product sieťach (Poon a Domingos, 2011; Delalleau a Bengio, 2011). Niektoré nové modely využívajú násobenie vektorov a matíc aktivácií po prvkoch na vybraných miestach modelu (Erhan a spol., 2009; Zhu a spol., 2018; Diba a spol., 2017; Schenck a Fox, 2018). Na blogosfére možno nájsť články rôznej kvality, ktoré popisujú konkrétne využitie násobenia v hlbokých neurónových sieťach a ukazujú jeho výhody na konkrétnych prípadoch. Ide teda o stále aktívny a sľubný smer výskumu.

3 Náš model

Náš model sa podobá na pí-sigma sieť (Ghosh a Shin, 1992), s tým rozdielom, že na skrytej vrstve aplikujeme hyperbolický tangens ako aktivačnú funkciu, čím dostaneme skryté aktivácie z intervalu -1 až 1. Tie potom na výstupnej vrstve násobíme. Váhy v našom modeli nie sú fixné, ani nefigurujú ako exponenty, ale sú do výpočtu zapojené tak, aby v jednom z extrémov na danej aktivácii skrytej vrstvy nezáležalo (zredukuje sa na jednotku, akoby bola umocnená na 0) a v druhom extréme vystupovala v pôvodnej hodnote (bez zmeny, akoby bola umocnená na 1). Tento prepočet, ktorým nahrádzame exponenciáciu z predošlých modelov, je v skutočnosti polynomiálny. Táto vlastnosť nám zaručí, že sa nikde vo výpočte nevyskytnú komplexné čísla, a zároveň umožňuje modelu naučiť sa, ktoré skryté neuróny „má zmysel" použiť pre dosiahnutie požadovaného výstupu. Schematické zobrazenie modelu ukazujeme na Obr. 1.

V našom súčasnom výskume sme zvolili hyperbolický tangens preto, lebo reprezentácia logických hodnôt pravdy a nepravdy pomocou hodnôt -1 a $+1$ umožňuje priamočiaro nasimulovať logickú funkciu XOR pomocou násobenia, pretože násobenie znamienok sa správa presne ako XOR. Pre iné úlohy bude vhodné zvoliť inú aktivačnú funkciu.

3.1 Dopredný beh

Pre vstup $\boldsymbol{x}$, vrátane trénovateľného prahu (bias), a váhy w^{hid} a w^{out} počítame aktiváciu skrytých neurónov ako

$$h_i = \tanh(\sum_j w_{ij}^{\text{hid}} \cdot x_j), \tag{1}$$

a z toho aktiváciu výstupných neurónov ako

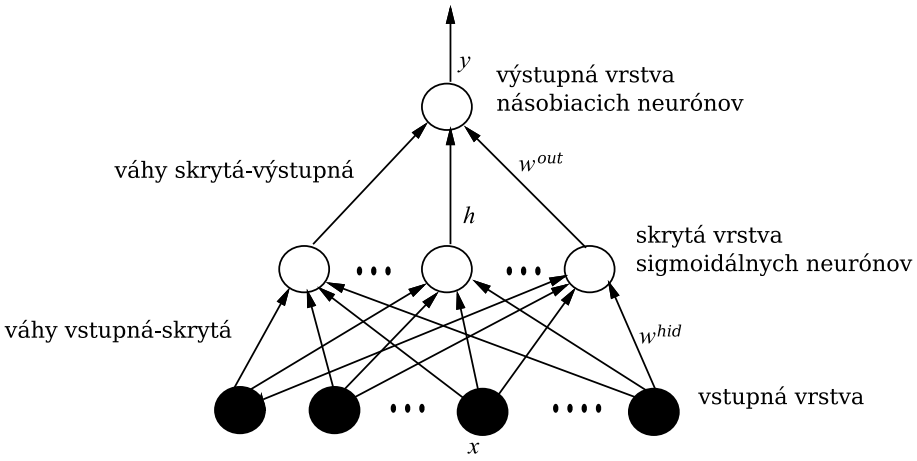
$$y_i = \prod_j \left(1 - \sigma(w_{ij}^{\text{out}})(1 - h_j)\right), \tag{2}$$

prípadne

$$y_i = \prod_j f(h_j, \sigma(w_{ij}^{\text{out}})), \tag{3}$$

kde funkciu f definujeme ako

$$f(h_j, d) = 1 - d(1 - h_j) \tag{4}$$



Obr. 1: Schéma nášho modelu adaptovaná od Ghosh a Shin (1992).

a σ je logistická funkcia.

Náš model teda na výstupnej vrstve používa analógiu produktového neurónu (Durbin a Rumelhart, 1989), ktorá však umocnenie vstupu na váhu nahrádza polynomiálnou funkciou f s jedným nelineárnym parametrom:

$$f(h, \sigma(w)) = 1 - \sigma(w)(1 - h) \tag{5}$$

Takýto prístup zachováva niektoré vlastnosti umocnenia, konkrétne $h^1 = f(h, 1) = h$ a $h^0 = f(h, 0) = 1$ avšak nezachováva vlastnosť $0^d = 0$. Namiesto toho v prípade nulového vstupu dostávame $f(0, d) = 1 - d$.

Náš spôsob aplikácie váhy pomocou funkcie f je navyše spojitý a spojito derivovateľný pre všetky nulové vstupné hodnoty, kým 0^0 je nedefinované a teda bod nespojitosti a bez derivácie. Aplikáciou logistickej funkcie na váhu docielime, aby váha mohla nadobúdať ľubovoľné reálne hodnoty, ktoré ale v extrémoch zodpovedajú exponentom 0 a 1. Vďaka tomu sa môžu výstupné váhy učiť bez obmedzení gradientovými metódami, ktoré by ich inak dostali mimo žiadúcich hodnôt.

3.2 Spätné šírenie chyby

Pre výstupnú vrstvu môžeme pomocou metódy gradientového zostupu pre strednú kvadratickú chybu odvodiť nasledovné pravidlo pre učenie výstupných váh:

$$\frac{\partial E}{\partial w_{ij}^{\text{out}}} = (d_i - y_i) \left(\prod_{k \neq j} 1 - \sigma(w_{ik}^{\text{out}})(1 - h_k) \right) \tag{6}$$
$$(h_j - 1)\sigma(w_{ij}^{\text{out}})(1 - \sigma(w_{ij}^{\text{out}}))$$

a spätné šírenie chyby na skrytú vrstvu:

$$\frac{\partial E}{\partial h_i} = \sum_k (d_k - y_k) \left(\prod_{l \neq i} 1 - \sigma(w_{kl}^{\text{out}})(1 - h_l) \right) \sigma(w_{ki}^{\text{out}}) \tag{7}$$

Čiastočné násobenie v pravidlách učenia môžeme pre urýchlenie výpočtu nahradiť delením v prípade, že sa vo výpočte nevyskytne delenie nulou:

$$\prod_{k \neq j} 1 - \sigma(w_{ik}^{\text{out}})(1 - h_l) = \frac{y_i}{1 - \sigma(w_{ij}^{\text{out}})(1 - h_j)} \tag{8}$$

Takto prešírenú chybu potom použijeme na učenie vstupných váh:

$$\Delta w_{ij}^{\text{hid}} \sim \frac{\partial E}{\partial w_{ij}^{\text{hid}}} = \frac{\partial E}{\partial h_i}(1 - h_i^2)x_j \tag{9}$$

Náš model nie je nutné ďalej obmedzovať ani stabilizovať, možno k nemu pristupovať po všetkých stránkach ako ku viacvrstvovému perceptrónu. Násobiacu vrstvu je možné zapájať do siete aj iným spôsobom, napríklad striedať sumačné a násobiace vrstvy, najskôr násobiť a potom sčítavať a podobne. Taktiež je možné funkciu f modifikovať alebo preškálovať podľa vlastností, ktoré chceme dosiahnuť a model bude vyzerať veľmi podobne. V našich experimentoch sme zatiaľ tieto možnosti neskúmali.

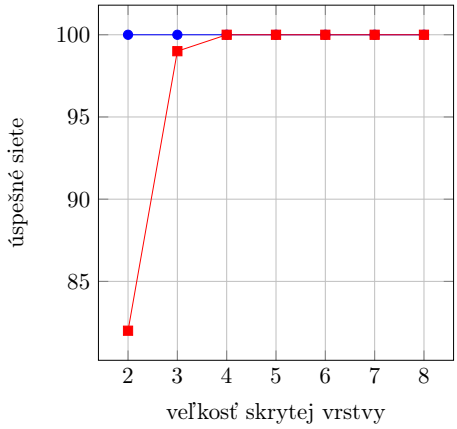
4 Experimenty

Pre základné porovnanie sme použili základný viacvrstvový perceptrón (MLP) bez akejkoľvek regularizácie. Náš model sme naimplementovali úpravou pôvodného MLP, aby bolo porovnanie čo najvernejšie. Z našich predbežných výsledkov vyberáme problém XOR a jeho všeobecnejšiu verziu pre viac vstupov, čiže paritu. Problém parity, teda určenia, či má binárne číslo párny alebo nepárny počet jednotiek, je klasická, pomerne ťažká úloha pre neurónové siete a používa sa na testovanie a overovanie nových modelov.

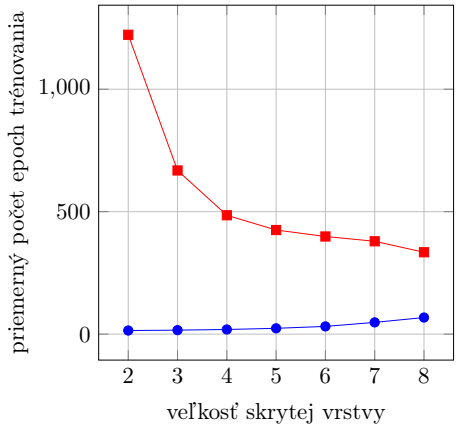
V našich experimentoch vyhodnocujeme konvergenciu, čiže koľko sietí zo 100 dospeje k stabilnému riešeniu. Za natrénovanú považujeme sieť, ktorá dáva správny výstup desať po sebe idúcich epoch.

Na natrénovanie dávame sieťam maximálne 3 tisíc epoch. Uvádzame aj priemerný počet epoch trénovania všetkých 100 sietí, čiže aj neúspešných. Pre oba typy siete používame jednotnú rýchlosť učenia $\alpha = 0.9$ a Gaussovskú inicializáciu váh s distribúciou $\mathcal{N}(0; 0, 5)$.

Na Obr. 2 a Obr. 3 zobrazujeme vplyv veľkosti skrytej vrstvy na konvergenciu sietí pri úlohe XOR. Pri takzvanom minimálnom XOR-e s 2 skrytými neurónmi pozorujeme jednoznačnú výhodu nášho modelu. Zatiaľ neexistuje nám známy model trojvrstvovej siete bez reziduálnych váh, ktorý sa v minimálnej architektúre s 2 skrytými neurónmi naučí tento problém na 100%, t.j. že všetky trénované siete nájdu riešenie. V tabuľke 1 zobrazujeme výsledky prvotných experimentov pre paritu $2 - 7$, kde parita 2 predstavuje XOR. Pre MLP sme už pre paritu vyššiu ako 5 nenašli optimálnu veľkosť skrytej vrstvy. Pri danom obmedzení počtu trénovacích epoch, ktoré sme stanovili vyššie, sa MLP úlohu nevedel naučiť.



Obr. 2: Vplyv veľkosti skrytej vrstvy na počet sietí konvergujúcich k riešeniu pre problém XOR. (červená - MLP, modrá - náš model)



Obr. 3: Vplyv veľkosti skrytej vrstvy na priemerný počet epoch trénovania pre problém XOR. (červená - MLP, modrá - náš model)

	Náš model		**MLP**	
N-parita	**h**	konverg.	**h**	konverg.
2	2	100	4	100
3	4	100	9	100
4	6	100	12	91
5	7	100	50	44
6	12	100	0	-
7	15	99	0	-

Tab. 1: Optimálna minimálna veľkosť skrytej vrstvy **h** a maximálny počet sietí, ktoré skonvergovali k riešeniu zo 100 pre paritu na $2 - 7$ bitoch.

5 Záver

Naša práca kompenzuje niektoré nedostatky násobiacich neurónových sietí a ukazuje spôsob, ako možno trénovať váhy multiplikatívnych zložiek všeobecným spôsobom a bez komplexných čísel. Náš model je matematicky jednoduchý, ľahko implementovateľný a kombinovateľný s existujúcimi modelmi, s prísľubom zvýšenia ich efektivity v niektorých typoch úloh, ktoré sú pre klasické neurónové siete pomerne ťažko riešiteľné.

Poďakovanie

Za podporu ďakujeme Slovenskej spoločnosti pre kognitívnu vedu[1].

Literatúra

Andersen, R. A., Essick, G. K. a Siegel, R. M. (1985). Encoding of spatial location by posterior parietal neurons. *Science*, 230(4724):456–458.

Anthony, M. a Bartlett, P. L. (1999). *Neural network learning: Theoretical foundations*, vol. 9. cambridge university press Cambridge.

Bugmann, G. (1991). Summation and multiplication: two distinct operation domains of leaky integrate-and-fire neurons. *Network: Computation in Neural Systems*, 2(4):489–509.

Delalleau, O. a Bengio, Y. (2011). Shallow vs. deep sum-product networks. *Advances in Neural Information Processing Systems*, 24.

Diba, A., Sharma, V. a Van Gool, L. (2017). Deep temporal linear encoding networks. V *Proceedings of*

[1] https://cogsci.fmph.uniba.sk/sskv/

the IEEE conference on Computer Vision and Pattern Recognition, str. 2329–2338.

Duda, R. O., Hart, P. E. a Stork, D. G. (1973). *Pattern classification and scene analysis*, vol. 3. Wiley New York.

Durbin, R. a Rumelhart, D. E. (1989). Product units: A computationally powerful and biologically plausible extension to backpropagation networks. *Neural Computation*, 1(1):133–142.

Erhan, D., Bengio, Y., Courville, A. a Vincent, P. (2009). Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1.

Ghosh, J. a Shin, Y. (1992). Efficient higher-order neural networks for classification and function approximation. *International Journal of Neural Systems*, 3(04):323–350.

Giles, C. L. a Maxwell, T. (1987). Learning, invariance, and generalization in high-order neural networks. *Applied Optics*, 26(23):4972–4978.

Heywood, M. a Noakes, P. (1995). A framework for improved training of sigma-pi networks. *IEEE transactions on Neural Networks*, 6(4):893–903.

Mel, B. a Koch, C. (1989). Sigma-pi learning: On radial basis functions and cortical associative learning. *Advances in Neural Information Processing Systems*, 2.

Mel, B. W. (1994). Information processing in dendritic trees. *Neural Computation*, 6(6):1031–1085.

Nilsson, N. J. (1965). *Learning machines.* McGrawHill New York.

Poon, H. a Domingos, P. (2011). Sum-product networks: A new deep architecture. V *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, str. 689–690. IEEE.

Redding, N. J., Kowalczyk, A. a Downs, T. (1993). Constructive higher-order network that is polynomial time. *Neural Networks*, 6(7):997–1010.

Rumelhart, D. E., Hinton, G. E. a McClelland, J. L. (1986). *A general framework for parallel distributed processing*, vol. 1, str. 26. Cambridge, MA: MIT Press.

Schenck, C. a Fox, D. (2018). Spnets: Differentiable fluid dynamics for deep neural networks. V *Conference on Robot Learning*, str. 317–335. PMLR.

Schmitt, M. (2002). On the complexity of computing and learning with multiplicative neural networks. *Neural Computation*, 14(2):241–301.

Shin, Y. a Ghosh, J. (1995). Ridge polynomial networks. *IEEE Transactions on neural networks*, 6(3):610–622.

Spirkovska, L. a Reid, M. B. (1994). Higher-order neural networks applied to 2d and 3d object recognition. *Machine Learning*, 15(2):169–199.

Suga, N., Olsen, J. a Butman, J. (1990). Specialized subsystems for processing biologically important complex sounds: Cross-correlation analysis for ranging in the bat's brain. V *Cold Spring Harbor symposia on quantitative biology*, vol. 55, str. 585–597. Cold Spring Harbor Laboratory Press.

Zhu, J., Zeng, H., Du, Y., Lei, Z., Zheng, L. a Cai, C. (2018). Joint feature and similarity deep learning for vehicle re-identification. *IEEE Access*, 6:43724–43731.

Intrinsic motivation based on feature extractor distillation

Matej Pecháč and Igor Farkaš

Faculty of Mathematics, Physics and Informatics
Comenius University in Bratislava, Slovak Republic
matej.pechac@gmail.com

Abstract

Reinforcement learning can solve decision-making problems and teach an AI agent to behave in an environment according to a pre-designed reward function. However, such an approach becomes very problematic if the reward function is too sparse and the agent does not come across the reward during the environmental exploration. The solution to such a problem may be in equipping the agent with an intrinsic motivation, which will provide informed exploration, during which the agent must is likely to also encounter external reward. Novelty detection is one of the promising branches of intrinsic motivation research. We present a two variants of novelty detector based on distillation of feature extractor which is learned by contrastive loss. The results show that such an approach can achieve faster growth and higher external reward for the same training time compared to the baseline model, which implies improved exploration in a very sparse reward environment. The source code is available at `https://github.com/Iskandor/MotivationModels`

1 Introduction

The development of reinforcement learning (RL) methods has achieved much success over the last decade, since together with advances in computer vision (Krizhevsky et al., 2012; He et al., 2016), it became possible to teach agents to solve various tasks, play computer games (Mnih et al., 2013), even surpassing human players (Mnih and et al., 2015). Nevertheless, these are still concrete single tasks. Training times are very long and the agents need a lot of resources. Coping with complex (continuous) environments such as our world is still a challenge. There are several research opportunities. One is the search for more efficient learning methods. Another is hardware development, which attempts to adapt to the requirements of neural networks that are currently being used in the RL field.

The most popular approach to make RL more efficient is based on *intrinsic motivation* (IM) (Baldassarre et al., 2014). IM has a strong psychological motivation (Ryan and Deci, 2000), observed in children during development. If we want to achieve an open-ended development with artificial agents, we have to master this first step and equip them with an ability to generate their own goals and acquire new skills. Therefore, computational approaches concerned with IMs and open-ended development provide the potential in this direction leading to more intelligent systems, in particular those capable of improving their own skills and knowledge autonomously and indefinitely (Baldassarre et al., 2014).

In particular, we provide three main contributions: First, we improved the stability of the Spatio-Temporal DeepInfoMax algorithm (Anand et al., 2019) in the conditions of an incomplete dataset (online learning), which we use for feature extractor training. Second, we propose a new model for novelty detection (inspired by Burda et al. (2018b)), which serves as a source of intrinsic motivation based on the distillation of the feature extractor. Third, we hypothesize that a properly trained feature extractor can serve both as a source of features for the actor and critic and as a source of internal motivation, leading to a simplification of the model. We also managed to perform the first tests supporting this idea.

2 Related work

The concept of intrinsic (and extrinsic) motivation was first studied in psychology (Ryan and Deci, 2000), and later entered the RL literature where the first taxonomy of computational models appeared in Oudeyer and Kaplan (2009). Following this taxonomy, we can divide the concept of motivation into external and internal, depending on the mechanism that generates motivation for the agent. If the source of motivation comes from outside, we are talking about *external* motivation, and it is always associated with a particular goal in the environment. If the motivation is generated within the structures that make up the agent, it is an *internal* motivation.

Another dimension for the differentiation, extrinsic or intrinsic, is less obvious. *Extrinsic* motivations pertain to behaviors whenever an activity is done in order to attain some separable outcome. Some variability exists in this context, since these behaviors can vary in the extent to which they represent self-determination (see the details in Ryan and Deci (2000)). On the other hand, *intrinsic* motivation is defined as doing an activity for its inherent satisfactions rather than for some separable consequence (or instrumental value). It has been operationally defined in various ways, backed up by dif-

ferent psychological theories, which point to some uncertainty in what IM exactly means. Nevertheless, Baldassarre (2019) offers a solution of an operational definition of IMs as processes that can drive the acquisition of knowledge and skills in the absence of extrinsic motivations. Furthermore, he proposes (and explains why) a new term of *epistemic motivations* as a suitable substitution for intrinsic motivations.

According to the prevailing view, the computational approaches to IM can be divided into two main categories with adaptive motivations. *Knowledge-based* approach is focused on acquisition of knowledge of the world and draws on the theory of drives, theory of cognitive dissonance and optimal incongruity theory. *Competence-based* approach focuses on acquisition of skills by motivating the agent to achieve a higher level of performance in the environment, which means to acquire desired actions to achieve self-generated goals. Its psychological basis includes the theory of effectance and the theory of flow.

The knowledge-based category is commonly divided into *prediction-based* and *novelty-based* approaches. Prediction-based approaches often use a forward model (e.g. Stadie et al. (2015); Bellemare et al. (2013); Pathak et al. (2017)) or a variational autoencoder Kingma and Welling (2013) to compute the prediction error (for more details, see Burda et al. (2018a)). The novelty-based approaches monitor the state novelty and the intrinsic signal is based on its value. The first models were based on count-based approach Tang et al. (2017). This method is impractical for large or continuous state spaces and it was extended by introducing pseudo-count and neural density models Ostrovski et al. (2017); Martin et al. (2017); Machado et al. (2018). A similar method to pseudo-count was used by a random network distillation (RND) model (Burda et al., 2018b) with a lower complexity.

Contrastive learning (Chopra et al., 2005) is a machine learning technique used to learn the general features of a dataset without labels by teaching the model which data points are similar or different. Several different objective functions were proposed e.g. Noise Contrastive Estimation (NCE) (Gutmann and Hyvärinen, 2010), InfoNCE (van den Oord and andOriol Vinyals, 2018), multi-class N-pair loss (Sohn, 2016). Contrastive learning also started to be used in the field of state representation learning (Lesort et al., 2018) and is proving to be a suitable method for creating feature space (Anand et al., 2019) and also finds its use in reinforcement learning (Srinivas et al., 2020).

3 Methods

The decision making problem in the environment using RL is formalized as a Markov decision process which consists of a state space $\mathcal{S}$, action space $\mathcal{A}$, transition function $\mathcal{T}_{s,a,s'} = p(s_{t+1} = s' | s_t = s, a_t = a)$, reward function $\mathcal{R}_{sas'}$ and a discount factor γ. The main goal of the agent is to maximize the discounted return $R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$ in each state, where r_t is immediate external reward at time t. Stochastic policy is defined as a state dependent probability function $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$, such that $\pi_t(s, a) = p(a_t = a | s_t = s)$ and $\sum_{a \in \mathcal{A}} \pi(s, a) = 1$ and the deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$ is defined as $\pi(s) = a$.

An agent following the optimal policy π^* maximizes the expected return R. The methods searching for the optimal policy can be divided into on-policy (family of actor–critic algorithms), and off-policy (family of Q-learning algorithms) methods. Actor–critic algorithms are based on two separate modules: an *actor* approximates agent's policy π and generates actions and a *critic* estimates the state value function V^π defined as

$$V^\pi(s) = \sum_a \pi(s, a) \sum_{s'} \mathcal{T}_{s,a,s'} \left[\mathcal{R}_{s,a,s'} + \gamma V^\pi(s') \right]$$

or action-state value function Q^π defined as

$$Q^\pi(s, a) = \sum_{s'} \mathcal{T}_{s,a,s'} \left[\mathcal{R}_{s,a,s'} + \gamma V^\pi(s') \right]$$

The actor then updates its policy to maximize return R based on critic's value function estimations.

3.1 Random Network Distillation model

The RND model (Burda et al., 2018b) has two components: randomly initialized (and fixed) target network Φ_T that generates random features, and the learning network Φ_L that tries to predict them. Intrinsic motivation is computed as the prediction error, defined as

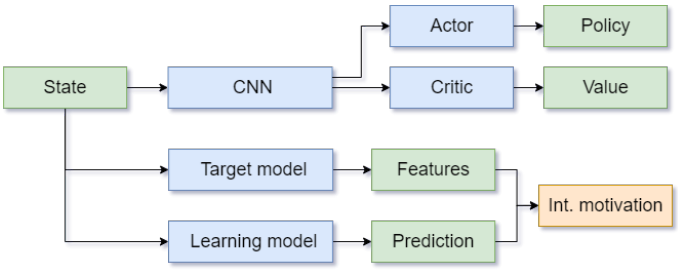
$$r_{\text{intr}} = \frac{1}{2} \| (\Phi_L(s_t) - \Phi_T(s_t)) \|^2 \qquad (1)$$

using the Euclidean norm. The model is simple and successful in the environments with sparse reward but has two serious disadvantages: (1) It is necessary to properly initialize the random network; and (2) over time, the signal of intrinsic motivation disappears due to sufficient adaptation of the learning network (a phenomenon that could be called generalization).

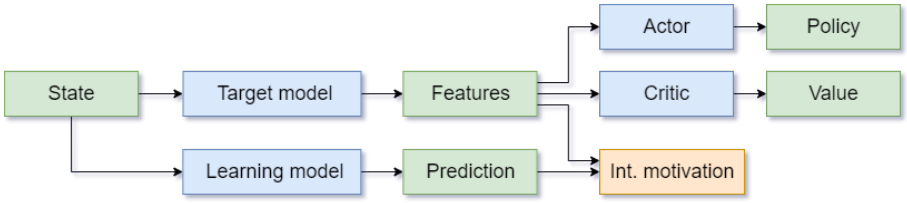
3.2 RND model extended with action

A simple extension that we propose is to add an action to the input, yielding the RNDa model. The randomly initialized target network Φ_T and the learning network Φ_L have two branches, one using convolutional neural network (CNN) to process the state and the other using multi-layerd perceptron (MLP) for action, and at the end it mixes both feature vectors into one representation. Intrinsic motivation is then computed as

$$r_{\text{intr}} = \frac{1}{8} \| (\Phi_L(s_t, a_t) - \Phi_T(s_t, a_t)) \|^2 \qquad (2)$$



(a) Feature extractor distillation separated model (FEDs)



(b) Feature extractor distillation model (FED)

Fig. 1: The schema of the feature extractor distillation models. In both variant, the target model (implemented as a convolutional neural network, CNN) transforms the raw state vector into a feature vector. The learning model (also a CNN) tries to predict the feature vector of the target model and the error between the two vectors serves as an intrinsic motivation signal. **(a)** In FEDs variant, the actor and the critic learn from a fixed CNN, whereas in **(b)** FED variant, they use the feature vector of the target model for further policy generation and value function estimation.

with an aim to increase the complexity of the input and thus prevent an early decline in intrinsic motivation. However, the key problems of the RND model are not solved.

3.3 Feature extractor distillation (FED)

We propose two models based on concept of distillation of feature extractor instead of randomly initialized network like RND. The schematic representation of both models is shown in Fig. 1.

By feature extractor we mean a module that learns to create a meaningful feature space according to a certain loss function, which is independent of other modules that consume the features of this extractor, and these modules could act properly. In other words, the error gradient is not back-propagated to the feature extractor from modules that use its feature vectors as input. By distillation we mean the transmission of the transformation represented by one neural network to another, whereby both networks will generate similar outputs for the same inputs.

In the first stage (denoted as Feature extractor distillation separated - FEDs - see Fig. 1a), we decided to design and test a model similar to the RND model, but with the addition of an objective function for target network training. Such a model has a module for learning policy and value function, consisting of a CNN that feeds two MLPs in roles of actor and critic. The goal of the second module is to generate internal motivation, which consists of target network Φ_T returning feature vectors and a learning network Φ_L that learns the same transformation and returns estimates of the feature vectors generated by target network. Both networks are CNNs. To this point, the architecture coincides with the RND model. The difference is that we added the learning rule for the target network. Following Anand et al. (2019), we use the Spatio-Temporal DeepInfoMax (ST-DIM) algorithm leveraging multi-class N-pair losses (Sohn, 2016):

$$\mathcal{L}_{\mathrm{GL}} = -\sum_{i=1}^{I}\sum_{j=1}^{J} \log \frac{\exp(g_{i,j})}{\sum_{s_t^* \in S_{\mathrm{next}}} \exp(g_{i,j})} \tag{3}$$

$$\mathcal{L}_{\mathrm{LL}} = -\sum_{i=1}^{I}\sum_{j=1}^{J} \log \frac{\exp(f_{i,j})}{\sum_{s_t^* \in S_{\mathrm{next}}} \exp(f_{i,j})} \tag{4}$$

where $f(.) = f(s_t, s_{t+1})$ and $g(.) = g(s_t, s_{t+1})$ are score functions for local-local objective $\mathcal{L}_{\mathrm{LL}}$ and global-local objective $\mathcal{L}_{\mathrm{GL}}$ respectively. Function $g_{i,j}$ is defined as non-normalized cosine similarity between transformed global feature $\Phi_T(s_t)$ and local feature $\Phi_T^{(l,i,j)}(s_{t+1})$ of intermediate layer l in Φ_T, where (i, j) is spatial location. Analogically $f_{i,j}$ is non-normalized cosine similarity between transformed local features $\Phi_T^{(l,i,j)}(s_t)$ and $\Phi_T^{(l,i,j)}(s_{t+1})$. Details of this algorithm are provided in Anand et al. (2019). S_{next} corresponds to the set of next states, (s_t, s_{t+1}) represents a pair of consecutive states, (s_t, s_{t*}) represents a pair of non-consecutive states and I, J are the width and height from output shape of intermediate convolutional layer of the target network. The resulting loss function is then

defined as

$$\mathcal{L} = \frac{1}{IJ}(\mathcal{L}_{\mathrm{GL}} + \mathcal{L}_{\mathrm{LL}}) \tag{5}$$

Following this objective function, the target network becomes a good feature extractor adapting to new states discovered by the agent. However, after initial tests, we found that the feature space formed by such an objective function tends to grow exponentially from a certain point until it eventually explodes. We provide a more detailed analysis of this problem in Section 5. The solution to this problem was to find a suitable regularization that would add to the existing loss function. In total, we tested three regularization terms:

1. Maximize entropy H to smooth logits represented by functions f and and g:

$$H_{\mathrm{GL}} = -\sum_{i=1}^{I}\sum_{j=1}^{J}\sigma(g_{i,j}) \cdot \log\sigma(g_{i,j})$$

$$H_{\mathrm{LL}} = -\sum_{i=1}^{I}\sum_{j=1}^{J}\sigma(f_{i,j}) \cdot \log\sigma(f_{i,j})$$

   where $\sigma(.)$ is standard softmax function, and the overall loss

$$\mathcal{L}_{\mathrm{reg}} = -(H_{\mathrm{GL}} + H_{\mathrm{LL}}). \tag{6}$$

2. Minimize L_2-norm of global features:

$$\mathcal{L}_{\mathrm{reg}} = \|\Phi_T(s_t)\| \tag{7}$$

3. Minimize L_2-norm of logits represented by functions f and g:

$$\mathcal{L}_{\mathrm{reg}} = p_{\mathrm{GL}} + p_{\mathrm{LL}} = \sum_{i=1}^{I}\sum_{j=1}^{J}(\|f_{i,j}\| + \|g_{i,j}\|) \tag{8}$$

According to the test results we decided to use the third option (eq. 8) and minimize the L_2-norm logits f and g. The final objective function, with scaling parameter $\beta = 0.001$, was defined as

$$\mathcal{L} = \frac{1}{IJ}(\mathcal{L}_{\mathrm{GL}} + \mathcal{L}_{\mathrm{LL}} + \beta\mathcal{L}_{\mathrm{reg}}) \tag{9}$$

In the second stage (FED, see Fig. 1b), we propose to replace CNN, which is used by the actor and the critic, with the target network, which is trained using the ST-DIM algorithm. We assume that the features it generates could be suitable for the successful functioning of the actor and the critic. This would reduce the number of networks needed and reduce the model complexity. The target network would serve both to generate intrinsic motivation and as an input for actors and critics.

In both models, the definition of intrinsic reward is the same:

$$r_{\mathrm{intr}} = \frac{1}{n}\|(\Phi_T(s_t) - \Phi_L(s_t))\|^2 \tag{10}$$

4 Experiments

For experiments, we chose Montezuma's Revenge environment for the Atari console. This is an environment with a very sparse reward, where it is almost impossible to find an optimal policy without internal motivation. The agent's goal is to overcome obstacles in individual rooms and to obtain and use items. The agent receives a reward of +1 for each increase in the score, regardless of its size. It does not receive any other reward or punishment. The state space consists of 4 consecutive frames of pixels on grey scale, so the total dimension of the state space is 4×96×96×256. The action space is discrete, consisting of 18 actions, of which 5 make sense, the others have no impact on the environment.

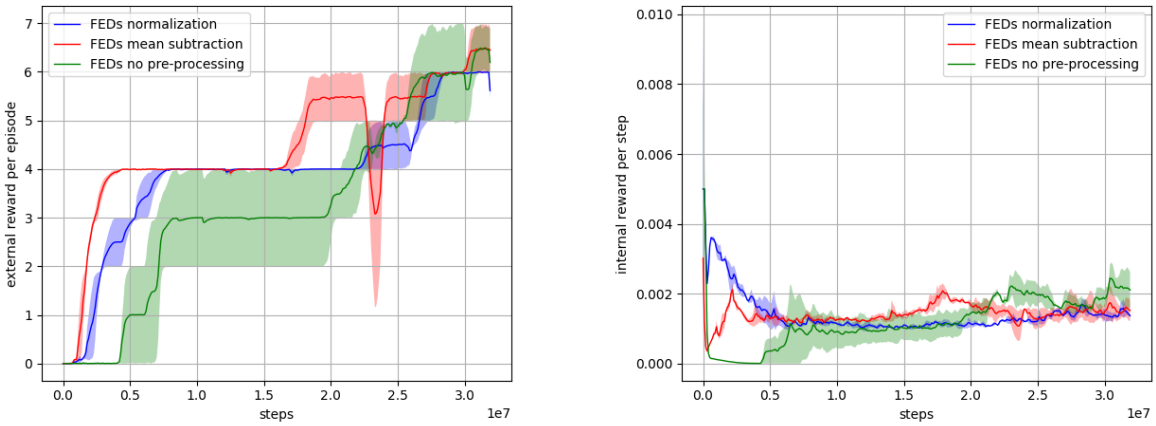
4.1 Training setup

All agents were trained using the PPO algorithm (Schulman et al., 2017) with 128 environments and we used Adam algorithm (Kingma and Ba, 2015) to optimize the parameters of all modules. The basic agent consists of an actor and a critic, which are two MLPs sharing a common CNN that processes the video input. The critic has two outputs (heads), one for estimating the value function for the external and the other for the internal reward. The discount factor γ for external reward is 0.998 and for internal reward 0.99. The motivational part consists of two CNNs (target and learning network), which receive pre-processed input (see 4.2) from 1 frame. The learning network has two more linear layers to have an increased capacity over the target network. In the case of the FED model, the motivation module contains only one CNN (learning network) and uses a feature extractor as the target network, which is also connected to the actor and the critic. Feature extractor receives on the input 4 consecutive frames. More hyper-parameters and further details of the learning process and architecture of modules can be found in our source codes.

4.2 State pre-processing study

The state before entering the motivation module of FEDs model can undergo pre-processing. In the case of the FED model, it is necessary to enter the full raw state. We tested three pre-processing methods: (1) Normalization of the state using the running mean and standard deviation, (2) Subtraction of the running mean value from the state, (3) Without pre-processing.

We trained two agents for each method in 32M steps. The test results in Tab. 1 and Fig. 2 show that state pre-processing did not have a significant effect on agent's performance (maximum reward achieved), only on the speed. This also agrees with our assumption that operations such as subtraction of the mean or normalization should be able to find the network itself trained using the contrastive loss function. Therefore it is not



(a) External reward

(b) Intrinsic reward

Fig. 2: Evolution of rewards in case of FEDs model with three state pre-processing methods.

necessary for the designer to put them into the learning process explicitly. These conclusions will still need to be confirmed by statistical analysis and show that there is no significant difference in the results achieved.

Tab. 1: Average cumulative reward per episode for all three pre-processing methods and average intrinsic reward per step.

Method	External reward	Intrinsic reward
norm.	3.60 ± 0.14	0.0016 ± 0.00008
mean sub.	4.13 ± 0.12	0.0013 ± 0.00005
none	2.31 ± 0.20	0.0009 ± 0.0001

Tab. 2: Average cumulative external reward per episode and mean intrinsic reward per step for tested agents. Compared on 128M steps and 32M steps (because the FED model has not yet been trained in 128M steps).

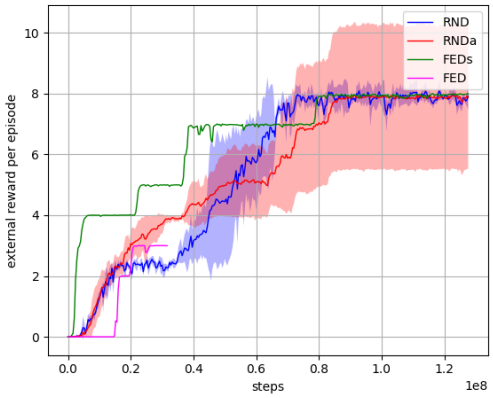
Agent (128M)	External reward	Intrinsic reward
RND	4.78	0.051
RNDa	5.15	0.014
FEDs	6.62	0.001
Agent (32M)		
RND	0.93	0.086
RNDa	1.43	0.033
FEDs	2.11	0.001
FED	1.08	0.005

4.3 Results

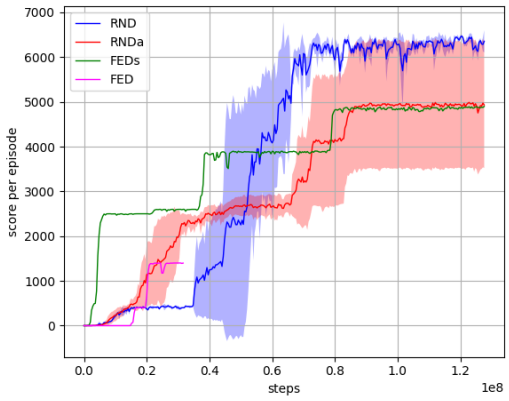
We trained three agents for the RND and RNDa model, one agent for the FEDs model and one agent for the FED model, trained in 32M steps. The results shown in Fig. 3 and Tab. 2 reveal that there is no significant difference between the RND and RNDa models, since both were able to achieve the same average reward at approximately the same time during the 128M step training. The addition of an action to the input did not delay the disappearance of the intrinsic reward (see Fig. 3c), as we had anticipated, but it brought about changes in the learned policy, where the agent discovered other sources of reward, as can be seen in Fig. 3b. The FEDs model achieved a faster increase in an external reward and also showed greater stability (although the sample is not large enough to warrant such a claim). The FED model achieved a higher intrinsic motivation compared to the FEDs model (compare Tab. 2 and Fig. 3c, which we attribute to the differences in their inputs - 4 frames vs 1 frame). For the FED model, the challenge is to learn to predict feature vectors for 4 frames, where e.g. two states may have the same last frame but the previous 3 may be different and from this point of view this input appears different for the FED model, while for FEDs, that only takes the last frame, it would appear to be the same. Therefore, the FED model explored the environment more slowly, but according to preliminary results, it does not seem to have a significant impact on the agent's overall performance. Compared to the RND model, it nevertheless achieved better results.

5 Discussion

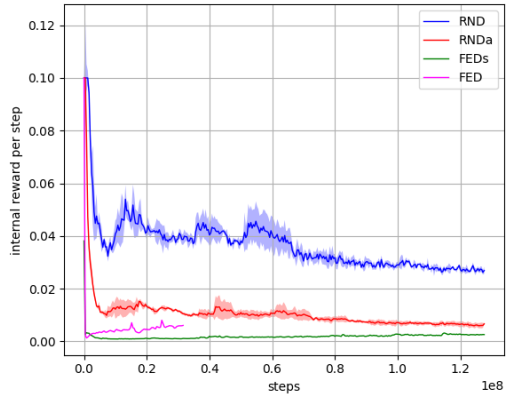
We have introduced a simple extension (RNDa) of the Random Network Distillation model and two variants of the model with intrinsic motivation derived from the



(a) External reward



(b) Score



(c) Intrinsic reward

Fig. 3: Evolution of external reward, score and intrinsic reward that agents received during learning. The agent FEDs grew faster than the RND and RNDa agents and achieved the same highest reward. In the score, we see that the RND, RNDa and FEDs each found a different policy and collected different items that led to different scores.

distillation of the feature extractor – the FEDs (which is similar to RND) and the FED model. Both proposed variants try to eliminate the identified shortcomings of the RND model – the need for good initialization and the loss of the motivational signal caused by the adaptation of the learning network. We also simplified he whole architecture with the FED model and used the ST-DIM algorithm to train the target network in both FED variants.

Our experiments revealed that if the ST-DIM algorithm works on an incomplete dataset that takes on new samples (the authors probably did not test it in such conditions), there is an instability and an exponential increase of activity in the feature space at certain moments. This is related to the use of cross-entropy loss function in its core (which does not limit the values of inputs – logits), where derivatives can reach large values and subsequently inflate the entire feature space.

We also found from observations that old states occur at the edge of the feature space and thus reach greater values than the new states that appear closer to zero. This requires further analysis, though. Therefore, we think that if a new state emerges that differs significantly from all previous ones, there may be a large growth of the entire feature space, which is further accelerated by large values of feature representations of the old states located on the edge. For this reason, we had to introduce a regularization expression into the loss function of the ST-DIM algorithm.

During the development of the model, it turned out that it is best to minimize the L_2-norm of logits that enter the cross-entropy. In addition, we tried to maximize the entropy of the distributions generating the respective logits and minimize the L_2-norm of global features. However, both described approaches failed to sufficiently stabilize the algorithm.

In the experiments, we tested the overall performance of the agents. The RNDa model did not make a big difference in agent performance compared to the RND model, which served us as a baseline. Neither did we observe the expected longer decline in internal motivation compared to RND. FEDs proved to be very promising and outperformed both RND and RNDa.

We also compared the effect of state preprocessing on the performance of the FEDs model. It turned out that the state pre-processing is not necessary since it has no significant effect on the agent's results. The FED model is proving to be a viable solution that may not grow as fast as the FEDs, but may ultimately achieve the same results. From a computational point of view, however, this is a simpler and faster solution, where it is not necessary to duplicate the CNN processing of the image input for other modules.

In the future, we plan to analyze more accurately the behavior of the feature space during training and verify the performance of FED models in other environments. We are also considering extending the input

of the FED model by an action, similar to the RNDa model. There is also room for fine-tuning the hyper-parameters, which could further improve the results.

References

Anand, A., Racah, E., Ozair, S., Bengio, Y., Côté, M., and Hjelm, R. D. (2019). Unsupervised state representation learning in atari. CoRR, abs/1906.08226.

Baldassarre, G. (2019). Intrinsic motivations and open-ended learning. arXiv:1912.13263v1 [cs.AI].

Baldassarre, G., Stafford, T., Mirolli, M., Redgrave, P., Ryan, R. M., and Barto, A. (2014). Intrinsic motivations and open-ended development in animals, humans, and robots: An overview. Frontiers in Psychology.

Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. (2013). The arcade learning environment: An evaluation platform for general agents. Journal of Artificial Intelligence Research, 47:253–279.

Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., and Efros, A. A. (2018a). Large-scale study of curiosity-driven learning. arXiv:1808.04355.

Burda, Y., Edwards, H., Storkey, A., and Klimov, O. (2018b). Exploration by random network distillation. arXiv:1810.12894.

Chopra, S., Hadsell, R., and LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 1, pages 539–546 vol. 1.

Gutmann, M. and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In 13th International Conference on Artificial Intelligence and Statistics, volume 9, pages 297–304.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Conference on Computer Vision and Pattern Recognition.

Kingma, D. P. and Ba, J. L. (2015). Adam: A method for stochastic optimization. In International Conference on Learning Representations.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. arXiv:1312.6114.

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. Neural Information Processing Systems, 25.

Lesort, T., Rodríguez, N. D., Goudou, J., and Filliat, D. (2018). State representation learning for control: An overview. CoRR, abs/1802.04181.

Machado, M. C., Bellemare, M. G., and Bowling, M. (2018). Count-based exploration with the successor representation. arXiv:1807.11622.

Martin, J., Sasikumar, S. N., Everitt, T., and Hutter, M. (2017). Count-based exploration in feature space for reinforcement learning. arXiv:1706.08090.

Mnih, V. and et al. (2015). Human-level control through deep reinforcement learning. Nature, 518:529–533.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing Atari with deep reinforcement learning. arXiv:1312.5602.

Ostrovski, G., Bellemare, M. G., van den Oord, A., and Munos, R. (2017). Count-based exploration with neural density models. In International Conference on Machine Learning, pages 2721–2730.

Oudeyer, P.-Y. and Kaplan, F. (2009). What is intrinsic motivation? a typology of computational approaches. Frontiers in Neurorobotics, 1:6.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. arXiv:1705.05363.

Ryan, R. and Deci, E. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. Contemporary Educational Psychology, 25(1):54–67.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. arXiv:1707.06347.

Sohn, K. (2016). Improved deep metric learning with multi-class n-pair loss objective. In Advances in Neural Information Processing Systems, volume 29. Curran Associates, Inc.

Srinivas, A., Laskin, M., and Abbeel, P. (2020). CURL: contrastive unsupervised representations for reinforcement learning. CoRR, abs/2004.04136.

Stadie, B. C., Levine, S., and Abbeel, P. (2015). Incentivizing exploration in reinforcement learning with deep predictive models. arXiv:1507.00814.

Tang, H. et al. (2017). #Exploration: A study of count-based exploration for deep reinforcement learning. In Advances in Neural Information Processing Systems, pages 2753–2762.

van den Oord, A. and andOriol Vinyals, Y. L. (2018). Representation learning with contrastive predictive coding. CoRR, abs/1807.03748.

BESST: Brno Extended Speech and Stress Test

Jan Pešán

Ústav počítačové grafiky a multimédií, Fakulta informačních technologií, Vysoké učení technické v Brně
Email: ipesan@fit.vutbr.cz

Vojtěch Juřík

Psychologický ústav, Filozofická fakulta Masarykovy univerzity, Brně
Email: jurik.vojtech@mail.muni.cz

Abstrakt

Detekce stresu je tradičním tématem v oblasti automatického zpracování řeči. Historicky problematickou oblastí v rámci přístupu tzv. hlubokého učení je nedostatek kvalitních referenčních dat pro trénink umělých systémů. V aktuálním výzkumu jsme aplikovali psychologickou metodologii do oblasti IT, abychom mohli shromáždit potřebná empirická data vhodná pro efektivní trénink hlubokých neuronových sítí v kontextu řečové zátěže. Modely neuronových sítí, poháněné adekvátními datovými vstupy, mohou významně podpořit klasifikaci a detekci stresu při automatickém zpracování řeči. Za tímto účelem byl vyvinut Brněnský rozšířený test řeči a zátěže (BESST), který je rozšířenou adaptací původního protokolu Maastrichtského akutního zátěžového testu (MAST). Upravená metodika BESST má za cíl maximalizovat sběr řečových výstupů od účastníků v různých stresových kontextech. Navrhovaná metodologie představená v tomto článku představuje funkční a škálovatelný nástroj pro sběr klíčových datových sad nezbytných pro detekci stresu pomocí technik hlubokého učení.

1 Strojové zpracování řečového projevu

Strojové učení (machine learning; ML), respektive celá oblast umělé inteligence (artificial intelligence; AI), se v posledních několika desetiletích nepopiratelně prosazuje jako nejrychleji rostoucí odvětví průmyslu. Neustálý rozvoj v této oblasti výrazně posunul také možnosti rozpoznávání a klasifikace různých aspektů řečového projevu člověka. I když oblast výzkumu zpracování řeči sahá již do 80. let 20. století, relativně vysoká technologická náročnost problematiky bránila využití této technologie pro širokou veřejnost a vývoj se zaměřoval především na korporátní a státní sféru. Později nicméně zavedení výkonných kapesních zařízení ve formě chytrých telefonů významně ovlivnilo celkovou situaci a rozpoznávání řeči našlo uplatnění u koncového uživatele. Masové přijetí technologií pro zpracování řeči začalo často uváděným milníkem, tedy představením aplikace Siri - Hlasového asistenta pro iPhony od firmy Apple. Nejrozšířenější podoblastí u zpracování řeči je bezpochyby automatické rozpoznávání řeči (automatic speech recognition; ASR), které se používá k automatickému přepisu řeči ze zvukového záznamu. Oblast ASR s postupným vývojem naplňuje přísliby technologie AI, protože zdůrazňuje dostupnost, jednoduchost, spolehlivost a rychlost, a často jsou zastoupeny všechny tyto faktory najednou. Kromě standardních funkcionalit jako je vytváření přepisů řečového projevu plní ASR další důležité úkoly, kam patří např. identifikace pohlaví mluvčího (GID), identifikace jazyka (LID) či osobní identifikace mluvčího (SID). Metadata týkající se přepisů řeči nabývají v rámci celé problematiky na významu. Stáváme se svědky formujícího se trhu s metadaty o emocích, zdraví, věku či dalších zásadních údajů o osobách, které jsou rozpoznatelné z řečového projevu. Mezi relativně novou oblast, která ale díky svojí relevanci v klíčových oblastech lidského konání zažívá neméně dynamický rozvoj, se řadí detekce stresu.

1.1 Identifikace stresu z řečového projevu

Stres ovlivňuje nejen způsoby každodenního uvažování člověka o světě, případně pak jeho osobní a pracovní návyky a výsledky, někdy ale dopadá přímo také na fyzický vzhled člověka a zdravotní stav. Jakkoli je pojem stres mnohoznačný, v současné době reprezentuje zastřešující pojmem zastupující zkušenosti člověka, které způsobují určitý druh napětí v lidském organismu. Abychom uvedli několik spouštěčů, víme, že kognitivní přetížení, fyzická bolest, vzrušení, deprese, intoxikace, infekce, strach a další faktory mohou vyvolat stresovou reakci. Vzhledem k tomu, že žijeme v informačním věku, stále více běžné kognitivní přetížení je důležitým faktorem ovlivňujícím naše rozhodovací procesy. Často se potýkáme se složitými problémy, které vyžadují soustředěné řešení, ale při tom jsme neustále zahlcováni přísunem konkurenčních podnětů, jsme rozptylování množstvím dalších vstupů. Pobyt v náročném prostředí tak může vést ke kritickým chybám v lidském úsudku a rozhodování, což nabývá na významu především v exponovaných oblastech, jakými

jsou např. chirurgie, pilotování letadla, autobusu, řízení letového provozu, či řízení procesu jaderného štěpení. Systémy, které umožní včas odhalit rizikovou situaci vedoucí ke kognitivnímu či emočnímu zahlcení se stávají životně důležitou oblastí vývoje ML. Detekce stresu v rámci ASR je nicméně náročný, v odborných kruzích dlouhá léta diskutovaný problém. Historicky největším problémem u hlubokého učení (deep learning; DL) byl nedostatek kvalitních referenčních dat, na základě kterých by bylo možné systém trénovat a dosáhnout tak uspokojivých predikcí o budoucích situacích.

1.2 Východiska výzkumu

V tomto příspěvku představujeme a popisujeme specifickou výzkumnou metodologii, která má za cíl shromáždit potřebná empirická data vhodná pro efektivní trénink hlubokých neuronových sítí v kontextu řečové zátěže. Předpokládáme, že modely neuronových sítí poháněné adekvátními datovými vstupy významně prohloubí možnosti detekce a klasifikace stresu v procesu ASR. Za tímto účelem byl pro české prostředí vyvinut tzv. Brněnský rozšířený test řeči a zátěže (BESST), který je rozšířenou adaptací původního Maastrichtského akutního zátěžového testu (MAST) Smeets a spol. (2012). Tato adaptovaná metodika optimalizuje proces sběru řečových výstupů osob uvedených do stresových situací. Pro tento účel bylo navrženo a otestováno experimentální prostředí, které zahrnuje nahrávání několika kamerami a mikrofony v situacích, kdy participanti vykonávají sérii úkolů vyvolávajících stres. V rámci prvního úkolu je navozován stres fyziologický, a to ponořením nedominantní ruky účastníka do ledové vody, zatímco je zaznamenán jeho volný řečový projev. Druhým kontextem je tzv. duální task založený na metodě RSPAN (reading span task) Daneman a Carpenter (1980), který zvyšuje kognitivní zátěž, zatímco účastník řeší a čte textový rébus. Kromě audio a video záznamu ze situace se snímá také elektrokardiograf prostřednictvím zařízení Faros 180 a galvanický kožní odpor prostřednictvím náramku Empatica E4. Subjektivní stavy účastníci sami reportují prostřednictvím vlastních odhadů zátěže (vyjádřeno na Likertově škále; málo 1 - 9 velmi), a to po každém úkolu. Psychologicky je stres určován pomocí škál vnímání stresu Perceived Stress Scale 14 (PSS14) Cohen a spol. (1983), State Trait Anxiety Inventory Y2 (STAI-Y2) Hedberg (1972) a NASA Task Load Experience (NASA TLX) Hart a Staveland (1988). Navrhovaná metodologie blíže popisovaná v tomto článku představuje funkční a škálovatelný nástroj pro sběr klíčových datových sad nezbytných pro detekci stresu pomocí technik DL. Právě škálovatelnost, ale např. také kulturní neutralita celého testu, kdy je zajištěno navozování stresu kulturně nezávislými způsoby, umožňuje přenesení celé metodologie do dalších výzkumných prostředí či kulturních kontextů a nabízí možnosti měření na dalších populacích jak v Evropském (např. Slovensko), tak celosvětovém měřítku (např. Japonsko).

2 Metoda - Brněnský rozšířený zátěžový a řečový test

Brněnský rozšířený zátěžový a řečový test (BESST) je modifikací Maastrichtského akutního zátěžového testu (MAST). Úpravy především usilují o maximalizaci řečové produkce mluvčího. V původní proceduře MAST mluvčí komentuje mentální aritmetický úkol (mental arithmetic task; MAT) a slovní zásoba je tak omezena pouze na mluvená čísla. V rámci adaptace BESST byly rozšířeny obě hlavní komponenty MAST, a to jak test vnoření ruky do ledové vody (Hand Immersion Task; HIT), tak MAT. HIT je vystavěn způsobem, kdy participanti dostanou za úkol slovně popisovat obrázky, které jsou jim prezentovány na monitoru před nimi. Pro tento účel bylo shromážděno 20 veřejných obrazů z Národní galerie ve Washingtonu DC[1], a 20 fotografií z webu Unsplash[2]. Oba zdroje poskytují tyto grafiky pod licencí Creative Common a tudíž zdarma k volnému použití. Každý obrázek je zobrazen na obrazovce po předem stanovenou (nicméně variabilní) dobu, která ale v žádném případě nepřesáhne 20 sekund na jeden snímek. Mentální aritmetika v úloze MAT je nahrazena upravenou verzí úlohy RSPAN, konkrétně jsou participantům prezentovány věty, ve kterých jsou některá slova nahrazena emotikony (obrázky), a slovní bloky tak představují určité rébusy. Tím, že požádáme účastníka, aby přečetl celou větu plynule, zaměstnáváme aktivní systém zpracování a vyhledávání lexikálního obsahu. Rébusy jsou zobrazovány participantovi formou plynule se vynořujících slov. Participant je požádán, aby věty nahlas přečetl, jakmile je rozluští. Cílem je zajistit spontánní a plynulou řečovou produkci. Rébusy obsahují dva až pět emotikonů a jejich náročnost postupně narůstá. Vzhledem k tomu, že je celý protokol BESST z velké části založen na původním protokolu MAST, provedené adaptace související s řečí byly diskutovány přímo s autorem MAST. Obecně řečeno, protokol BESST se skládá z následujících částí, které budou rozebrány níže:

1. Instrumentace a úvod
2. Získávání dat
3. De-briefing a de-instrumentace

Každý experimentální běh mapující jednoho participanta je veden dvěma instruktory a trvá přibližně jednu hodinu. Denní kapacita pro experiment je tedy cca devět lidí.

[1] https://www.nga.gov
[2] https://www.unsplash.com

2.1 Instrumentace a úvod

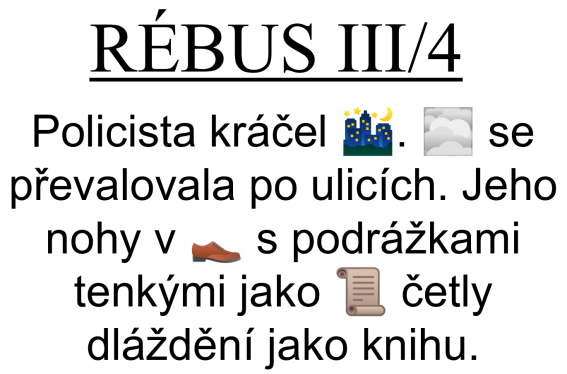
V první fázi jsou účastníci uvítáni a požádáni, aby podepsali formulář informovaného souhlasu, ve kterém vyjadřují, že do celého experimentu vstupují dobrovolně s vědomím zachování vlastní anonymity. Po podpisu souhlasů jsou připojeny měřící přístroje participantovi přímo na tělo (EKG, EDA a mikrofon) a zkontroluje se správná funkce (nahrávání) všech zařízení. Od tohoto okamžiku začíná záznam celé relace. Účastník vyplní první z auto-evaluačních dotazníků na vnímaný stress (PSS14). Získaná data z tohoto dotazníku slouží k odhadu obecné úrovně úzkosti ve srovnání s ostatními účastníky. Dále je uvedena prezentace s pokyny pro participanta a zároveň zde instruktor zodpoví případné dotazy.

2.2 Sběr dat

Sběr dat začíná třemi minutami relaxace, kdy jsou účastníci požádáni, aby zavřeli oči a odpočívali. V době relaxace ustanovujeme na základě fyziologického měření základní stav krátkodobé biologické odezvy participanta. Poté vyplňuje participant dotazník mapující aktuální stav úzkosti (STAI-Y2). STAI-Y2 se používá k odhadu úrovně úzkosti před samotným testem a bezprostředně po něm. Následuje zácvik samotného testu bez ponoření ruky do ledové vody. V této části se jedná o to, aby participant porozuměl všem úkolům obsaženým v nadcházející části. Po každém jednom úkolu zároveň účastník uvádí subjektivně vnímanou obtížnost tohoto úkolu na Likertově škále. Dále pak účastník plní střídavě úkoly HIT a MAT a celkově absolvuje 9 úkolů. Každý úkol MAT obsahuje jiný rébus s různým počtem slov a emotikonů a každý HIT obsahuje jinou sadu obrázků a délku ponoření ruky, příklad je zobrazen v obrázku 1. Bezprostředně po posledním úkolu participant opětovně vyplní STAI-Y2, kde je zaznamenána úroveň aktuálního stresu ihned po testu. Poté následuje 3-minutová relaxace. Po relaxačním období participant pro zhodnocení celkové obtížnosti celého experimentu BESST vyplní dotazník NASA-TLX. Na fotografiích 2 a 3 je zachycen reálný experiment za běhu.

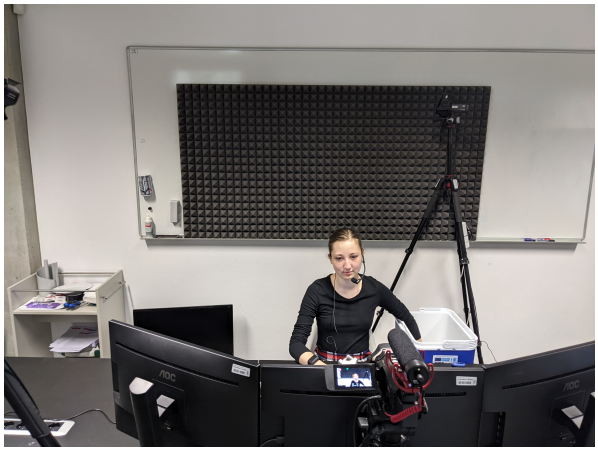
2.3 Debriefing a de-instrumentace

Po ukončení experimentu experimentátor položí několik připravených otevřených otázek v rámci polostrukturovaného rozhovoru s participantem a je proveden celkový debriefing. V této fázi je vzhledem k náročnosti experimentu kladen zásadní důraz na vyjasnění a zotavení účastníka z celého experimentálního běhu. Debriefingová část byla důkladně diskutována s expertem z oblasti psychologie. Na závěr jsou participantovi sejmuty veškeré měřící nástroje, je mu poděkováno a je odeslán pryč z místnosti.



Obr. 1: Příklad rébusu řešeného účastníky

Obr. 2: Instruktoři během experimentu



Obr. 3: Účastnice experimentu během HIT úkolu

3 Data

Výsledná datová sada se skládá z multimodálních datových toků u každé experimentální relace daného participanta. Do datového souboru jsou zahrnuty následující měření:

- Dvousvodový Elektrokardiogram (EKG) snímaný prostřednictvím zařízení Faros 180[3],
- Elektrodermální aktivita (EDA) snímaná prostřednictvím zařízení Empatica E4[4],
- Čtyři audio-vizuální streamy skrze kamery Panasonic HC-VX9805,
- Jeden vyhrazený zvukový stream z mikrofonu XY přes rekordér Zoom H4n[5],
- Jeden vyhrazený zvukový tok zaznamenaný pomocí náhlavní soupravy Shure SM-35XLR7[6] a Zoom H4n rekordér.

Kromě popsaných datových toků jsou schraňovány také výsledky z PSS14, pre- a post-experimentální měření na škále úzkosti (STAI-Y2) a post-experimentální dotazníky NASA-TLX. Každý záznam jednoho participanta obsahuje v průměru 55 GB dat. Průměrná čistá délka řeči je 11 minut na účastníka.

3.1 Výzkumný vzorek

Vzhledem ke skutečnosti, že sebraná data by měla být homogenní z hlediska jazykových dialektů a dalších možných odchylek způsobených socio-demografickými nebo zdravotními faktory, a protože samotný experiment pracuje s metodologií navozování fyziologického a psychologického stresu, účastníci vhodní pro participaci v experimentu by měli splňovat následující kritéria:

- žádné chronické onemocnění jako je cukrovka, vysoký krevní tlak, srdeční arytmie atd.
- v současné době neužívají žádné léky předepsané lékařem a nepodstupují žádný druh léčby
- jsou to rodilí mluvčí českého jazyka
- reprezentují cílovou věkovou kohortu

3.2 Etické aspekty výzkumu

V rámci takto strukturovaného měření je vzhledem ke skutečnosti, že je u participantů navozován stres, nezbytné vyřešit etické aspekty. V rámci popisovaného projektu se podařilo získat etický souhlas Etické komise pro výzkum. Souhlas pro námi uskutečněné měření byl získán od Etické komise Fakulty elektrotechnické, Ústav biomedicínského inženýrství. Součástí řízení pro získání etického souhlasu je vedle návrhu výzkumného protokolu také předložení konkrétní formy formuláře informovaného souhlasu, který má být před každým měřením participanta podepisován dotyčným participantem.

4 Automatická detekce stresu z mluveného projevu

Produkce řeči se skládá ze tří navazujících fází, které se dají definovat jako “konceptualizace”, “formulace” a “artikulace”. Ve fázi konceptualizace je utvářen záměr řečeného a je formován pre-verbální abstrakt řeči. Fáze formulování zahrnuje gramatickou, fonologickou a syntaktickou přípravu výpovědi. Fáze artikulace využívá soubor svalových pohybů uvnitř našich plic, krku a úst, aby byla vyprodukována požadovaná řeč. Tento proces tvoří východisko pro navržení postupů pro automatickou detekci stresu v mluveném projevu.

4.1 Produkce řeči a zpracování řečového signálu

Produkce řečového signálu tedy začíná v plicích, které vytlačují vzduch skrz hlasivky umístěné v hrtanu. Hlasové provazce, jsou-li sevřeny, mají schopnost periodicky otevírat nebo zavírat hrtan pro vytvoření základního tónu s řadou harmonií různých amplitud. V případě, že hlasivky nejsou omezeny, produkují aperiodický signál - šum. Zvuková vlna základního tónu pak prochází hltanem, ústní a nosní dutinou. Tyto dutiny tvoří rezonátor, kde různé části dutin, jako jsou rty, jazyk a patro upravují konečný zvuk řeči zeslabením nebo zesílením různých harmonií. Řečovou vlnu vytvořenou naším hlasovým traktem lze chápat jako systém postavený na filtrování zdroje zvuku. V teorii Source-Filter Fant (1970) je zdroj reprezentován plícemi, hrtanem a hlasivkami, které vytvářejí iniciální signál. Hltan a zbytek vokálního traktu potom představují uvažovaný filtr. Řečový signál je obvykle uložen jako sekvence řečových vzorků nazývaný vlnový průběh (waveform). Potřebné vzorky ve vlnovém průběhu jsou získávány vzorkováním analogového signálu ze záznamového zařízení, např. mikrofonu.

4.2 Extrakce metadat z řeči

Extrakce metadat z řeči má několik dobře zavedených kroků. Obecné schéma je znázorněno na Obr. 4.V dalších částech stručně popíšeme jednotlivé části.

[3] https://www.bittium.com/medical/bittium-faros

[4] https://www.empatica.com/research/e4/

[5] https://zoomcorp.com/en/us/handheld-recorders/handheld-recorders/h4n/

[6] https://www.shure.com/en-US/products/microphones/sm35

Obr. 4: Obecné schéma extraktoru metadat z řeči.

4.3 Předzpracování a extrakce příznaků

Než je zvukový záznam použit pro samotný trénink řečového modelu, obvykle je nejprve předzpracován způsobem, kdy jsou jeho příznaky (features) extrahovány ze syrových dat. Několik standardních kroků předběžného zpracování bývá přítomno z principu, např. Windowing a Diskrétní Fourierova transformace (DFT). Řeč, coby signál produkovaný lidským hlasovým traktem, má jako každý biologický systém svou vlastní setrvačnost. V rozmezí zhruba 25 milisekund může být vzorek řeči považován za stacionární. ”Windowing”vytváří sekvenci snímků s originálním průběhem zachycujícím řečovou produkci o délce právě 25 ms. Každý snímek je zde posunutý o 10 ms, aby se vytvořila sekvence snímků s přesahem 15 ms. Před provedením DFT se vyhladí okraje oken a odstraní se artefakty v DFT spektru aplikací Hammingova okan. Jako výsledek DFT získáme dvě složky: Amplitudové a Fázové spektrum. Fázové spektrum se dále neuvažuje, nicméně kvadratura amplitudového spektra pomáhá vygenerovat tzv. výkonové spektrum (Power spectrum). Mel-spektrum se pak získá vynásobením výkonového spektra filtry trojúhelníkového tvaru nazývanými Mel-filtry a sečtením koeficientů pod každým trojúhelníkem.

4.4 Příznaky související se stresem

Při snaze identifikovat typické příznaky reprezentované ve stresovém kontextu existuje pár variant ke zvážení. Patří mezi ně především:

4.4.1 Banky filtrů

Banky filtrů (Filterbank) reprezentují nejjednodušší a nejsnáze implementovatelné příznaky. Obvykle je za Filterbanky označováno Mel-spektrum, nicméně další filtry místo Mel-filtrů mohou být využity. Jednoduchost banek filtrů je často využívána u neuronových sítí, protože nepokládá žádné předpoklady u zdrojových dat. Neuronová síť se tedy může učit sama bez omezení reprezentace vyvozené z těchto příznaků.

4.4.2 Mel-Frekvenční cepstrální koeficienty

Mel-Frekvenční cepstrální koeficienty (MFCC) jsou zlatým standardem při zpracování řeči. Fungují poměrně dobře v různých podmínkách a jsou vytvořeny na základě banek filtrů. Jednotlivé kanály banek filtrů jsou de-korelovány provedením bodového součinu s bázemi Diskrétní kosinové transformace (DCT).

4.4.3 Prozodické příznaky

Prozodické příznaky reprezentují speciální druh příznaků, které mohou obsahovat různé nestandardní příznaky jako frekvence formantů, výšku tónu, intenzitu atd. Běžná sada příznaků pro analýzu prozódie je OpenSMILE Eyben a spol. (2010) extraktor příznaků.

4.4.4 Vysokoúrovňové příznaky

Příznaky založené na ASR - tyto vysokoúrovňové příznaky jsou založeny na ASR analýze mluvených projevů. Takovými příznaky mohou být kadence řečového projevu, průměrná délka slova nebo samohlásky, používání výplňkových slov, délka ticha mezi slovy atp. Vzhledem k povaze těchto příznaků, systém ASR pro cílový jazyk musí být k dipozici.

4.5 Modelování příznaků

Extrahované příznaky se obvykle používají k vytvoření vysoko-úrovňových modelů. Ty popisují statistickou reprezentaci sekvencí příznakových vektorů z řečových dat. Modely mohou popisovat řečová data na úrovni segmentů ve formě sumarizace nebo na úrovni jednotlivých příznakových rámců. Pro sumarizační modelování připadá v úvahu koncept i-vektor nebo x-vektor.

4.5.1 i-vektor

i-vektor model byl představený Najimem Dehakem a jeho kolegyDehak a spol. (2011) a pracuje s konceptem totální variability v prostoru příznaků. i-vektor je generativní model vycházející konceptuálně z Joint Factor analýzy (JFA) a Universal Background modelování - Gaussian Mixture Model (UBM-GMM) Reynolds a spol. (2000). Původně byl i-vektor model navržen pro identifikaci řečníka z nahrávky (Speaker Recognition; SRE) ale brzy se rozšířil i na další druh metadat.

4.5.2 x-vektor

Koncept sumarizačního vektoru k popisu požadované variability v řečovém segmentu byl brzy převeden do domény neuronových sítí. Sumarizace podle sekvence se ve světě neuronových sítí obvykle nazývá vtisknutí (embedding). David Snyder s kolegy představil koncept tzv. x-vektorů Snyder a spol. (2018). x-vektory reprezentují specifický řečový embedding pro rozpoznávání identity řečníka (SRE) z řečového segmentu.

4.6 Klasifikátory

Pro správnou klasifikaci výstupů vysokoúrovňových modelů je třeba použít klasifikátor. S ohledem na daný úkol, trénovací data a druh vysokoúrovňového modelu může být klasifikátorem cokoliv od jednoduchého systému založeného na vektorové vzdálenosti (cosine distance) až po pravděpodobnostní systém lineární diskriminační analýzy (Probabilistic Linear Discriminant Analysis; PLDA). Více informací o SRE a PLDA lze nalézt v zde Kenny (2010).

5 Závěr

V rámci toho příspěvku byla přiblížena problematika metodologie zaměřené na utváření vhodných vstupních dat pro trénink neuronových sítí v oblasti automatického rozpoznávání stresu v řečovém projevu. Příspěvek shrnuje současný stav problematiky ASR v oblasti detekce stresu a uvádí argumenty a návrhy pro další vývoj potřebných metod. Jednou z možných metod je protokol BESST, který coby rozšířená adaptace předchozích nástrojů (především MAST) umožňuje systematický sběr empirických dat potřebných pro trénink neuronových sítí. Protokol BESST je v rámci článku podrobně představen včetně popisu měřených proměnných a možných způsobů zpracování výsledných dat. Na základě navržené metodologie, která je škálovatelná a zároveň vzhledem k formě navození stresu také relativně kulturně nespecifická, je možné rozšířit sběry referenčních datasetů do dalších kulturních kontextů se zaměřením na specifické populace (např. Slovensko, Japonsko). Tímto způsobem je prospektivně možné kumulativně budovat potřebné datové sady, které mohou být využívány pro účely detekce stresu na základě hlasového projevu především v aplikační sféře.

Poděkování

Biomedicínský výzkumný hardware byl zapůjčen Ústavem biomedicínského inženýrství, Fakulty elektrotechniky a komunikačních technologií, VUT v Brně. Výzkum byl také realizován s podporou výzkumné laboratoře CEPCOS GREY lab, Psychologický ústav Filozofické fakulty Masarykovy univerzity v Brně.

Reference

Cohen, S., Kamarck, T. a Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior*, 24(4):385.

Daneman, M. a Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19(4):450–466.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P. a Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798.

Eyben, F., Wöllmer, M. a Schuller, B. (2010). Opensmile. V *Proceedings of the international conference on Multimedia - MM '10*. ACM Press.

Fant, G. (1970). *Acoustic theory of speech production*. No. 2. Walter de Gruyter.

Hart, S. G. a Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. V *Advances in Psychology*, str. 139–183. Elsevier.

Hedberg, A. G. (1972). Review of state-trait anxiety inventory. *Professional Psychology*, 3(4):389–390.

Kenny, P. (2010). Bayesian Speaker Verification with Heavy-Tailed Priors. V *Proc. The Speaker and Language Recognition Workshop (Odyssey 2010)*, str. paper 14.

Reynolds, D. A., Quatieri, T. F. a Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1):19–41.

Smeets, T., Cornelisse, S., Quaedflieg, C. W., Meyer, T., Jelicic, M. a Merckelbach, H. (2012). Introducing the maastricht acute stress test (mast): A quick and non-invasive approach to elicit robust autonomic

and glucocorticoid stress responses. *Psychoneuroendocrinology*, 37(12):1998–2008.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. a Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. V *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, str. 5329–5333.

Diferenciální evoluce s adaptací velikosti populace v závislosti na diverzitě

Radka Poláková
Slezská univerzita v Opavě
746 01 Opava, Na Rybníčku 626/1
Email: radka.polakova@fpf.slu.cz

Petr Bujok
Ostravská univerzita
701 03 Ostrava, Dvořákova 7
Email: petr.bujok@osu.cz

Abstrakt

V článku popisujeme nový mechanismus adaptace velikosti populace v algoritmu diferenciální evoluce. Navržený mechanismus je založen na lineárním snižování míry diverzity populace a dovoluje jak snížení velikosti populace tak i její zvýšení. Efektivitu několika variant algoritmu diferenciální evoluce s a bez adaptivního mechanismu jsme experimentálně porovnali na sadě testovacích funkcí vytvořených pro CEC 2014. Navíc jsme mechanismus porovnali s lineárním snižováním velikosti populace. Výsledky porovnání ukazují, že použití navrženého mechanismu je z hlediska efektivity hledání optima výhodné ve více než polovině testovaných úloh, naopak výsledky se implementací mechanismu zhorší jen zřídka.

1 Optimalizace diferenciální evolucí

Optimalizovat funkci, tj. nalézt bod globálního optima, lze matematickými prostředky. Tento proces je ale u některých funkcí, které např. nejsou diferencovatelné, obtížný nebo přímo nemožný. Avšak optimalizovat funkci lze také prostředky stochastickými, tedy s využitím stochastických algoritmů. Mezi tyto patří také algoritmus diferenciální evoluce (DE). Algoritmus byl poprvé předveden v Storn a Price (1997), práce má aktuálně více než 14000 citací.

Podle původního návrhu algoritmus diferenciální evoluce během výpočtu pro konkrétní funkci pracuje s konstantně nastavenými parametry. Od vzniku algoritmu byly vyvinuty mnohé jeho adaptivní verze. Jednou z nich je algoritmus LSHADE (Tanabe a Fukunaga, 2014), pro který autoři navrhli mechanismus lineárního snižování velikosti populace. Algoritmus LSHADE uspěl v optimalizační soutěži na kongresu CEC (Liang a spol., 2013, 2014). Mechanismus lineárního snižování velikosti populace se stal populárním a využívá jej mnoho verzí DE (Guo a spol., 2015; Awad a spol., 2016; Brest a spol., 2016, 2017), které vznikly po algoritmu LSHADE. Tyto verze se ukázaly efektivními, viz výsledky soutěží CEC - Liang a spol. (2015); Suganthan a spol. (2016); Awad a spol. (2017a). Z uvedeného plyne, že adaptace velikosti populace v algoritmu DE je důležitá. Vhodný výběr hodnoty parametru velikosti populace může podstatně zvýšit efektivitu algoritmu.

Náš návrh adaptace velikosti populace předložený v Poláková a spol. (2019) dovoluje jak zmenšování populace tak i její zvětšování a spočívá v udržování diverzity populace v míře adekvátní fázi výpočtu algoritmu, tj. na začátku výpočtu velkou, na konci výpočtu minimální a během výpočtu algoritmu lineárně se zmenšující. V tomto článku jsme se rozhodli se k popisu mechanismu vrátit, popsat jej v češtině a podat tak informace o něm širšímu publiku.

2 Diferenciální evoluce

Diferenciální evoluce (Storn a Price, 1997) je populační algoritmus pro globální optimalizaci. Pracuje s populací P množiny NP bodů. NP je velikost populace. Populace se během procesu hledání globálního optima vyvíjí. Prvky populace uvažujeme jako kandidáty na řešení. Populace je inicializována náhodně v celém prohledávaném prostoru $S = \Pi_{j=1}^{D}[a_j, b_j], a_j < b_j, j = 1, 2, \ldots, D$. D je dimenze problému. Po inicializaci populace bodů následuje opakování cyklu až do splnění podmínky k ukončení výpočtu. Ukončovací podmínka je často dána jako maximální možný počet výpočtů optimalizované funkce. V těle cyklu se k aktuální populaci P vytváří nová populace Q. Ke každému bodu x_i populace P je vytvořen y_i, tzv. pokusný bod. Jestliže platí, že $f(y_i) \leq f(x_i)$, kde f je optimalizovaná funkce, stává se prvkem nové populace Q pokusný bod y_i, pokud podmínka neplatí, je do populace Q vložen bod x_i.[1] Populace Q je na začátku každého běhu cyklu inicializována jako prázdná množina. Po vytvoření nové populace Q se tato populace Q stává populací P a není-li splněna podmínka ukončení algoritmu, cyklus se opakuje. Každý pokusný bod y_i je vytvořen s využitím operací mutace a křížení.

Mutací vzniká tzv. mutant vektor (bod v prostoru) v_i a to nejčastěji přidáním F-násobku rozdílu nebo rozdílů dvojice či několika dvojic nějakých bodů (prvků) populace k nějakému dalšímu bodu, tzv. základnímu bodu mutace. Vstupní parametr F ovlivňuje vzdálenost mutanta a základního bodu mutace. V DE

[1] Protože maximalizovat funkci g znamená minimalizovat funkci $-g$, můžeme se o optimalizaci bavit jako o minimalizaci.

existuje několik druhů mutace. Pokusný bod y_i je vytvořen ze dvou bodů, původního bodu populace x_i a mutanta v_i, křížením. V diferenciální evoluci je podle původního návrhu možné využít jeden ze dvou druhů křížení, binomické nebo exponenciální. Obě varianty křížení využívají vstupní parametr CR, který ovlivňuje podíl složek mutanta, které přecházejí do pokusného bodu. Kombinace mutace a křížení (tzv. DE-strategie), je často zkracována jako $DE/m/n/c$, kde m je použitá mutace, n je počet rozdílů (diferencí) využitých při vytváření mutanta a c je využitý typ křížení. DE-strategie společně s hodnotami parametrů F a CR se nazývá DE-nastavení.

3 Adaptace velikosti populace

Adaptace velikosti populace byla využita jako nástroj k řízení diverzity populace také v Arabas a spol. (1994). V práci je uvedena i myšlenka, že je-li populace příliš malá, může algoritmus předčasně konvergovat a naopak, když je populace příliš velká, může docházet k mrhání výpočetními zdroji. Autoři také zmiňují, že v různých fázích evolučního procesu může být výhodná jiná velikost populace.

Jeden z prvních mechanismů adaptace velikosti populace pro diferenciální evoluci byl uveden v Teo (2006). Autor navrhl dvě modifikace adaptivního mechanismu DESAP. V tomto mechanismu se s každým bodem populace ukládal ještě parametr velikosti populace, který se během výpočtu algoritmu adaptoval a po vytvoření celé nové generace populace se hodnoty tohoto parametru využily pro výpočet velikosti populace pro další průběh algoritmu. Jedna verze mechanismu DESAP pracuje s absolutní velikostí populace a druhá s relativní.

Dalších Z mechanismů adaptace velikosti populace v DE je mechanismus navržený v Brest a Maučec (2008). Algoritmus DE zde začíná pracovat s populací velkého rozsahu. Po určité části výpočtu se velikost populace zmenší na polovinu a to se opakuje až do konce běhu algoritmu. V Wang a Zhao (2013) a Zhu a spol. (2013) se velikost populace upravuje v závislosti na zlepšení či nezlepšení aktuálního řešení problému. V článku Salehinejad a spol. (2017) se pracuje s velikostí populace tak, že dojde-li v generaci k zlepšení řešení, velikost populace se může zmenšit nebo zůstává stejná. Když ke zlepšení nedojde, jsou do populace přidány nové body, či se velikost populace nemění.

Lineární snižování velikosti populace bylo navrženo k zlepšení efektivity algoritmu SHADE (Tanabe a Fukunaga, 2013), vznikl tak algoritmus LSHADE (Tanabe a Fukunaga, 2014). Tento způsob adaptace se objevil v mnoha DE-verzích, které byly navrženy po LSHADE, jedná se např. o iLSHADE (Brest a spol., 2016), jSO (Brest a spol., 2017).

V práci Awad a spol. (2017b) se pro snižování velikosti populace využívá následující schéma. NP se začíná snižovat až od poloviny výpočtu algoritmu. Pro každý bod je vypočítána jeho Eukleidovská vzdálenost od nejlepšího bodu, následně je populace podle těchto vzdáleností setříděna a pak, takto setříděná, je rozdělena do dvou stejně velkých částí. Populace se redukuje lineárně a odstraňují se body patřící do druhé části populace, té horší.

Další články z oblasti DE pracující s diverzitou populace jsou např. Weber a spol. (2009) nebo Yang a spol. (2013) nebo Yang a spol. (2014). Mechanismus uvedený v Gonuguntla a spol. (2015) se zaměřuje na úsporu výpočtů optimalizované funkce v zájmu vyšší efektivity algoritmu. V mechanismu navrženém pro algoritmu FDSADE (Tirronen a Neri, 2009) se zohledňuje diverzita hodnot optimalizované funkce v populaci. Obsáhlý přehled verzí DE zahrnující i verze adaptující velikost populace je možné nalézt v Neri a Tirronen (2010) a nebo např. v Piotrowski (2017).

4 Mechanismus adaptace velikosti populace založený na diverzitě

Diverzitu populace označíme DI a měříme následujícím vztahem

$$DI = \sqrt{\frac{1}{NP}\sum_{i=1}^{NP}\sum_{j=1}^{D}(x_{ij} - \overline{x}_j)^2}, \tag{1}$$

kde $\overline{x}_j$ je aritmetický průměr j-tých souřadnic aktuální generace populace bodů

$$\overline{x}_j = \frac{1}{NP}\sum_{i=1}^{NP} x_{ij}. \tag{2}$$

DI je odmocnina z průměrného čtverce vzdálenosti bodu populace a jejího těžiště $\overline{\mathbf{x}} = (\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_D)$. Je zřejmé, že $DI \geq 0$ a když se DI rovná 0, tak jsou všechny body v populaci totožné. Diverzitu počáteční generace populace bodů označme DI_{init}. Tuto DI_{init} použijme jako referenční hodnotu v definici relativní míry diverzity aktuální generace populace

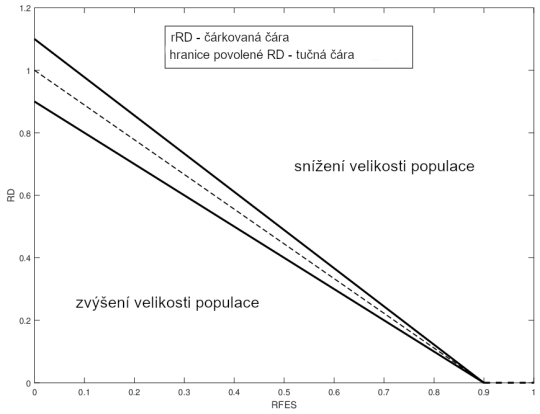
$$RD = \frac{DI}{DI_{init}}. \tag{3}$$

Relativní počet využitých vyhodnocení funkce je definováno následujícím výrazem

$$RFES = \frac{FES}{MaxFES}, \tag{4}$$

kde FES je aktuální počet využitých vyhodnocení optimalizované funkce a $MaxFES$ je celkový povolený počet vyhodnocení funkce během výpočtu. Velikost populace se mění v závislosti na aktuální relativní diverzitě populace. Relativní diverzitu RD jsem navrhli udržovat blízkou rRD. Tato vyžadovaná rRD je

lineárně se snižující od hodnoty 1 na začátku výpočtu až po hodnotu 0 na konci výpočtu algoritmu. Koncept blízkosti RD lineárně se snižující rRD je ilustrován na Obrázku 1. Velikost populace se mění v případě, že



Obr. 1: Lineárně se snižující rRD během výpočtu a hranice pro akceptovanou RD.

je RD vyšší než $1{,}1 \times rRD$, nebo naopak nižší než $0{,}9 \times rRD$. Toto pravidlo platí pro prvních devět desetin délky výpočtu algoritmu, v poslední desetině výpočtu je požadována nulová hodnota relativní diverzity. RD se vypočítává po každé nově vytvořené generaci populace. Když je splněna podmínka pro změnu velikosti populace, tj. relativní diverzita RD není v blízkosti rRD, NP vzroste o jedničku a do populace je přidán náhodně vygenerovaný bod z prohledávaného prostoru, to v případě, že RD je menší než $0{,}9 \times rRD$. Velikost populace NP se o 1 zmenší, tj. z populace je vyřazen nejhorší bod, v případě, že RD je větší než $1{,}1 \times rRD$.

Velikost populace je třeba udržovat v nějakém „rozumném“ intervalu, není možné, aby např. neomezeně rostla. Prohledávací proces tedy začíná s populací, jejíž velikost je rovna NP_{init} a je udržována v intervalu $\langle NP_{min}, NP_{max} \rangle$. Uvedené parametry jsme v závislosti na výsledcích předchozích experimentů (Poláková, 2017; Poláková a spol., 2017) nastavili na následující hodnoty, $NP_{init} = 50$, $NP_{min} = 8$, $NP_{max} = 5 \times D$.

5 Varianty diferenciální evoluce využité k testování navrženého mechanismu

K otestování efektivity navrženého mechanismu jsme zvolili 8 variant diferenciální evoluce. Jsou to: originální verze algoritmu DE, tři adaptivní verze, které se podle Das a Suganthan (2010), Das a Suganthan (2016) a Al-Dabbagh a spol. (2018) řadí mezi tzv. „state-of-the-art“ algoritmy, jmenovitě CoDE (Wang a spol., 2011b), EPSDE (Mallipeddi a spol., 2011) a jDE (Brest a spol., 2006). Pátým algoritmem zařazeným do našich testů je algoritmus *b6e6rl* (Tvrdík a Poláková, 2013), jedná se o efektivní verzi soutěživé DE navržené v Tvrdík (2006). Dalším algoritmem zahrnutým v testech je SHADE (Tanabe a Fukunaga, 2013), vítěz soutěže CEC 2013 (Loshchilov a spol., 2013). Modifikace tohoto algoritmu (Tanabe a Fukunaga, 2014) s implementovaným mechanismem lineárního snižování velikosti populace je také velmi úspěšným algoritmem (Liang a spol., 2014). Dalším algoritmem využitým v našich experimentech je algoritmus IDE (Tang a spol., 2015) a posledním algoritmem je algoritmus jSO (Brest a spol., 2017), který byl druhým algoritmem v pořadí [2] a současně nejlepší DE variantou na soutěži optimalizačních algoritmů na kongresu CEC 2017 (Awad a spol., 2017a).

Původní verze DE je v tomto článku využita s nejčastěji používanou *DE/rand/1/bin* strategií, nastavení vstupních parametrů je následující, $F = 0{,}8$, $CR = 0{,}5$. Algoritmus CoDE (Wang a spol., 2011b) vytváří pro každý bod vždy tři adepty na pokusný bod a jeden z nich pak vybírá na základě hodnoty optimalizované funkce v nich jako finální. CoDE vytváří adepty na pokusný bod třemi strategiemi, *DE/rand/1/bin*, *DE/rand/2/bin* a *DE/current-to-rand/1/-*, a dvojice parametrů k nim vybírá z následujících dvojic, $(1; 0{,}1)$, $(1; 0{,}9)$ a $(0{,}8; 0{,}2)$.

Algoritmus EPSDE (Mallipeddi a spol., 2011) využívá množinu strategií, množinu hodnot pro parametr F a množinu hodnot pro parametr CR. Každý bod populace má přiřazenu trojici parametrů (strategie, F, CR). Pokud je tato trojice úspěšná, tj. vytvoří pokusný bod lepší než bod původní, zůstává trojicí parametrů pro tento bod. V opačném případě, se buď generuje nová trojice nebo se vybírá náhodně některá trojice parametrů z úspěšných trojic, které se během celého výpočtu ukládají. Množina strategií obsahuje *DE/best/2/bin*, *DE/rand/1/bin* a strategii *DE/current-to-rand/1/-*, množina hodnot pro parametr F je $\{0{,}4; 0{,}5; 0{,}6; 0{,}7; 0{,}8; 0{,}9\}$ a množina hodnot pro parametr CR je $\{0{,}1; 0{,}2; 0{,}3; 0{,}4; 0{,}5; 0{,}6; 0{,}7; 0{,}8; 0{,}9\}$.

Algoritmus jDE (Brest a spol., 2006) je jednou z prvních a současně velmi efektivních adaptivních verzí DE. Pracuje se strategií *DE/rand/1/bin*. Při inicializaci populace se ke každému bodu populace inicializují i jeho vlastní parametry F a CR. Každý z nich se může s malou pravděpodobností před každým výpočtem pokusného bodu změnit - reinicializovat. V případě, že pomocí takto změněné dvojice parametrů se vytvoří úspěšný pokusný bod, tato nová dvojice parametrů F a CR se pak stává dvojicí parametrů příslušející tomuto prvku populace v dalším výpočtu. V případě opačném se příslušnému bodu navrací zpět jeho původní dvojice hodnot parametrů.

Soutěživá DE (Tvrdík, 2006) pracuje s několika nastaveními algoritmu DE, které mají na začátku

[2] Algoritmus jSO byl dokonce původně deklarován jako celkový vítěz soutěže.

výpočtu stejnou pravděpodobnost využití při tvorbě pokusného bodu. Čím je nastavení úspěšnější, tj. čím častěji vytváří pokusné body lepší než body původní, tím více se zvyšuje pravděpodobnost jejího využití. V případě, že některá z pravděpodobností je příliš malá, pravděpodobnosti se „resetují“ na navzájem rovnající se hodnoty. Algoritmus *b6e6rl* (Tvrdík a Poláková, 2013) pracuje se dvěma strategiemi (*DE/randrl/1/bin* a *DE/randrl/1/exp*), parametr F může nabývat jedné ze dvou hodnot $0{,}5$ a $0{,}8$ a parametr CR má tři možné hodnoty. Různé kombinace těchto hodnot F a CR vedou k 6 různým nastavením se strategií *DE/randrl/1/bin* a 6 různým nastavením se strategií *DE/randrl/1/exp*. V *b6e6rl* tedy soutěží těchto dvanáct nastavení DE.

SHADE (Tanabe a Fukunaga, 2013) je algoritmus postavený na základech algoritmu JADE (Zhang a Sanderson, 2009) s mutací *current-to-pbest/1* a archivem, do kterého se zapisují prvky populace, které byly v populaci přepsány svým pokusným bodem. V JADE se F a CR generují z Cauchyho resp. normálního rozdělení. První parametry těchto rozdělení (k využití v následující generaci) se počítají ze všech v aktuální generaci úspěšně využitých hodnot odpovídajících parametrů. V SHADE se využívá adaptace parametrů F a CR „zděděná“ z JADE, jen jsou zde navíc použity dvě kruhové paměti o velikosti H. V SHADE se z úspěšně využitých hodnot parametrů F a CR počítají opět první parametry - parametry polohy pro obě rozdělení, jen se zde uchovává posledních H takto vypočítaných hodnot. Při generování F a CR pro využití k vytvoření pokusného bodu se nejdříve náhodně zvolí jeden z indexů i, $i \in \{1, 2, 3, \ldots, H\}$, v kruhových pamětech a F je náhodné číslo z Cauchyho rozdělení s prvním parametrem z příslušné kruhové paměti s indexem i a CR je náhodné číslo z normálního rozdělení s prvním parametrem z příslušné kruhové paměti s indexem i. Všechny prvky v kruhových pamětech jsou inicializovány na hodnotu $0{,}5$. Druhý parametr obou využitých rozdělení je stále roven $0{,}1$.

Algoritmus IDE (Tang a spol., 2015) pracuje s populací bodů rozdělenou na dvě části, S (lepší) a I (horší). Velikost těchto dvou částí populace se za běhu algoritmu mění. Na začátku výpočtu je část S málo početná, naopak část I obsahuje celý zbytek populace. Takto zůstávají, co se týká velikosti, obě části populace až téměř do konce výpočtu algoritmu, následně velikost části populace S rychle roste, až je rovna velikosti populace. IDE využívá nově nadefinovanou mutaci. Dále i výpočet algoritmu je rozdělen na dvě části. V první části algoritmus využívá jako základní bod mutace náhodně vybraný bod populace. V druhé části běhu algoritmu se jako základní bod pro mutaci využívá aktuálně nejlepší bod populace. Navíc i využívaný typ mutace pracuje jinak pro část populace S a jinak pro část I. Parametry F a CR jsou nastaveny v závislosti na pořadí bodu odlišně pro každý bod populace, nejmenší hodnoty pro nejlepší bod a největší hodnoty pro nejhorší bod populace. Navíc je v IDE implementován mechanismus, kterým se předchází předčasné konvergenci.

Algoritmus jSO (Brest a spol., 2017) je vylepšené iLSHADE (Brest a spol., 2016), což je vylepšené LSHADE (Tanabe a Fukunaga, 2014). LSHADE je SHADE s lineárním snižováním velikosti populace. Algoritmus jSO má ve srovnání s LSHADE několik odlišných vlastností. jSO také využívá mechanismus lineárního snižování velikosti populace, ale výpočet začíná s $NP = 25 \times \sqrt{D} \times logD$ místo $18 \times D$. Parametr p, který využívá mutace *current-to-pbest/1*, se během výpočtu lineárně snižuje od $0{,}25$ do $0{,}125$, zatímco v LSHADE je p konstantní. Velikost kruhových pamětí je pro jSO nastavena na hodnotu 5. Ostatní vlastnosti jSO algoritmu najdete v Brest a spol. (2017).

6 Experimenty

Pro posouzení efektivity navrženého mechanismu adaptace velikosti populace jsme využili všech osmi variant DE popsaných v kapitole 5. V algoritmech DE, CoDE, EPSDE, jDE, *b6e6rl*, SHADE a IDE se neadaptuje velikost populace. Naproti tomu v algoritmu jSO je implementován mechanismus lineárního snižování velikosti populace. Z tohoto důvodu jsme pro každý z algoritmů DE, CoDE, EPSDE, jDE, *b6e6rl*, SHADE a IDE vytvořili další dvě verze. První s implementací mechanismu adaptace velikosti populace založeného na její diverzitě, tuto verzi jsme vždy označili předponou „d“, druhou s implementací lineárního snižování velikosti populace, tuto verzi jsme vždy označili předponou „L“. Takto jsme získali 21 různých algoritmů, tj. původních sedm algoritmů, sedm algoritmů s mechanismem diverzity (dDE, dCoDE, dEPSDE, atd.) a sedm algoritmů s lineárním snižováním velikosti populace (LDE, LCoDE, LEPSDE, LjDE, atd.). K takto získaným 21 algoritmům jsme ještě přidali jSO a algoritmus, který jsme získali z jSO odebráním mechanismu lineárního snižování velikosti populace a následnou implementací mechanismu řízení velikosti populace na základě diverzity, tento algoritmus jsme označili djSO.

Pro experimenty jsme zvolili testovací sadu 30 funkcí vytvořenou pro soutěž optimalizačních algoritmů uspořádanou v rámci kongresu CEC 2014 (Liang a spol., 2013). Maximální velikost populace při testech algoritmů s implementovaným námi navrženým mechanismem byla nastavena na hodnotu $NP_{max} = 5 \times D$. Nastavení ostatních parametrů každého z algoritmů jsme převzali z jejich původní definice. Všech 23 algoritmů jsme testovali na čtyřech úrovních dimenze, $D = 10, 30, 50, 100$. Pro každý z 23 testovaných algoritmů a každý z 30×4 optimalizačních problémů jsme provedli 51 opakování (51 běhů daného algoritmu). Tedy jsme získali 51 výsledků, tj. 51 minim, nalezených daným algoritmem k danému optimalizačnímu problému. Všech takto získaných 23×120 (2760) sad 51 výsledků jsme

zhodnotili níže uvedeným způsobem.

Všechny testované algoritmy byly implementovány v software Matlab 2010b a všechny výpočty byly provedeny na standardním PC s Windows 7 a konfigurací: Intel(R) Core(TM)i7-4790 CPU 3.6 GHz, 16 GB RAM. Všechny statistické výpočty byly provedeny v R software (R Core Team, 2015).

7 Výsledky experimentů

Nejdříve nás zajímal vliv implementace navrženého adaptivního mechanismu na efektivitu sedmi testovaných algoritmů, které v původní verzi velikost populace neadaptují. Porovnání algoritmů s pevně stanovenou velikostí populace pro celý běh algoritmu (DE, CoDE, EPSDE, jDE, *b6e6rl*, SHADE a IDE) s jejich variantami, které využívají mechanismus úpravy velikosti populace na základě diverzity (dDE, dCoDE, dEPSDE, djDE, *db6e6rl*, dSHADE a dIDE) je uvedeno v Tabulce 1, kde jsou shrnuty výsledky 840 dvouvýběrových Wilcoxonových testů. Z tabulky je zřejmé, že implementace adaptivního mechanismu založeného na diverzitě zvyšuje efektivitu testovaných algoritmů (kromě IDE, kde je počet vítězství 47) ve více než polovině z testovaných optimalizačních problémů. Počet vítězství d-verze je u všech algoritmů vyšší než počet proher. Pokud uvažujeme každou testovanou dimenzi samostatně, tak i zde platí téměř ve všech případech, že výher algoritmu s implementací navrženého adaptivního mechanismu je více než jeho proher, jedinou výjimkou je opět IDE v dimenzi $D = 100$, zde má dIDE o pět proher více než výher. Uvažujeme-li výhry a prohry pro všechny algoritmy dohromady, každou dimenzi zvlášť, ve všech třech vyšších dimenzích je podíl výher zhruba $2/3$. V dimenzi $D = 10$ je podíl výher sice menší, ale na druhou stranu je počet proher v této dimenzi minimální. Uvažujeme-li celkově všechny dimenze dohromady pro všechny testované algoritmy dohromady, je podíl výher verzí s navrženým mechanismem vyšší než 60 %, zatímco původní verze byly úspěšnější pouze zhruba v 11 % z testovaných úloh. Ve zhruba 26 % testovaných úloh nebyl rozdíl v efektivitě původní verze a jeho verze s implementací mechanismu adaptace velikosti populace na základě diverzity statisticky významný.

Dále nás zajímalo, zda je efektivnější lineární snižování velikosti populace nebo, zda k větší efektivitě námi testovaných verzí DE vede implementace mechanismu založeného na diverzitě populace. Porovnání algoritmů z těchto dvou skupin variant testovaných algoritmů je uvedeno v Tabulce 2, kde jsou shrnuty výsledky 960 dvouvýběrových Wilcoxonových testů.

Výsledky tohoto porovnání ukazují, že pro všechny algoritmy, kromě jSO, platí, že ve všech třech vyšších dimenzích je využití nově navrženého mechanismu adaptace velikosti populace výhodnější než využití mechanismu lineárního snižování velikosti populace. Fakt, že jSO si vede v tomto porovnání mnohem úspěšněji s lineárním snižováním velikosti populace, je pravděpodobně způsoben tím, že vhodné hodnoty parametrů tohoto algoritmu byly v průběhu jeho vývoje testovány právě s mechanismem lineárního snižování velikosti populace a změna mechanismu za jiný pravděpodobně velmi narušila jejich optimalitu.

alg.	dimenze	10	30	50	100	$\sum$
DE	# vítěz.	26	27	24	27	104
	# proher	0	0	0	0	0
	# $\approx$	4	3	6	3	16
jDE	# vítěz.	13	18	20	14	65
	# proher	1	2	5	11	19
	# $\approx$	16	10	5	5	36
IDE	# vítěz.	5	14	17	11	47
	# proher	1	4	4	16	25
	# $\approx$	24	12	9	3	48
SHADE	# wins	7	24	23	23	77
	# proher	0	0	3	5	8
	# $\approx$	23	6	4	2	35
b6e6rl	# vítěz.	9	19	20	18	66
	# proher	0	1	4	11	16
	# $\approx$	21	10	6	1	38
CoDE	# vítěz.	25	20	22	21	88
	# proher	0	4	5	4	13
	# $\approx$	5	6	3	5	19
EPSDE	# vítěz.	18	19	23	22	82
	# proher	0	2	4	6	12
	# $\approx$	12	9	3	2	26
$\sum$	# vítěz.	103	141	149	136	529
	# proher	2	13	25	53	93
	# $\approx$	105	56	36	21	218

Tab. 1: Počet vítězství a proher d-mechanismu proti pevně nastavené velikosti populace - výsledky na základě 840 výsledků výpočtu Wilcoxonova dvouvýběrového statického testu.

Když neuvažujeme jSO, které má pro dimenzi $D = 10$ podobné výsledky jako v ostatních dimenzích, pak pro dimenzi $D = 10$ jsou mezi testovanými algoritmy tři, pro které je využití námi navrhovaného mechanismu úpravy velikosti populace méně výhodné než využití lineárního snižování velikosti populace, jsou to jDE, SHADE a *b6e6rl*. Podíváme-li se však detailněji, jsou pro jDE a SHADE počty proher a výher v $D = 10$ srovnatelné a pouze pro algoritmus *b6e6rl* je počet proher v této dimenzi znatelně vyšší než počet výher. Pro ostatní čtyři testované algoritmy (DE, IDE, CoDE a EPSDE) je i v dimenzi $D = 10$ využití d-mechanismu výhodnější. Uvažujeme-li každý z testovaných algoritmů samostatně (dohromady ve všech dimenzích) jsou v našem porovnání pouze dva algoritmy (neuvažujeme-li jSO), pro které platí, že počet výher d-verze algoritmu není větší než polovina ze všech tes-

tovaných úloh. Jsou to jDE a SHADE. V obou těchto případech je však stále více výher d-verze algoritmu než-li jeho proher. Nyní se věnujme každé z dimenzí samostatně. V každé z dimenzí se objevuje zhruba 40 proher (z 240 případů) d-mechanismu. S rostoucím dimenzí se zvětšuje i počet výher d-mechanismu, naopak počet shod s rostoucí dimenzí klesá. Celkově v tomto porovnání d-mechanismus nad L-mechanismem vyhrál ve více než 60 % všech případů, prohrál ve zhruba 17 % případů a nevýznamný rozdíl ve výsledcích nastal ve zhruba 23 % případů.

alg.	dim.	10	30	50	100	$\sum$
DE	# vítěz.	23	30	29	27	109
	# proher	1	0	0	0	1
	# $\approx$	6	0	1	3	10
jDE	#vítěz.	8	12	16	20	56
	# proher	10	6	7	7	30
	# $\approx$	12	12	7	3	34
IDE	# vítěz.	17	27	28	28	100
	# proher	2	0	0	0	2
	# $\approx$	11	3	2	2	18
SHADE	#vítěz.	7	12	13	15	47
	# proher	8	8	7	10	33
	# $\approx$	15	10	10	5	40
b6e6rl	# vítěz.	3	16	23	25	67
	# proher	9	3	2	4	18
	# $\approx$	18	11	5	1	35
CoDE	# vítěz.	22	29	29	29	109
	# proher	0	1	1	1	3
	# $\approx$	8	0	0	0	8
EPSDE	# vítěz.	19	21	21	21	82
	# proher	0	2	2	5	9
	# $\approx$	11	7	7	4	29
jSO	# vítěz.	3	4	6	6	19
	# proher	16	14	17	17	64
	# $\approx$	11	12	7	7	37
$\sum$	# vítěz.	101	151	165	171	589
	# proher	46	34	36	44	160
	# $\approx$	93	55	39	25	211

Tab. 2: Počet vítězství a proher d-mechanismu proti L-mechanismu - výsledky na základě 960 výsledků výpočtu Wilcoxonova dvouvýběrového statického testu.

Výsledky všech 23 testovaných variant algoritmu diferenciální evoluce byly porovnány Friedmanovým testem, do testu vstupoval vždy medián ze všech 51 nalezených minim. Nulová hypotéza o shodě efektivity algoritmů byla pro všechny čtyři dimenze zamítnuta s p-hodnotou menší než $2,2 \times 10^{-16}$. Výsledky tohoto porovnání jsou zachyceny v Tabulce 3, jsou zde uvedena průměrná pořadí algoritmu pro každou dimenzi. Algoritmy jsou seřazeny od nejlepšího k nejhoršímu podle průměrného pořadí přes všechny 4 dimenze. V závorkách jsou uvedena pořadí algoritmů v rámci dané dimenze. Uvažujeme-li celkové pořadí algoritmů, vidíme, že až na jSO pro všechny testované algoritmy platí, že jejich d-verze je úspěšnější než originální algoritmus a také než jeho L-verze. Pro algoritmy SHADE, jDE, *b6e6rl*, EPSDE je pořadí verzí (d-verze, L-verze, fixní velikost populace). Pro algoritmy IDE, CoDE a DE je pak pořadí verzí (d-verze, fixní velikost populace, L-verze).

Z tabulky je také zřejmý celkový vítěz testu, je jím algoritmu jSO, který byl taky úspěšný na soutěži CEC 2017. jSO je v celkovém pořadí následován algoritmem djSO a třetí algoritmus v pořadí je d-verze algoritmu SHADE. Celkové pořadí algoritmů v Tabulce 3 odpovídá očekávané efektivitě optimalizačních algoritmů. jSO je úspěšný algoritmus, který byl vytvořen v roce 2017. SHADE a IDE jsou efektivní algoritmy, které vznikly několik málo let před jSO. Na druhé straně algoritmy EPSDE a CODE vznikly někdy okolo roku 2010.

Algoritmus	$D = 10$		$D = 30$		$D = 50$		$D = 100$		avg
jSO	6,9	(1)	4,4	(1)	4,3	(1)	4,5	(1)	5,0
djSO	9,9	(8)	5,4	(2)	5,2	(2)	5,8	(2)	6,6
dSHADE	8,5	(4)	6,2	(3)	6,1	(3)	6,0	(3)	6,7
LSHADE	7,5	(3)	6,6	(4)	6,9	(4)	7,1	(4)	7,0
dIDE	7,3	(2)	7,2	(5)	8,0	(5)	9,4	(8)	8,0
djDE	9,9	(9)	9,1	(8)	8,2	(6)	8,5	(5)	8,9
IDE	8,8	(6)	8,4	(6)	10,0	(9)	9,1	(6)	9,1
db6e6rl	10,7	(11,5)	8,7	(7)	8,5	(7)	10,1	(11)	9,5
dEPSDE	10,4	(10)	9,6	(9)	9,3	(8)	9,7	(10)	9,7
LjDE	8,6	(5)	10,4	(10)	10,3	(10)	10,3	(12)	9,9
SHADE	10,7	(11,5)	11,7	(13)	10,8	(12)	9,5	(9)	10,7
jDE	12,5	(15)	12,0	(14)	10,7	(11)	9,2	(7)	11,1
Lb6e6rl	9,2	(7)	10,7	(11)	11,6	(13)	13,5	(16)	11,3
b6e6rl	12,8	(16)	11,3	(12)	12,5	(14)	11,0	(13)	11,9
LEPSDE	13,8	(17)	13,9	(16)	13,1	(15)	11,9	(14)	13,2
EPSDE	14,7	(18)	13,0	(15)	13,7	(16)	12,2	(15)	13,4
LIDE	11,7	(13)	14,1	(17)	13,9	(17)	15,1	(17)	13,7
dCoDE	12,3	(14)	15,7	(18)	14,9	(18)	15,9	(18)	14,7
dDE	16,1	(20)	17,4	(20)	17,0	(20)	15,9	(19)	16,6
CoDE	18,0	(21)	17,2	(19)	16,3	(19)	17,8	(20)	17,3
LCoDE	16,0	(19)	20,1	(21)	20,6	(21)	20,6	(21)	19,3
DE	20,5	(23)	21,0	(22)	21,7	(22)	21,3	(22)	21,1
LDE	19,2	(22)	22,2	(23)	22,7	(23)	21,8	(23)	21,5

Tab. 3: Pořadí algoritmů v každé z dimenzí (podle výsledků Friedmanova testu) a průměrné pořadí algoritmů.

8 Závěr

V článku jsme navrhli nový adaptivní mechanismus pro úpravu velikosti populace v algoritmu diferenciální evoluce. Mechanismus je založen na řízení míry diverzity populace v diferenciální evoluci. V porovnání s oblíbeným a v poslední době často využívaným adaptivním mechanismem lineárního snižování velikosti populace v DE náš mechanismus nedovoluje pouze snižování velikosti populace, ale také její zvyšování. Námi navržený mechanismus jsme experimentálně porovnali se zmiňovaným lineárním snižováním velikosti populace.

Provedené testy nově navrženého mechanismu přinesly slibné výsledky. Takováto adaptace velikosti populace v DE, v porovnání s využitím mechanismu lineárního snižování velikosti populace, vede

v převažující části testovaných úloh k vyšší efektivitě procesu hledání optima.

V budoucnu bychom se chtěli zabývat implementací navrženého mechanismu do dalších evolučních algoritmů a také hledáním takové modifikace této adaptace, která zlepší i výsledky jednoho z aktuálně nejefektivnějších verzí algoritmu diferenciální evoluce, tedy algoritmu jSO.

Poděkování

Tento příspěvek je financován ze Strukturálních a investičních fondů Evropské unie OP VVV, z projektu „Zvýšení kvality vzdělávání na Slezské univerzitě v Opavě ve vazbě na potřeby Moravskoslezského kraje“, CZ.02.2.69/0.0/0.0/18_058/0010238.

Tento příspěvek je věnován našemu společnému školiteli doc. Josefu Tvrdíkovi, CSc., který nás před necelými dvěma lety opustil. Nebyl to jen náš školitel, ale i dlouholetý kamarád a především úžasný člověk. Je nám ctí, že jsme se mohli učit právě od něj.

Literatura

Al-Dabbagh, R. D., Neri, F., Idris, N. a Baba, M. S. (2018). Algorithmic design issues in adaptive differential evolution schemes: Review and taxonomy. *Swarm and Evolutionary Computation*, 43:284–311.

Arabas, J., Michalewicz, Z. a Maluwka, J. (1994). GAVaPS - a genetic algorithm with varying population size. V *Proceedings of IEEE Congress on Evolutionary Computation, 1994*, str. 73–78. IEEE.

Awad, N. H., Ali, M. Z., Liang, J. J., Qu, B. a Suganthan, P. N. (2017a). CEC17 special session on single objective numerical optimization single bound constrained real-parameter numerical optimization.

Awad, N. H., Ali, M. Z. a Suganthan, P. N. (2017b). Ensemble of parameters in a sinusoidal differential evolution with niching-based population reduction. *Swarm and Evolutionary Computation*, 39.

Awad, N. H., Ali, M. Z., Suganthan, P. N. a Reynolds, R. G. (2016). An ensemble sinusoidal parameter adaptation incorporated with l-shade for solving cec2014 benchmark problems. V *IEEE Congress on Evolutionary Computation 2016*, str. 2958–2965.

Brest, J., Greiner, S., Boškovič, B., Mernik, M. a Žumer, V. (2006). Self-adapting control parameters in differential evolution: A comparative study on numerical benchmark problems. *IEEE Transactions on Evolutionary Computation*, 10:646–657.

Brest, J. a Maučec, M. S. (2008). Population size reduction for the differential evolution algorithm. *Appl Intell*, 29:228–247.

Brest, J., Maučec, M. S. a Boškovič, B. (2016). iL-SHADE: Improved L-SHADE algorithm for single objective real-parameter optimization. V *IEEE Congress on Evolutionary Computation 2016*, str. 1188–1195.

Brest, J., Maučec, M. S. a Boškovič, B. (2017). Single objective real-parameter optimization: Algorithm jSO. V *IEEE Congress on Evolutionary Computation 2017*, str. 1311–1318.

Das, S. a Suganthan, P. N. (2010). Differential evolution: A survey of the state-of-the-art. *IEEE Transactions on Evolutionary Computation*, 15:4–31.

Das, S. a Suganthan, P. N. (2016). Recent advances in differential evolution: An updated survey. *Swarm and Evolutionary Computation*, 27:1–30.

Gonuguntla, V., Mallipeddi, R. a Veluvolu, K. C. (2015). Differential evolution with population and strategy parameter adaptation. *Mathematical Problems in Engineering*, 2015.

Guo, S.-M., Yang, C.-C., Hsu, P.-H. a Tsai, J. S.-H. (2015). A seff-optimization approach for L-SHADE incorporated with eigenvector-based crossover and successful-parent-selecting framework on CEC2015 benchmark set. V *IEEE Congress on Evolutionary Computation (CEC) 2015 Proceedings*, str. 1003–1010.

Holland, J. (1992). Genetic algorithms - computer programs that evolve in ways that resemble natural selection can solve complex problems even their creators do not fully understand. *Scientific American*, str. 66–72.

Liang, J. J., Qu, B. a Suganthan, P. N. (2013). Problem definitions and evaluation criteria for the CEC 2014 special session and competition on single objective real-parameter numerical optimization. [online] http://www.ntu.edu.sg/home/epnsugan/.

Liang, J. J., Qu, B. a Suganthan, P. N. (2014). Ranking results of CEC14 special session and competition on real-parameter single objective optimization. [online] http://www3.ntu.edu.sg/home/epnsugan/.

Liang, J. J., Qu, B., Suganthan, P. N. a Chen, Q. (2015). CEC15 competition on learning-based real-parameter single objective optimization.

Loshchilov, I., Stuetzle, T. a Liao, T. (2013). Ranking results of CEC13 special session and competition on real-parameter single objective optimization. [online] http://www3.ntu.edu.sg/home/epnsugan/.

Mallipeddi, R., Suganthan, P. N., Pan, Q. K. a Tasgetiren, M. F. (2011). Differential evolution algorithm with ensemble of parameters and mutation strategies. *Applied Soft Computing*, 11:1679–1696.

Neri, F. a Tirronen, V. (2010). Recent advances in differential evolution: A survey and experimental analysis. *Artificial Intelligence Review*, 33:61–106.

Piotrowski, A. P. (2017). Review of differential evolution population size. *Swarm and Evolutionary Computation*, 32:1–24.

Poláková, R. (2017). Controlling population size in differential evolution by diversity mechanism. V *International Conference on Artificial Intelligence and Soft Computing: LNAI 10245 Artificial Intelligence and Soft Computing - Part 1*, str. 408–417, Heidelberg. Springer.

Poláková, R., Tvrdík, J. a Bujok, P. (2017). Adaptation of population size according to current population diversity in differential evolution. V *2017 IEEE Symposium Series on Computational Intelligence (SSCI) Proceedings*, str. 2627–2634. IEEE.

Poláková, R., Tvrdík, J. a Bujok, P. (2019). Differential evolution with adaptive mechanism of population size according to current population diversity. *Swarm and Evolutionary Computation*, 50(100519).

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Salehinejad, H., Rahnamayan, S. a Tizhoosh, H. R. (2017). Self-adaptive differential evolution algorithm using population size reduction and three strategies. *Applied Soft Computing*, 52:812–833.

Storn, R. a Price, K. (1997). Differential evolution - A simple and efficient heuristic for global optimization over continuous spaces. *J. Global Optimization*, 11:341–359.

Suganthan, P. N., Ali, M. Z. a Awad, N. H. (2016). CEC16 special session on single objective numerical optimization single parameter - operator set based case.

Tanabe, R. a Fukunaga, A. (2013). Success-history based parameter adaptation for differential evolution. V *IEEE Congress on Evolutionary Computation 2013*, str. 71–78.

Tanabe, R. a Fukunaga, A. (2014). Improving the search performance of SHADE using linear population size reduction. V *IEEE Congress on Evolutionary Computation 2014*, str. 1658–1665.

Tang, L., Dong, Y. a Liu, J. (2015). Differential evolution with an individual-dependent mechanism. *IEEE Transactions on Evolutionary Computation*, 19:560–574.

Teo, J. (2006). Exploring dynamic self-adaptive populations in differential evolution. *Soft Computing*, 10:673–686.

Tirronen, V. a Neri, F. (2009). Differential evolution with fitness diversity self-adaptation. *Nature-Inspired Algorithms for Optimization*, str. 199–234.

Tvrdík, J. (2006). Competitive differential evolution. Matoušek, R. a Ošmera, P. (zost.), V *MENDEL 2006, 12th International Conference on Soft Computing*, str. 7–12, Brno.

Tvrdík, J. a Poláková, R. (2013). Competitive differential evolution applied to CEC 2013 problems. V *IEEE Congress on Evolutionary Computation 2013*, str. 1651–1657.

Wang, H., Rahnamayan, S. a Wu, Z. (2011a). Adaptive differential evolution with variable population size for solving high-dimensional problems. V *IEEE Congress on Evolutionary Computation*, str. 2626–2632.

Wang, X. a Zhao, S. (2013). Differential evolution algorithm with self-adaptive population resizing mechanism. *Mathematical Problems in Engineering*.

Wang, Y., Cai, Z. a Zhang, Q. (2011b). Differential evolution with composite trial vector generation strategies and control parameters. *IEEE Transactions on Evolutionary Computation*, 15:55–66.

Weber, M., Neri, F. a Tirronen, V. (2009). Distributed differential evolution with explorative–exploitative population families. *Genetic Programming and Evolvable Machines*, 10:343–371.

Yang, M., Cai, Z., Li, C. a Guan, J. (2013). An improved adaptive differential evolution algorithm with population adaptation. V *GECCO '13 Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation*, str. 145–152.

Yang, M., Li, C., Cai, Z. a Guan, J. (2014). Differential evolution with auto-enhanced population diversity. *IEEE Transactions on Cybernetics*, 45:302–315.

Zhang, J. a Sanderson, A. C. (2009). JADE: Adaptive differential evolution with optional external archive. *IEEE Transactions on Evolutionary Computation*, 13:945–958.

Zhu, W., Tang, Y., Fang, J. a Zhang, W. (2013). Adaptive population tuning scheme for differential evolution. *Information Sciences*, 223:164–191.

Prediktory vizuálnej predstavivosti: senzorická senzitivita, všímavosť a osobnostné dimenzie

Alexandra Ružičková (1), Lenka Jurkovičová (2, 3), Jan Páleník (1) a Vojtěch Jurík (1)

1 Psychologický ústav, Filozofická fakulta, Masarykova univerzita, Arna Nováka 1, Brno
2 CEITEC Masarykova Univerzita, Kamenice 5, Brno
3 1. neurologická klinika, Lékařská fakulta, Masarykova univerzita, Pekařská 53, Brno

alexandra.ruzickova@mail.muni.cz, jurkovicova.lenka@mail.muni.cz, jan.palenik@mail.muni.cz, jurik.vojtech@mail.muni.cz

Abstrakt

Vizuálna predstavivosť, tj. vnímanie obrazových predstáv v mysli, je diskutovaná nielen ako adaptívny kognitívny mechanizmus pre efektívne fungovanie ľudského organizmu, ale zároveň aj ako užitočný nástroj v oblastiach psychológie a kognitívnych vied, akými sú trénovanie pamäti, terapia alebo vývoj umelej inteligencie. V rámci tohto príspevku sme empiricky skúmali vplyv možných prediktorov na živosť vizuálnej predstavivosti, konkrétne to boli senzorická senzitivita, všímavosť (mindfulness) a osobnostné dimenzie Big Five. Dáta boli zozbierané v rámci dotazníkového šetrenia, kde bol respondentom prezentovaný Dotazník živosti vizuálnej predstavivosti, Glasgowský dotazník senzorickej senzitivity, Freiburgský inventár všímavosti, a Päťfaktorový osobnostný inventár NEO-FFI. Analýza dát vykonaná na odpovediach od 92 participantov ukázala ako najsilnejší prediktor všímavosť. Signifikantnými prediktormi boli aj senzorická senzitivita a otvorenosť voči skúsenosti. Výsledky štúdie okrem iného naznačujú, že tréningom všímavosti by mohlo byť možné zlepšovať úroveň vizuálnej predstavivosti.

1 Úvod

Pomerne veľkú časť kôry ľudského mozgu zaberajú centrá zodpovedné za vizuálnu percepciu vnemov (Grill-Spector & Malach, 2004). Aktivita týchto oblastí pri percepcii sa zároveň značne zhoduje s aktivitou mozgu pri vizuálnom imaginovaní (Dijkstra et al., 2017; Lee, Kravitz & Baker, 2012). Kvôli primárnosti vizuálneho spracovávania informácií sa v súčasnosti spomedzi všetkých ľudských zmyslov skúma najviac vizuálna percepcia a vizuálna predstavivosť (VP), ktorá je s ňou prepojená (Pearson, 2019).

VP je top-down proces umožňujúci opätovne manipulovať so stimulmi, ktoré sme už raz percipovali bottom-up procesmi a vnímať tak "obrazy v mysli" (Dijkstra et al., 2017). Na sformovanie vnemu bottom-up procesmi a jeho moduláciu procesmi top-down má nezanedbateľný vplyv zameranie pozornosti, emočné spracovávanie podnetov (Comte et al., 2016) alebo úroveň excitability vizuálneho kortexu mozgu (Reinhart et al., 2016). V našom výskume sme sa pokúsili preskúmať vplyv vybraných prediktorov zakladajúcich sa na uvedených psychických procesoch. Výsledkami sme chceli prispieť k diskusii o VP, široko využívanej napríklad v terapii (Curtis, 2016), pri trénovaní pamäte (napr. technika pamäťového paláca; Huttner & Robbert, 2018), v robotike (Di Nuovo & Cangelosi, 2015) alebo pri výskume výpovedí očitých svedkov (Măirean, 2015).

2 Prediktory vizuálnej predstavivosti

Živosť vizuálnej predstavivosti (ŽVP) sa pohybuje na spektre od afantázie (kompletnej neschopnosti vytvárať vizuálne predstavy) po hyperfantáziu (tzv. eidetickú predstavivosť; kedy sa živosť predstáv prakticky rovná živosti skutočného percipovaného stimulu; Zeman et al., 2020). Predstavy môžu byť ďalej dobrovoľné − vyvolané zámerne, alebo nedobrovoľné ako halucinácie a sysnestetické predstavy (Pearson & Westbrook, 2015).

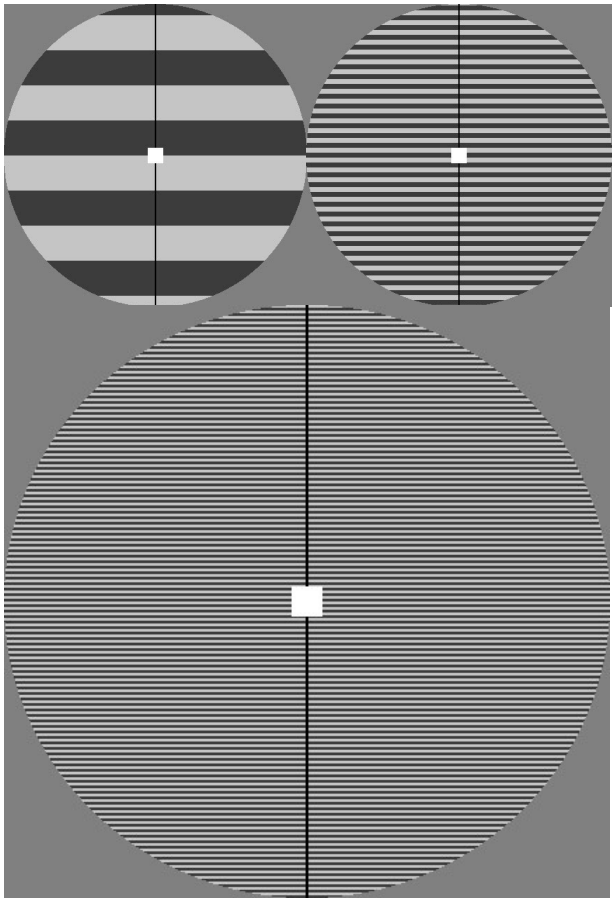
Niekoľko výskumov na synestetikoch alebo afantastikoch poukazuje na spojitosť medzi intenzitou vnímania podnetov, resp. senzorickou senzitivitou (SS) a ŽVP (Dance, Ward & Simner, 2021a; Dance, Ward & Simner, 2021b). Synestézia a afantázia sa tiež spájajú so zvýšeným množstvom autistických čŕt (Dance, Ward & Simner, 2021a) a viac autistických čŕt v bežnej populácii tesne koreluje so zvýšenou SS (Robertson, 2012). SS môže byť zvýšená v dvoch smeroch. Hypersenzitivita predstavuje citlivé vnímanie a zvýšenú

pozornosť voči stimulom, ktoré môžu viesť až k pocitom diskomfortu a zahltenia (Robertson, 2013). Typická je zle emočne regulovaná, neprimeraná reaktivita voči stimulom (Herbert, 2018). U hyposenzitivity je to naopak − reakcie na stimuly sú podpriemerné, ale tendencia dopĺňať si stimuláciu umelo a venovať tak podnetom viac pozornosti sa tu objavuje tiež. Paradoxne platí, že hypersenzitivita a hyposenzitivita sa vyskytujú spoločne (Robertson, 2013). Poškodená pozornosť, vyššia SS a poruchy autistického spektra (PAS) sa spolu často vyskytujú pri ADHD (Bijlenga et al, 2017).

Výskum SS a VP je v súčasnosti zasadený do výskumu excitability vizuálneho kortexu mozgu (Dance, Ward & Simner, 2021b; Keogh, Bergmann & Pearson, 2020). Autori Dance, Ward a Simner (2021b) skúmali ako súvisí SS, VP a miera excitability mozgu pomocou dotazníkov a *Pattern Glare tasku*. *Pattern Glare task* je zraková úloha, ktorá je niektorými autormi považovaná za spoľahlivý indikátor excitability primárneho vizuálneho kortexu mozgu (viz Obr. 1). Výsledky tohto výskumu spájajú nižšiu excitabilitu vizuálneho kortexu s nižšou SS a horšou VP. Nižšia senzorická senzitivita tu bola participantmi reportová a vyplývala tiež zo slabšej reakcie na *Pattern Glare task* (nižšia excitabilita vizuálneho kortexu).

Oproti tomu, autori Keogh, Bergmann a Pearson (Keogh, Bergmann & Pearson, 2020) dali participantom úlohu na predstavivosť a excitabilitu ich vizuálneho kortexu V1-V3 skúmali funkčnou magnetickou rezonanciou (fMRI), transkraniálnou magnetickou stimuláciou (TMS) a transkraniálnou stimuláciou priamym prúdom (tDCS). Na základe toho určili, že excitabilita vizuálneho kortexu je počas *resting state* (pokojový stav) nižšia u ľudí s lepšou VP. Autori podotýkajú, že participanti s nižšou excitabilitou vizuálneho kortexu počas *resting state* boli schopní pri imaginovaní zvýšiť aktivitu svojho vizuálneho kortexu viac ako participanti s vysokou počiatočnou úrovňou excitability vizuálneho kortexu a odkazujú na štúdiu, ktorá tvrdí, že počas VP rastie excitabilita vizuálneho kortexu (Cattaneo et al., 2011).

Ľudská psychika má prediktívnu povahu (Anālayo, 2019). Vnútorné konštruovanie sveta (top-down) je faktor vplývajúci na percipovanie senzorického zážitku (bottom-up). Všímavosť (mindfulness) je spôsob vnímania sveta a seba samého založený na neposudzujúcej pozornosti venovanej stimulom a kontrole emočných reakcií na nich (Takahashi et al., 2020). Zmienené je hlavná charakteristika konceptu všímavosti a tiež rozdiel medzi všímavým a senzoricky senzitívnym vnímaním sveta. Ďalší rozdiel medzi všímavosťou a SS je v možnosti trénovania všímavého vnímania podnetov, vedúceho ku všeobecnému zlepšeniu kognitívnych procesov, napr. pozornosti (Gallant, 2016) alebo k zlepšeniu emočnej regulácie (Chiesa, Calati, & Serretti, 2011). Avšak podobne ako to je v prípade SS, aj všímavosť spočíva v intenzívnom vnímaní podnetov a súvisí so živšou VP.



Obr. 1. Ukážka *Pattern Glare tasku* s kruhovými terčami. Pri *Pattern Glare tasku* sa používajú široké (kontrolné) pruhy (0,5 cpd [cycle per degree]; vpravo hore), stredne široké pruhy (3 cpd; vľavo hore) a pod nimi sú vyobrazené najužšie pruhy (11 cpd; Braithwaite, Mevorach & Takahashi, 2015).

Veľmi podobné charakteristikám konceptu všímavosti sú charakteristiky osobnostnej dimenzie podľa *Big Five*, tzv. otvorenosti. Otvorenejší ľudia sú, okrem iného, vnímavejší voči vonkajším aj vnútorným stimulom (McCrae & Costa, 1989) a zároveň sú schopní lepšie regulovať vlastné emócie (Morawetz, Alexandrowicz & Heekeren, 2017). Emočná regulácia sa spája s ďalšími dvoma dimenziami *Big Five* − neuroticizmom a extroverziou. Neuroticizmus je asociovaný so zlou emočnou reguláciou − horšou schopnosťou kognitívne kontrolovať, inhibovať či prehodnocovať negatívne

emócie (Yang et al., 2020). Extroverzia umožňuje častejšie zažívanie pozitívnych emócií, pretože sa pravdepodobne spája s lepšou schopnosťou kontrolovať vlastné emócie (Finley & Schmeichel, 2019). Aktuálny výskum naznačuje spojitosť medzi introverziou, afantáziou a autistickými črtami, a tiež spojitosť medzi hyperfantáziou a otvorenosťou (Milton et al., 2021). Neuroticizmus pravdepodobne neovplyvňuje ŽVP priamo, avšak je možné, že s ňou súvisí nepriamo (McDougall & Pfeifer, 2012).

3 Zhrnutie a ciele výskumu

Cieľom nášho výskumu (Ružičková, 2022) bolo zostaviť lineárny regresný model z možných prediktorov ŽVP. Ako prediktory ŽVP sme si na základe dostupnej literatúry zvolili senzorickú senzitivitu, všímavosť a tri osobnostné dimenzie Big Five − otvorenosť, neuroticizmus a extroverziu. Očakávali sme, že živšiu VP by mohla predikovať vyššia SS, rozvinutejšia všímavosť, nižší neuroticizmus a extroverzia. Zaujímalo nás, aký bude rozdiel medzi rýdzo dispozičnými možnými prediktormi ŽVP (SS, otvorenosť, neuroticizmus, extroverzia) a možným prediktorom "všímavosť", ktorá je trénovateľná. Zároveň sme chceli preskúmať, ako silno budú ŽVP predikovať koncepty súvisiace so zameriavaním pozornosti (SS, všímavosť, otvorenosť) s excitabilitou vizuálneho kortexu (SS) alebo ako jej mieru ovplyvní úroveň regulácie emočného prežívania.

4 Metódy

4.1. Participanti

Výskumný súbor pozostával z 92 participantov (49 žien, 39 mužov, 4 pohlavie neuviedli; M vek = 24,86, SD = 5,05). Väčšina participantov (N = 88) žije dlhodobo v Českej republike, traja uviedli ako miesto svojho trvalého pobytu Slovenskú republiku a jeden Belgicko. Všetci participanti sa na pokročilej úrovni dorozumievali v češtine. Participanti boli do výskumu pozvaní v rámci medzinárodného výskumného projektu zameraného na výskum vedomia COST Action (CA18106 − The neural architecture of consciousness), ktorému boli prispôsobené kritériá vylučujúce účasť. Participanti boli neurotypickí, so zdravým/ na normu korigovaným zrakom a sluchom. Zároveň museli spĺňať kritériá pre vyšetrenie v magnetickej rezonancii a kvôli absolvovaniu *Pattern Glare tasku* nesmeli trpieť migrénou. Projekt bol schválený Etickou komisiou pre výskum Masarykovej univerzity pod identifikačným kódom EKV-2020-094 a prebiehal vo výskumnom centre CEITEC Masarykova univerzita pod názvom *Paměť a vědomí*.

4.2. Nástroje

Výskum bol realizovaný formou online dotazníkového šetrenia a použité dotazníky boli celkom štyri: Dotazník živosti vizuálnej predstavivosti (*Vividness of Visual Imagery Questionnaire–2*; VVIQ–2; Marks, 1995), Glasgowský dotazník senzorickej senzitivity (*Glasgow Sensory Questionnaire*; GSQ; Robertson & Simmons, 2013), Freiburský inventár všímavosti (*Freiburg Mindfulness Inventory*; FMI; Wallach et al., 2006) a NEO päťfaktorový osobnostný inventár (*NEO Five-Factor Inventory*; NEO-FFI; McCrae & Costa, 1989).

4.3. Procedúra

Nábor participantov prebiehal pomocou inzerátu na sociálnej sieti a v univerzitnom informačnom systéme, a tiež formou informačného e-mailu poslaného vybraným osobám registrovaným v databáze dobrovoľníkov pre účasť na projektoch Centra neurověd CEITEC MU. Participanti sa do výskumu *Paměť a vědomí* väčšinou zapájali prostredníctvom náboru na sociálnej sieti. V prípade registrácie do výskumu bol participantom ďalším e-mailom poslaný odkaz na sadu dotazníkov preložených do češtiny, ktoré mali pred fyzickou návštevou vyplniť z domu. Spolu s nimi obdržali inštrukcie ohľadom výskumu a informovaný súhlas s účasťou vo výskume a spracovaním osobných údajov. Druhá časť výskumu prebiehala na CEITEC MU. Po úspešnom absolvovaní celého merania boli participanti odmenení čiastkou 1000 Kč.

4.4. Spracovanie a analýza dát

Dáta boli spracované prostredníctvom matematického software R-4.1.2 (R Core Team, 2021). Ako metódu na preskúmanie vzťahov medzi premennými sme si zvolili lineárnu regresnú analýzu. Predpoklady, ktoré sme pre využitie lineárneho regresného modelovania overovali boli: dostatočnosť veľkosti výskumnej vzorky; odľahlé pozorovania; normalita rozloženia premenných; korelácie medzi premennými; rozloženie reziduálov výsledného lineárneho regresného modelu.

5 Výsledky

Po vypočítaní korelácií medzi premennými sme z analýz vyradili osobnostnú dimenziu neuroticizmus, nakoľko nebol splnený predpoklad jeho korelácie s ŽVP (r = -0,04; p = 0,73). U extraverzie sa ukázalo, že ŽVP predikuje významne iba v samostatnom modeli (beta = 0,27, p < 0,01; 95% CI [0,07; 0,48]) a v kombinácii s ostatnými prediktormi nie je významným prediktorom ŽVP. Výsledný lineárny regresný model zahŕňal prediktory: senzorická senzitivita, všímavosť a otvorenosť. Spolu tieto prediktory vysvetľovali 19 % rozptylu predikovanej premennej (ŽVP) a celkovo bol regresný model štatisticky významný (F(3, 88) = 7,99; p < 0,001). Kompletný dátový report je možné nájsť v bakalárskej práci Alexandry Ružičkovej (2022). Prehľad výsledkov lineárnej regresie je uvedený v Tab. 1.

Tab. 1. Výsledný lineárny regresný model. Tabuľka zobrazuje mieru vplyvu prediktorov (FMI = Freiburský inventár všímavosti; GSQ = Glasgowský dotazník senzorickej senzitivity; dimenzia dotazníka NEO-FFI: NEOotvor = Otvorenosť) na závislú premennú (VVIQ; Dotazník živosti vizuálnej predstavivosti).

		VVIQ		
Prediktory	*Odhad*	*Beta*	*CI*	*p hodnota*
Konštanta	1,91	0,00	-0,19 – 0,19	0,913
GSQ	0,32	0,22	0,03 – 0,41	**0,022**
FMI	1,09	0,29	0,09 – 0,49	**0,005**
NEOotvor	0,89	0,21	0,01 – 0,41	**0,041**
Pozorovania 92				
R^2/R^2 adjustované 0,214/0,187				

6 Diskusia

Cieľom nášho výskumu bolo zistiť mieru vplyvu vybraných prediktorov na ŽVP. VP je nástroj hojne využívaný v terapeutickej praxi (Curtis, 2016). Od dobrej VP závisí efektívne uplatňovanie pamäťových techník (Huttner & Robbert, 2018). Výskum v oblasti robotiky učí umelú inteligenciu predstavivosti, aby bola schopnejšia orientovať sa v sociálnej interakcii (Di Nuovo & Cangelosi, 2015). Študovanie vplyvu predstavivosti je rovnako dôležité v situáciách, kedy sa človek stáva svedkom kriminálnych činov (Măirean, 2015). Ukazuje sa, že za určitých podmienok by mohla úroveň živosti VP skresľovať spomienky na prežité udalosti.

Naším príspevkom k týmto poznatkom je, že živšiu VP môžeme očakávať u ľudí senzoricky senzitívnych, všímavých a otvorených. Najsilnejší predpoklad je u ľudí, ktorí sú všímaví (mindful). Z toho plynie, že VP je možné do istej miery zlepšovať tréningom všímavého vnímania sveta. Druhý záver, ktorý je z našich výsledkov možné vyvodiť je, že VP ovplyvňujú skôr prediktory súvisiace s venovaním zvýšenej pozornosti, resp. s intenzívnejším vnímaním podnetov. Premenné, ktoré sa spájajú primárne s emočnou reguláciou (neuroticizmus, extroverzia) neboli s ŽVP v štatisticky významnom vzťahu.

Medzi limity nášho výskumu patrí fakt, že sme pre analýzu mali prístupné dáta iba z vybraných dotazníkov pre výskumný projekt *Paměť a vědomí*. Pretože sme nemali dotazník zisťujúci úroveň autistického kvocientu (AQ), nedokážeme spoľahlivo tvrdiť, že napr. senzoricky senzitívni participanti pochádzali z bežnej populácie, hoci boli psychiatrické poruchy jedným z vylučovacích kritérií. Participanti si nadpriemerného AQ nemuseli byť vedomí. Pri tomto type výskumu − tj. dotazníkovom šetrení − je tiež možné použiť väčšie množstvo nástrojov merajúcich rovnaký koncept. V našom prípade sme použili jeden nástroj pre každý koncept. Pre objektivizáciu výsledkov by bolo v budúcnosti vhodné použiť viac nástrojov skonštruovaných rôznymi autormi.

Prezentovaný výskum ďalej plánujeme rozšíriť o neurovedné dáta zhromaždené v rámci výskumu *Paměť a vědomí*. Naskytá sa nám príležitosť využiť výsledky z merania excitability primárneho vizuálneho kortexu V1 a insuly rovnakej vzorky participantov magnetickou rezonančnou spektroskopiou (MRS). Tieto výsledky môžeme ďalej porovnávať so záznamom z elektroencefalografie (EEG) počas prezentovania *Pattern Glare tasku*. Vďaka záznamu *resting state* a súčasne dátam z Pattern Glare tasku máme možnosť skúsiť objasniť protichodné výsledky štúdií dvoch skupín autorov − Dancea, Warda a Simnera (2021b) a Keogha, Bergmanna a Pearsona (2020). Po preštudovaní výsledkov týchto autorov totiž vzniká nejasnosť, či živšia VP súvisí skôr s nižšou alebo vyššou excitabilitou vizuálneho kortexu mozgu.

Zároveň by sme sa chceli pokúsiť vysvetliť vzťah *Pattern Glare tasku* s excitabilitou meranou pomocou MRS, kde sme zistili opačný smer asociácie naprieč našimi dvoma vzorkami. Vzhľadom k možnosti významných rozdielov mezi MRS meraniami jednej osoby v rôznych situáciách by sme sa sústredili na prítomnosť interakcií s intervenujúcimi premennými. Okrem psychologických dát je možné využiť aj behaviorálne a demografické údaje a metadáta merania ako je jeho denná doba alebo trvanie. Pokiaľ by sme našli významný faktor alebo kombináciu faktorov, mohli by

byť ovplyvnené všetky experimentálne dáta získané pomocou MRS.

10 Poďakovanie

Výskum bol realizovaný s podporou zdieľaného laboratória MAFIL na CEITEC MU pod záštitou MŠMT ČR (LM2018129 Czech-BioImaging), ktoré je súčasťou Euro-BioImaging (www.eurobioimaging.eu) ALM a Medical Imaging Node (Brno, CZ). Výskum bol spolufinancovaný z podpory Špecifického univerzitného výskumu, ktorý poskytuje MŠMT ČR. Dáta boli získané v rámci projektu COST Action CA18106 The neural architecture of consciousness.

Literatúra

Anālayo, B. (2019). In the seen just the seen: mindfulness and the construction of experience. Mindfulness, 10(1), 179-184. https://doi.org/10.1007/s12671-018- 1042-9

Bijlenga, D., Tjon-Ka-Jie, J. Y. M., Schuijers, F., & Kooij, J. J. S. (2017). Atypical sensory profiles as core features of adult ADHD, irrespective of autistic symptoms. European Psychiatry, 43, 51-57. https://doi.org/10.1016/j.eurpsy.2017.02.481

Braithwaite, J. J., Mevorach, C., & Takahashi, C. (2015). Stimulating the aberrant brain: Evidence for increased cortical hyperexcitability from a transcranial direct current stimulation (tDCS) study of individuals predisposed to anomalous perceptions. Cortex, 69, 1–13. https://doi.org/10.1016/j.cortex.2015.03.023

Cattaneo, Z., Pisoni, A., Papagno, C., & Silvanto, J. (2011). Modulation of visual cortical excitability by working memory: effect of luminance contrast of mental imagery. Frontiers in Psychology, 2, 29. https://doi.org/10.3389/fpsyg.2011.00029

Chiesa, A., Serretti, A., & Jakobsen, J. C. (2013). Mindfulness: Top–down or bottom–up emotion regulation strategy? Clinical psychology review, 33(1), 82-96. https://doi.org/10.1016/j.cpr.2012.10.006

Comte, M., Schön, D., Coull, J. T., Reynaud, E., Khalfa, S., Belzeaux, R., ... & Fakra, E. (2016). Dissociating bottom-up and top-down mechanisms in the cortico-limbic system during emotion processing. Cerebral cortex, 26(1), 144-155. https://doi.org/10.1093/cercor/bhu185

Curtis, R. (2016). The use of imagery in psychoanalysis and psychotherapy. Psychoanalytic Inquiry, 36(8), 593-602. https://doi.org/10.1080/07351690.2016.1226033

Dance, C. J., Jaquiery, M., Eagleman, D. M., Porteous, D., Zeman, A., & Simner, J. (2021a). What is the relationship between Aphantasia, Synaesthesia and Autism? Consciousness and Cognition, 89, 103087. https://doi.org/10.1016/j.concog.2021.103087

Dance, C. J., Ward, J., & Simner, J. (2021b). What is the Link Between Mental Imagery and Sensory Sensitivity? Insights from Aphantasia. Perception, 50(9), 757–782. https://doi.org/10.1177/03010066211042186

Dijkstra, N., Bosch, S. E., & van Gerven, M. A. (2017). Vividness of visual imagery depends on the neural overlap with perception in visual areas. Journal of Neuroscience, 37(5), 1367-1373. https://doi.org/10.1523/JNEUROSCI.3022- 16.2016

Di Nuovo, A., & Cangelosi, A. (2015, December). Artificial mental imagery in cognitive robots interaction [Paper presentation]. In 2015 IEEE Symposium Series on Computational Intelligence, South Africa, Cape Town. https://doi.org/10.1109/SSCI.2015.23

Finley, A. J., & Schmeichel, B. J. (2019). Aftereffects of self-control on positive emotional reactivity. Personality and Social Psychology Bulletin, 45(7), 1011- 1027. https://doi.org/10.1177/0146167218802836

Gallant, S. N. (2016). Mindfulness meditation practice and executive functioning: Breaking down the benefit. Consciousness and cognition, 40, 116-130. https://doi.org/10.1016/j.concog.2016.01.005

Grill-Spector, K., & Malach, R. (2004). The human visual cortex. Annual Review of Neuroscience, 27, 649-677. https://doi.org/10.1146/annurev.neuro.27.070203.144220

Hebert, K. R. (2018). Sensory processing styles and eating behaviors in healthy adults. British journal of occupational therapy, 81(3), 162-170. https://doi.org/10.1177/0308022617743708

Huttner, J. P., & Robbert, K. (2018). The role of mental factors for the design of a virtual memory palace [Paper presentation]. In Twenty-fourth Americas Conference on Information Systems, New Orleans, 2018.

Keogh, R., Bergmann, J., & Pearson, J. (2020). Cortical excitability controls the strength of mental imagery. elife, 9, e50232. https://doi.org/10.7554/eLife.50232

Lee, S. H., Kravitz, D. J., & Baker, C. I. (2012). Disentangling visual imagery and perception of real-world objects. Neuroimage, 59(4), 4064-4073. https://doi.org/10.1016/j.neuroimage.2011.10.055

Măirean, C. (2015). False memory for positive and negative life events. The role of mental imagery. Romanian Journal of Psychology, 17(1).

Marks, D. F. (1995). New directions for mental imagery research. Journal of Mental Imagery, 19(3-4), 153–167.

McCrae, R. R., & Costa, P. T. (1989). The NEO-PI/NEO-FFI Manual supplement. Odessa.

McDougall, S., & Pfeifer, G. (2012). Personality differences in mental imagery and the effects on verbal memory. *British Journal of Psychology*, *103*(4), 556-573. https://doi.org/10.1111/j.2044-8295.2011.02094.x

Milton, F., Fulford, J., Dance, C., Gaddum, J., Heuerman-Williamson, B., Jones, K., ... & Zeman, A. (2021). Behavioral and Neural Signatures of Visual Imagery Vividness Extremes: Aphantasia versus Hyperphantasia. Cerebral Cortex Communications, 2(2), 1-5. https://doi.org/10.1093/texcom/tgab035

Morawetz, C., Alexandrowicz, R. W., & Heekeren, H. R. (2017). Successful emotion regulation is predicted by amygdala activity and aspects of personality: A latent variable approach. Emotion, 17(3), 421– 441. https://doi.org/10.1037/emo0000215

Pearson, J. (2019). The human imagination: the cognitive neuroscience of visual mental imagery. Nature Reviews Neuroscience, 20(10), 624-634. https://doi.org/10.1038/s41583-019-0202-9

Pearson, J., & Westbrook, F. (2015). Phantom perception: voluntary and involuntary nonretinal vision. Trends in Cognitive Sciences, 19(5), 278-284. https://doi.org/10.1016/j.tics.2015.03.004

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.Rproject.org/

Reinhart, R. M., Xiao, W., McClenahan, L. J., & Woodman, G. F. (2016). Electrical stimulation of visual cortex can immediately improve spatial vision. Current Biology, 26(14), 1867-1872. https://doi.org/10.1016/j.cub.2016.05.019

Robertson, A. E., & Simmons, D. R. (2013). The relationship between sensory sensitivity and autistic traits in the general population. Journal of Autism and 83 Developmental disorders, 43(4), 775-784. https://doi.org/ 10.1007/s10803- 012-1608-7

Robertson, E. E. (2012). *Sensory experiences of individuals with autism spectrum disorder and autistic traits: A mixed methods approach* [Doctoral dissertation, University of Glasgow]. University of Glasgow.

Ružičková, A. (2022). *Senzorická senzitivita, osobnostné dimenzie a mindfulness ako prediktory živosti vizuálnej predstavivosti* [Bachelor´s thesis, Masaryk University]. Institutional Repository at the Masaryk University. https://is.muni.cz/th/p1vs7/

Takahashi, T., Kawashima, I., Nitta, Y., & Kumano, H. (2020). Dispositional Mindfulness Mediates the Relationship Between Sensory-Processing Sensitivity and Trait Anxiety, Well-Being, and Psychosomatic Symptoms. Psychological Reports, 123(4), 1083–1098. https://doi.org/10.1177/0033294119841848

Walach, T., Buchheld, N., Buttenmuller, V., Kleinknecht, N., & Schmidt, S. (2006). Measuring mindfulness: The Freiburg Mindfulness Inventory (FMI). Personality and Individual Differences, 40(8), 1543–1555. https://doi.org/ 10.1016/j.paid.2005.11.025

Yang, J., Mao, Y., Niu, Y., Wei, D., Wang, X., & Qiu, J. (2020). Individual differences in neuroticism personality trait in emotion regulation. Journal of Affective Disorders, 265, 468-474. https://doi.org/10.1016/j.jad.2020.01.086

Zeman, A., Milton, F., Della Sala, S., Dewar, M., Frayling, T., Gaddum, J., ... & Winlove, C. (2020). Phantasia – the psychological significance of lifelong imagery vividness extremes. Cortex, 130, 426–440. https://doi.org/10.1016/j.cortex. 2020.04.003

Úloha stelesnenia a pohľadu v interakcii človeka a robota

Sabína Samporová, Kassandra Friebe, Kristína Malinovská
Katedra aplikovanej informatiky, FMFI,
Univerzita Komenského v Bratislave
Mlynská dolina, 84248 Bratislava
Email: {samporova1,friebe1,rebrova1}@uniba.sk

Matěj Hoffmann
Katedra kybernetiky, Fakulta elektrotechnická,
České vysoké učení technické v Praze
Karlovo náměstí 13, 121 35 Praha 2
Email: matej.hoffmann@fel.cvut.cz

Abstrakt

Výskumná oblasť interakcie medzi človekom a robotom sa primárne zaoberá tým ako ľudia vnímajú roboty. Moderný výskum v tejto oblasti sa začína orientovať aj na klasické fenomény psychológie, ako je napríklad sociálny pohľad. V príspevku popisujeme experimentálne prostredie a procedúru na overenie vplyvu fyzickej prítomnosti robota a sociálnych stimulov, ako je naznačovanie pohľadom, na to, ako človek robota vníma a stotožňuje sa s ním. Načrtneme tiež perspektívu interakcie človeka a robota vo virtuálnej realite.

1 Úvod

Vo svetle technologického pokroku dnešnej doby sa roboty stávajú bežnou súčasťou nášho života a spoločnosti. Stále viac pozornosti sa dostáva odvetviu skúmania interakcie medzi človekom a robotom (Human-robot interaction, HRI, Bartneck a spol., 2020), ktoré spája psychológiu, kognitívnu vedu a robotiku. HRI často pracuje s humanoidnými robotmi, či agentami a skúma mnoho faktorov, ktoré formujú sociálnu interakciu, no prepojenie medzi HRI a fenoménmi klasickej sociálnej kognície nájdeme v literatúre len zriedkavo. V našej práci spájame a skúmame klasický fenomén sociálneho pohľadu a vnímanie humanoidného robota v HRI.

Veľmi aktuálna téma, ktorá sa nás všetkých dotkla za uplynulej pandémie Covid 19 je otázka, či a ako môže fyzická prítomnosť ovplyvniť naše interakcie. Intuitívne môžeme potvrdiť, že fyzická prítomnosť v tom istom priestore je základným faktorom, ktorý ovplyvňuje to, ako interagujeme a vnímame nášho partnera v interakcii. Keď túto otázku rozšírime na oblasť HRI, dostávame sa k skúmaniu vplyvu stelesnenia robota na jeho vnímanie človekom. Pýtame sa, či naša intuícia platí aj pri interakcii ľudí s robotmi v rôznych spodobneniach.

Motiváciou pre takéto skúmanie je aj možnosť nového výskumného prospektu, skúmania interakcie človeka a robota vo virtuálnej realite (VR). Ak by sme dokázali potvrdiť, že kvalita vnímania a stotožnenia sa človeka s robotom, ako s partnerom, je rovnaká vo VR ako pri fyzických robotoch, otvoria sa nové smery aj pre výskum a budovanie umelej inteligencie. Trénovanie robotov totiž najbežnejšie prebieha práve v simulovanom prostredí, kde človek ako partner pre kolaboratívnu úlohu zasiahnuť nemôže. Prepojením VR a kognitívnej robotiky môžeme získať väčšiu bezpečnosť pre človeka, ako aj pre robota.

Pre naše výskumné účely sme navrhli experiment a experimentálne prostredie pre tri rôzne formy prezentácie robota a to fyzická, teleprezenčná a virtuálna. V príspevku popíšeme experimenty a zistenia z oblasti HRI, ktoré inšpirovali náš výskum. Ďalej predstavíme náš experiment a experimentálne prostredie, ktoré sme navrhli a čiastočne vytvorili. Nakoniec stručne popíšeme predbežné výsledky našich experimentov.

2 Stelesnenie, prítomnosť a sociálny pohľad v interakcii človeka a robota

Stelesnenie (embodiment) robota poukazuje nie len na hardvér robota, teda z akých mechanických častí, senzorov a aktuátorov sa robot skladá. Do otázky stelesnenia vstupujú aj softvérové súčasti robota, keďže práve tie sú kľúčové v tom, ako robot vystupuje a aký dojem urobí na človeka, s ktorým interaguje. Obe tieto časti ponúkajú možnosti a zároveň kladú obmedzenia na to, ako dokáže robot interagovať s ľuďmi vo svete, v ktorom sa nachádza. Ak sa robot fyzicky dostatočne podobá na človeka, je pre ľudí jednoduchšie s ním komunikovať tak, ako s človekom (Bartneck a spol., 2020). Tým pádom sa ľuďom prirodzene prenesú skúsenosti a očakávania z medziľudskej interakcie do interakcie s robotom.

Zatiaľ, čo stelesnenie poukazuje na formu, telo a vlastnosti agenta, prítomnosť (presence) hovorí o prítomnosti robota v rovnakom fyzickom priestore ako sa nachádza človek. Ako ukázali Li (2015), roboty, ktoré sa nachádzajú v rovnakom fyzickom priestore ako ľudia, sú hodnotení pozitívnejšie a interakcia s nimi je úspešnejšia v porovnaní s robotmi prezentovanými na obrazovke. Jedným z vysvetlení tohto javu môže byť, že fyzicky prítomné roboty vnímame

s ohľadom na ich spoločenské postavenie, a preto sa k nim správame tak, ako by sme sa správali k iným sociálnym aktérom. Napríklad Leyzberg a spol. (2012) ukázali, že sa participanti učili efektívnejšie, ak ich učil fyzicky prítomný robot, v porovnaní s robotom na videu, ale aj s nezúčastneným ľudským hlasom.

V sérii interakčných štúdií inštruovali Liu a spol. (2017) participantov spolupracovať s robotom tak, že posúvali tanier, na ktorý mal robot zhodiť nejaký objekt. Robot bol prezentovaný buď na obrazovke alebo vo virtuálnej realite. Ľudia, ktorí interagovali s robotom vo VR, boli výrazne rozhodnejší a presnejší pri určení pozície, na ktorú majú tanier posunúť. Zatiaľ, čo tieto výsledky naznačujú, že VR by mohla byť novým nástrojom pre HRI, je dôležité preskúmať, ako na to vplývajú rôzne faktory interakcie a porovnať túto možnosť prezentácie nielen s teleprítomnosťou, ale aj s fyzickou prítomnosťou. Li a spol. (2019) pri porovnávaní fyzickej a virtuálnej podoby robota ukázali, že jeho prezentácia formovala proxemiku interakcie a našli významné rozdiely medzi týmito dvoma formami prezentácie. Napriek tomu, že vyššie uvedené výsledky naznačujú, že prezentovanie robota vo VR môže ovplyvniť špecifické výstupné premenné v HRI experimentoch, spektrum premenných na štúdium je široké a je potrebné aj replikovať a potvrdiť existujúce výsledky (Wijnen a spol., 2020).

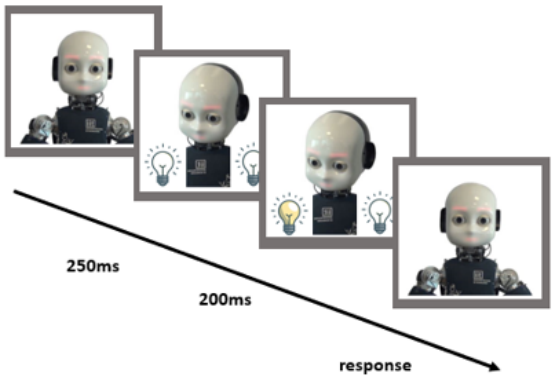
Sociálne interakcie sú pre ľudí spontánne a prirodzené. Ich podstatou je komplexná výmena rôznych sociálnych signálov, ako je ukazovanie, držanie tela a plynulosť pohybu. Narozdiel od iných sociálnych signálov má pohľad očí špecifický význam pri prijímaní a odovzdávaní informácií medzi ľuďmi (Risko a spol., 2016). Hovoríme o takzvanom sociálnom pohľade. Náš pohľad môže zaujať pozornosť druhého (Kuhn a spol., 2009) a vnímanie pohľadu inej osoby nám môže poskytnúť informáciu o tom, kam je zameraná jej pozornosť (Frischen a spol., 2007). Proces v ktorom využijeme pohľad očí na zameranie našej pozornosti sa volá gaze cueing a hovorí sa o ňom ako o prekurzore zdieľanej pozornosti (Emery, 2000).

Veľký význam sociálneho pohľadu sa s cieľom vyvolať väčší záujem u ľudí premietol aj do dizajnu sociálnych robotov, akým je napríklad Kismet (Breazeal a Scassellati, 1999). Zatiaľ jedinú nám známu systematickú štúdiu, ktorá skúma vzťah stelesnenia, prítomnosti a komunikácie mimikov tváre, vykonali Mollahosseini a spol. (2018). Participanti tu interagovali s jedným zo štyroch typov agentov, z ktorých každý sa líšil svojím stelesnením a prítomnosťou (virtuálny agent, fyzický robot, teleprezenčný robot a človek). Výsledky Mollahosseini a spol. (2018) naznačujú, že zatiaľ čo spôsob stelesnenia nemá vplyv na vizuálnu reč, rozpoznávanie pohľadu očí rozhodne týmto ovplyvnené je.

3 Špecifikácia experimentu

V našom experimente je, rovnako ako u Wiese a spol. (2018), pred participantom postavený robot a zariadenie na vytváranie svetelných stimulov. V našom prípade striedame formu prezentácie robota a to fyzickú, teleprezenčnú a virtuálnu. Na pravej a ľavej strane od robota sa nachádzajú dve lampy, ktoré slúžia na vytváranie svetelných stimulov. Ako reakciu na stimul má participant čo najrýchlejšie stlačiť jedno z dvoch tlačidiel, a to to, ktoré je na rovnakej strane ako lampa, ktorá zasvietila.

Jedno meranie reakcie sa skladá zo štyroch fáz. V prvej fáze sa robot pozerá priamo vpred, na miesto kde sa nachádza participant. Tým sa snaží nadviazať očný kontakt a zaujať participanta. Po 250 milisekundách prejde do druhej fázy, kedy robot vykoná pohyb hlavy a očí buď doprava alebo doľava. Tento pohyb znamená upriamenie pohľadu na svetlo na strane, na ktorú sa hýbal. Podľa toho, na ktorej strane neskôr zasvieti svetlo rozlišujeme valídne (súhlasné) a nevalídne (nesúhlasné) nápovedy. Schematické zobrazenie fáz merania zobrazujeme na Obr. 1.



Obr. 1: Fázy merania reakcie a valídna nápoveda.

Po 200 milisekundách, v tretej fáze, sa zasvieti jedna z lámp, čím sa vytvorí svetelný stimul. Vtedy má participant čo najrýchlejšie zareagovať stlačením tlačidla pre príslušnú (ľavú alebo pravú) stranu, podľa toho, kde svieti svetlo. V poslednej fáze sa robot vráti do iniciálnej polohy, kde opäť nadväzuje očný kontakt a pozerá priamo na participanta. Súčasťou každého merania je odmeranie a zapísanie reakčného času, t.j. času od zasvietenia lampy po stlačenie klávesy.

Každé meranie má priradenú jednu zo špecifických konfigurácií, ktoré určujú smer pohybu robota iCub a lampu, ktorá zasvieti. Jedna konfigurácia sa skladá z označenia strany, na ktorú ukáže robot, a označenia strany, na ktorej zasvieti lampa. V experimente sa vyskytujú štyri rôzne druhy konfigurácií:

- pohľad vľavo, stimul vľavo
- pohľad vpravo, stimul vpravo

- pohľad vľavo, stimul vpravo
- pohľad vpravo, stimul vľavo

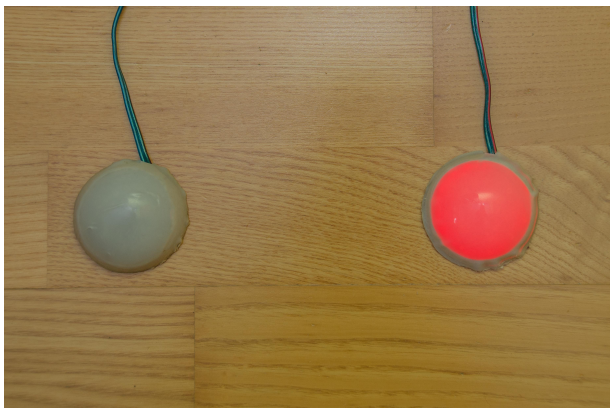
Experiment s jedným participantom je zložený z 80 meraní reakcie. V týchto 80 meraniach je zastúpených po 20 z každého vyššie popísaného druhu konfigurácií. Konfigurácie sú na začiatku preusporiadané do náhodného poradia. Keď je účastík pripravený, spustí sa program ovládajúci robota, zariadenie na vytváranie svetelných stimulov a meranie reakcií. Každé meranie sa vykoná podľa konfigurácie, ktorá mu prislúcha. Zakaždým sa do výstupného súboru uloží poradové číslo konfigurácie, samotná konfigurácia, rýchlosť a správnosť reakcie.

4 Návrh a realizácia experimentálneho prostredia

4.1 Zariadenie na vytváranie svetelných stimulov

Kľúčovou časťou experimentu je meranie rýchlosti reakcie na stimuly. Ako vhodné riešenie na vytváranie stimulov sme použili svetelné led pásy, ktoré boli ovládané pomocou dosky Arduino Nano. Finálna podoba našich experimentálnych lámp je zobrazená na Obr. 2

Mechanizmus ovládania svetiel bol jednoduchý. Arduino komunikovalo s počítačom cez sériový port, pričom dostávalo inštrukcie s nami špecifikovaným významom. Významy inštrukcií boli: zapnutie ľavého svetla, zapnutie pravého svetla, vypnutie ľavého svetla, vypnutie pravého svetla. Bezprostredne po prijatí inštrukcie Arduino doska vykonala danú akciu.



Obr. 2: Zariadenie na vytváranie svetelných stimulov

4.2 Robot iCub

Robot iCub (Metta a spol., 2010) je humanoidný robot určený predovšetkým na výskumné účely, akými sú najmä návrh a testovanie algoritmov umelej inteligencie, ktoré vyžadujú fyzicky stelesneného robota, ktorý sa svojou stavbou tela a rozsahom pohybov podobá človeku. Je vytvorený ako otvorený projekt a celý softvér pre prácu s robotom je open-source. Vyvinutý bol na Italian Institute of Technology, Genoa. iCub má približne 1 meter, váži 22 kilogramov a svojimi pohybovými schopnosťami sa najviac podobá na tri a pol ročné dieťa. iCub má viac než 50 ovládateľných kĺbov, čo umožňuje vykonávanie veľmi rôznorodých pohybov.

Pre náš experiment bolo dôležité, aby robot pôsobil čo najprirodzenejšie a čo možno najviac sa jeho pohyby podobali ľudským. Robota iCub sme vybrali nie len preto, lebo sa podobá na človeka, ale najmä preto, lebo má schopnosť hýbať hlavou, a ako jeden z mála podobne veľkých humanoidov dokáže hýbať očami a žmurkať (Roncone a spol., 2016; Lehmann a spol., 2016).

4.3 Ovládanie robota

Experiment sme vykonávali s robotom iCub v dvoch formách stelesnenia. Prvou bola fyzická forma, kde išlo o skutočného robota. Druhou bola digitálna forma, kde išlo o robota iCub v simulátore (Tikhanoff a spol., 2008). Dve formy robota v experimentálnom prostredí zobrazujeme na Obr. 3. Implementačný jazyk pre obe formy bol C++. Obe formy používali na ovládanie robota iCub platformu YARP. Softvér využíval nami navrhnuté a skonštruované svetlá.

Na zobrazenie a možnosť pohybu robota sme využívali softvér Simulátor robota iCub - iCubSIM. Vytvorili sme konzolovú aplikáciu, ktorá komunikovala s platformou YARP a tým ovládala pohyb robota. Komunikovala aj s Arduinom, čím zabezpečovala vytváranie stimulov. Reakcia na stimul sa zaznamenávala stláčaním klávesy na klávesnici. Robot sa v tejto forme zobrazoval na obrazovke počítača, pomocou ktorej participant sleduje zmenu jeho pohybu.

Ovládanie fyzického robota je riešené rovnako, pomocou YARP. Preto sme program konzolovej aplikácie s malými úpravami niektorých konštánt (rýchlosti, zrýchlenia, oneskorenie) upravili tak, aby bol pohyb fyzického robota rovnaký ako robota v simulátore. Zdrojový kód pre meranie experimentu s oboma formami robota a zariadeniami na vytváranie stimulov je dostupný v GitHub repozitári [1].

4.4 Prepojenie robota s virtuálnou realitou

Súčasťou nášho výskumu bola aj snaha o skúmanie virtuálneho stelesnenia humanoidného robota a miera vplyvu tejto formy na vnímanie sociálnych podnetov. V súčastnosti je prepojenie robota s virtuálnym headsetom pomerne novou a nie veľmi dobre preskúmanou úlohou. Naša snaha bola prepojiť robota iCub s virtuálnym headsetom HTC Vive Pro. Vyhľadali sme niekoľko existujúcich projektov, ktoré spojením rôznych technológií vytvorili prepojenie robota a virtuálnej reality.

[1] `github.com/Sabka/icub-hri-cuing`

Obr. 3: Setup digitálneho a fyzického prostredia

Unity3D Robotics Toolkit
Prvým riešením bol projekt Unity3D Robotics Toolkit. Je to open-source nástroj prepájajúci Robot Operating System (ROS) s prostredím Unity3D. Prepojenie je riešené pomocou existujúceho ROSbridge, ktorý je obsiahnutý v softvéri ROS. Tento projekt implementuje ROSbridge interface v Unity3D na strane človeka a prepája ho s ROSbridge node na strane ROS a teda robota (Krupke a spol., 2017).

VRUI MDF a OpenVR headset ROS
Ďalší softvér, ktorý sme skúmali je VRUI MDF. Tento prepája HTC Vive so simulátorom Gazebo, ktorý je veľmi vhodný pre prácu s iCubom. Na prepojenie používa VRui server, ROS, SteamVR for Linux a OpenCV. Projekt bol určený pre výuky na univerzite.

Autori Shi a McGhan (2020) uvádzajú, že tento setup funguje dobre pre niektoré edukačné projekty a pre menej komplexné robotické modely. Tento softvér má ale problém so simuláciou komplexnejších robotov, čo sa prejavuje spomalením simulácie. Preto úplne nevyhovuje na testovanie humanoidných robotov. Uvádzajú aj analýzu spomalenia.

Od projektu VRUI MDF bol odvodený softvér OpenVR headset ROS. Architektúra ostáva podobná, s tým rozdielom, že komponent VRui server bol vymenený za softvér OpenVR. OpenVR zaznamenáva informácie o zmene polohy virtuálneho zariadenia a posiela obraz do virtuálneho headsetu. Ide o stále živý projekt, ktorý sa vyvíja. OpenVR headset ROS bol otestovaný aj s virtuálnym headsetom HTC Vive Pro, jeho ovládačmi a base stanicami Lighthouse 1.0 aj 2.0., ktorý využívame.

V rámci nášho projektu sme softvér OpenVR headset ROS bližšie preskúmali, avšak keďže ide o experimentálny softvér, narazili sme na prekážky, ktoré sme zatiaľ nedokázali vyriešiť. Nedostatky sme nahlásili autorom projektu, ktorí ho na základe týchto podnetov opravujú a zlepšujú.

5 Predbežné výsledky experimentu

Naše predbežné výsledky naznačujú, že komunikáciu pohľadom, ako základný jav ľudskej sociálnej kognície, možno nájsť aj v interakciách s humanoidnými robotmi. Napriek predpokladom z existujúcich štúdií sa zdá, že rôzne spôsoby prezentácie robota nemenia silu efektu nápovedy pohľadom. Jedným z možných vysvetlení je prominencia humanoidnej formy robota iCub a dominancia vplyvu sociálneho pohľadu na percepciu stimulov. V súlade s tým sú aj dojmy našich participantov, ktorí vnímali fyzického robota podobne ako robota na obrazovke.

6 Záver

V našej práci prezentujeme experiment a experimentálne prostredie pre skúmanie vplyvu formy stelesnenia a prítomnosti robota na mieru efektu sociálneho pohľadu u človeka. Pre náš experiment sme navrhli naše vlastné hardvérové aj softvérové riešenie pre fyzicky a teleprezenčne stelesneného robota. Okrem toho sme preskúmali možnosti rozšírenia experimentu na virtuálnu formu prezentácie robota.

Z predbežných výsledkov meraní usudzujeme, že vplyv humanoidnej formy robota a sila efektu sociálneho pohľadu sú silnejšie než forma prezentácie robota. Zistenie, že aj teleprítomný iCub môže byť rovnako dobrým partnerom pre interakciu ako fyzický, je veľmi pozitívne hlavne v oblasti kognitívnej robotiky, kde vývin kognitívnych systémov pre robota prebieha hlavne v simulovanom prostredí a používanie fyzických robotov je obmedzené. Dôležitou úlohou do budúcna, je implementácia a skúmanie možností virtuálnej reality.

Výskum vzťahu medzi stelesnením a prítomnosťou robota môže mať veľké opodstatnenie v praktickej interakcii medzi človekom s robotom. Naše výsledky môžu napomôcť tomu, na aké účely sa bude využívať daná forma stelesnenia robota, ale aj priniesť nové otázky v oblasti kognitívnej vedy a HRI.

Poďakovanie

Tento príspevok vznikol za podpory Grantovej agentúry České republiky, projekt č. 20-24186X. Ďalej sa chceme srdečne poďakovať Jakubovi Rozlivkovi a Lukášovi Rustlerovi za pomoc s programovaním robota iCub a Adamovi Rojíkovi za organizáciu náboru participantov. Za podporu tiež ďakujeme Slovenskej spoločnosti pre kognitívnu vedu SSKV[2].

[2] https://cogsci.fmph.uniba.sk/sskv/

Literatúra

Admoni, H. a Scassellati, B. (2017). Social eye gaze in human-robot interaction: A review. *Journal of Human-Robot Interaction*, 6(1):25–63.

Bartneck, C., Belpaeme, T., Eyssel, F., Kanda, T., Keijsers, M. a Sabanovic, S. (2020). *Human-Robot Interaction: An Introduction*. Cambridge University Press.

Breazeal, C. a Scassellati, B. (1999). How to build robots that make friends and influence people. V *Proceedings 1999 IEEE/RSJ international conference on intelligent robots and systems. Human and environment friendly robots with high intelligence and emotional quotients*, vol. 2, str. 858–863. IEEE.

Emery, N. J. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & biobehavioral reviews*, 24(6):581–604.

Frischen, A., Bayliss, A. P. a Tipper, S. P. (2007). Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychological bulletin*, 133(4):694.

iCub Tech IIT (2022). icub technical docs. `https://icub-tech-iit.github.io/documentation/`.

Krupke, D., Starke, S., Einig, L., Zhang, J. a Steinicke, F. (2017). Prototyping of immersive hri scenarios. str. 537–544.

Kuhn, G., Tatler, B. W. a Cole, G. G. (2009). You look where i look! effect of gaze cues on overt and covert attention in misdirection. *Visual Cognition*, 17(6-7):925–944.

Lehmann, H., Roncone, A., Pattacini, U. a Metta, G. (2016). Physiologically inspired blinking behavior for a humanoid robot. V *International Conference on Social Robotics*, str. 83–93. Springer.

Leyzberg, D., Spaulding, S., Toneva, M. a Scassellati, B. (2012). The physical presence of a robot tutor increases cognitive learning gains. V *Proceedings of the annual meeting of the cognitive science society*, vol. 34.

Li, J. (2015). The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, 77.

Li, R., van Almkerk, M., van Waveren, S., Carter, E. a Leite, I. (2019). Comparing human-robot proxemics between virtual reality and the real world. V *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, str. 431–439. IEEE.

Liu, O., Rakita, D., Mutlu, B. a Gleicher, M. (2017). Understanding human-robot interaction in virtual reality. V *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, str. 751–757. IEEE.

Martini, M. C., Buzzell, G. A. a Wiese, E. (2015). Agent appearance modulates mind attribution and social attention in human-robot interaction. V *International conference on social robotics*, str. 431–439. Springer.

Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., Von Hofsten, C., Rosander, K., Lopes, M., Santos-Victor, J. a spol. (2010). The icub humanoid robot: An open-systems platform for research in cognitive development. *Neural networks*, 23(8-9):1125–1134.

Mollahosseini, A., Abdollahi, H., Sweeny, T. D., Cole, R. a Mahoor, M. H. (2018). Role of embodiment and presence in human perception of robots' facial cues. *International Journal of Human-Computer Studies*, 116:25–39.

Risko, E. F., Richardson, D. C. a Kingstone, A. (2016). Breaking the fourth wall of cognitive science: Real-world social attention and the dual function of gaze. *Current Directions in Psychological Science*, 25(1):70–74.

Roncone, A., Pattacini, U., Metta, G. a Natale, L. (2016). A Cartesian 6-DoF gaze controller for humanoid robots. V *Robotics: science and systems*, vol. 2016.

Shi, Z. a McGhan, C. L. R. (2020). Affordable virtual reality setup for educational aerospace robotics simulation and testing. *Journal of Aerospace Information Systems*, 17(1):66–69.

Tikhanoff, V., Cangelosi, A., Fitzpatrick, P., Metta, G., Natale, L. a Nori, F. (2008). An open-source simulator for cognitive robotics research: the prototype of the icub humanoid robot simulator. V *Proceedings of the 8th workshop on performance metrics for intelligent systems*, str. 57–61.

Wiese, E., Weis, P. P. a Lofaro, D. M. (2018). Embodied social robots trigger gaze following in real-time hri. *2018 15th International Conference on Ubiquitous Robots (UR)*, str. 477–482.

Wijnen, L., Lemaignan, S. a Bremner, P. (2020). Towards using virtual reality for replicating hri studies. V *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, str. 514–516.

Umělý život jako umělecký a estetický problém

Aleš Svoboda

Univerzita Karlova, Fakulta humanitních studií
Pátkova 2137/5, 182 00 Praha 8 – Libeň
ales.svoboda@fhs.cuni.cz

Abstrakt

Od zrodu počítačového, respektive digitálního umění se generativní umění snažilo uvolnit původní striktní determinismus počítačových programů do rámcově určených, avšak značně nepředvídatelných výsledků. Lze však v tomto případě hovořit o inteligenci, respektive o inteligentním uměleckém rozhodování? Přes veškerou náhodnost jsou výsledky předem omezeny, byť do nedozírně velkých souborů. Ani postupy, které se hlásí k umělé inteligenci – ať už tradičním symbolickým modelováním (AARON Harolda Cohena) nebo využitím umělých neuronových sítí (projekty skupiny Obvious) – nenaplňují imperativ otevřenosti umění jako systému. Zdá se ovšem, že projekty využívající principy umělého života mohou z vlastní podstaty pracovat se stále složitějšími pravidly a stále komplikovanější strukturou, a přitom si při vší otevřenosti ponechat vnitřní konzistenci.

1 Člověk – rozum, jazyk, umění a umělá inteligence

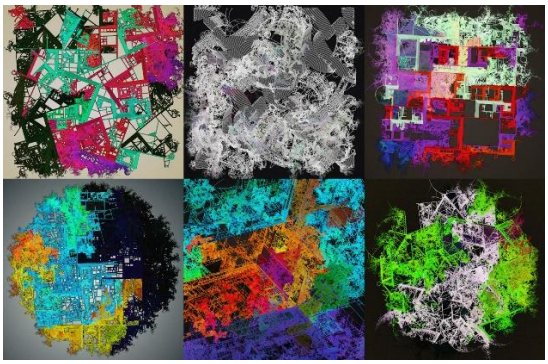
Lidská sebereflexe živočišného druhu se postupně upnula především na schopnost rozumného uvažování, racionality, diskurzivního a komplexního propojování stále obsáhlejšího objemu zapamatovaných a ověřených poznatků a jejich postupného bezrozporného strukturování. To samozřejmě doprovázelo poznání, že takový proces je umožněn existencí znakových systémů, především znakového systému k tomu účelu nejpropracovanějšího – jazyka. Noosféra je pak interagujícím prostorem vyplněným technologií a kulturou jako lidskými produkty, lidskou extenzí a záštitou. Umění, jako exkluzivní oblast kultury, se stalo znakovým systém, jenž především udržuje svoji otevřenost, takže v porovnání s jazykem jde o znakový systém mnohem dynamičtější. Umění za to ovšem občas platí nemalou daň ztrátou srozumitelnosti, která se projevuje především v obdobích překotných společenských proměn.

S příchodem technologie nového stupně – výpočetních strojů – se probudila i nová naděje na vytvoření autonomní, na člověku nezávislé inteligence. Mezníkem se stává ustavení oboru umělé inteligence na letním workshopu v Darthmouthu v roce 1956. V jisté paralele s Turingovým testem naznačil obsah umělé inteligence jako vědy Marvin Minsky: „…vytváření strojů nebo systémů, které budou při řešení určitého úkolu užívat takového postupu, který – kdyby ho dělal člověk – bychom považovali za projev jeho inteligence.“ Definice se samozřejmě za téměř sedmdesát let specifikovala do mnoha variant a obohatila o mnohé aspekty, ale zůstává to podstatné: softwarová simulace inteligentního lidského chování. Vladimír Mařík například hovoří o postupech a algoritmech, „…které ve svém důsledku vedou k určitému napodobení projevů inteligentního chování člověka (Mařík 1993: 15).“ Co je inteligentní chování stále zůstává definičně poněkud nejisté a umělá inteligence bere lidskou inteligenci nejen jako výzvu k napodobení, ale i jako objekt analýzy. Předpokladem je většinou vytváření vnitřních modelů (znakových struktur) okolí, které umožňují predikci a účelná rozhodnutí pro požadovanou činnost. Podobně o umělé inteligenci přemýšlí i Margaret Bodenová: „Umělá inteligence (AI) se snaží o to, aby počítače dělaly onen druh věcí, které dělají mysli. Některé z nich (například usuzování) obvykle charakterizujeme jako „inteligentní“. Jiné zase (například „vidění“) nikoliv. Ale všechny vyžadují duševní dovednosti – jako vnímání, asociaci, predikci, plánování, řízení motoriky – které umožňují lidem a živočichům dosahovat jejich cílů. Inteligence není jednorozměrná, ale je bohatě strukturovaným prostorem odlišných schopností zpracovávání informací. Proto používá AI mnoho různých postupů, určených pro mnoho různých úkolů (Boden 2016: 1).“

Umělecká tvorba se většinou považuje za činnost založenou na inteligenci, protože je jednak spjata (ať už intuitivně nebo vědomě a diskurzivně) se zapamatovanou a naučenou kulturní tradicí, jednak je také cílevědomá, ve svých podstatných projevech dosahuje jak estetických, tak uměleckých cílů, byť zpočátku nejasných a tušených.

2 Generativní počítačové umění

Využívání počítačů pro uměleckou tvorbu přineslo především zcela specifickou oblast umění – tzv. generativní počítačové umění. Z počáteční snahy napodobit stávající nezobrazivá umělecká díla se vyvinula velmi plodná a lákavá snaha nalézt různé algoritmy, které se zahrnutím jistého stupně náhodnosti použitých parametrů budou produkovat nečekané vizuální struktury. Lev Manovich, významný americký novomediální teoretik, v souvislosti s výstavou *Abstraction Now!*, která se uskutečnila ve vídeňském Künstlerhausu v roce 2003, konstatoval aktuální paradigmatický posun od „modernistické redukce“ založené v 10. letech 20. století k algoritmicky řízené komplikovanosti abstraktních obrazů. Paralelu viděl v obdobném posunu od klasického determinismu newtonovské fyziky ke studiu komplexity v současné vědě (Manovich 2004).
Philip Galanter pak definicí generativního umění ustavuje zastřešující uměleckou praxi, „...ve které umělec používá nějaký systém, jako třeba soubor pravidel přirozeného jazyka, počítačový program, stroj nebo jinou procedurální konstrukci, která je uvedena do pohybu s určitým stupněm nezávislosti, aby přispěla k tvorbě díla nebo do hotového uměleckého díla vyústila (Galanter 2016: 146–180).“ Díla generativního umění, jejich zásadní položku dnes činí počítačově vytvářená díla, jsou pak charakteristická skladbou velkého počtu značně proměnlivých elementů, přičemž jejich „chování“ lze sice rámcově předepsat programem, ale nad jistou mez již není zcela možné předpovědět všechny důsledky. Tvůrce manipuluje s celým procesem jen na základě značné intuice. Dostáváme se tak vlastně na práh „doslovné“ vizuální algoritmizace.



Obr. 1. Frank Force: Chaospills, 2022 – příklad generativního počítačového umění.

Generativní počítačové umění nicméně přináší otázku: nakolik je umělecký software tohoto druhu vůči svému tvůrci opravdu autonomní?

3 Případová studie: Cohen vs. Obvious

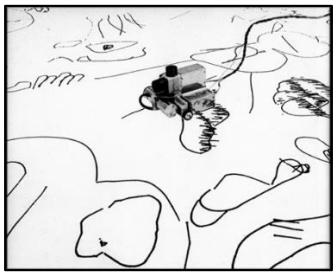
Svět umění i počítačová věda dnes sdílí dva výrazné příklady umělecké tvorby, které se odvolávají na umělou inteligenci. Jednak je jím dlouhodobý projekt Harolda Cohena AARON, jednak zcela nedávný úspěch pařížské umělecké skupiny Obvious. Jestliže Cohenův software průběžně stále komplikovaněji pracoval s rozšiřující se databází na principu tradiční umělé inteligence, tedy se symbolicko-reprezentačním, algoritmickým, z centra řízeným přístupem, potom skupina Obvious využila umělou neuronovou síť, která se opakovanou interakcí s předkládaným vizuálním materiálem dokázala vyškolit k uměleckému výsledku.

3.1 Harold Cohen, AARON

Harold Cohen (1928–2016) byl britský umělec usazený ve Spojených Státech, který se v roce 1966 účastnil spolu s dalšími britskými umělci Benátského bienále.[1] V roce 1968 odjel jako hostující profesor na roční stáž na Kalifornskou univerzitu v San Diegu (UCSD), již se záměrem seznámit se s možnostmi využití počítače pro uměleckou práci. Program, který pro počítačovou malbu vytvořil, poprvé veřejně předvedl v Los Angeles County Museu v r. 1972. Od roku 1973 pak pracoval v Laboratoři umělé inteligence na Stanford University (Cohen 2002).

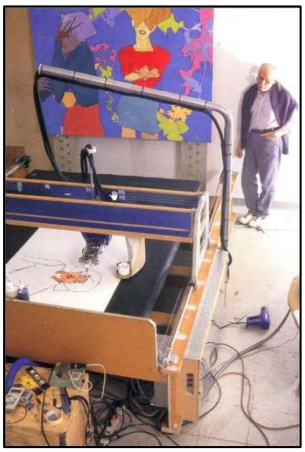
Obr. 2. Kreslící stroj Harolda Cohena v roce 1979. Archiv Harolda Cohena.

[1] Společně s Haroldem Cohenem vystavoval v pavilonu Velké Británie jeho bratr Bernard, Anthony Caro, Robyn Denny a Richard Smith.



Obr. 3. Harold Cohen – abstraktní motivy ze 70. let a kreslicí zařízení.

Program AARON jde cestou budování symbolicko-reprezentačního repertoáru obsaženého v paměti a algoritmů, které vybírají adekvátní rozhodnutí v daných uzlových bodech. Cohen tak víc než 40 let program stále přepisoval, postupně ho vybavoval stále novými komplikovanějšími postupy, které vycházely z dosažených výsledků. Od původních abstraktních motivů dospěl k zobrazivosti, k vegetaci a postavám, které podle jeho slov realizovaly viděné bez vidění.



Obr. 4. Harold Cohen při tisku barevného figurálního uměleckého díla v roce 1994.

Obr. 5. Harold Cohen: 020514, 2003 – figurální motivy.

3.2 Obvious, The next Rembrandt

Pařížská umělecká skupina Obvious se v roce 2016 rozhodla vytvořit obraz, který by respektoval osobní malířský styl nizozemského umělce 17. století Rembrandta van Rijn. Ve spolupráci s Delft University of Technology, Mauritshuis v Haagu a Rembrandt House Museem v Amsterdamu shromáždila skeny 346 Rembrandtových děl a umělou neuronovou síť vlastně naučila analyzovat, srovnávat a hodnotit prvky portrétů, a tím v případné konstrukci nového díla se výběru přiklánět k typičtějším tvarovým a barevným variantám malířova stylu.

Obr. 6. Obvious – analýza a komparace prvků portrétu.

Takto připravená umělá neuronová síť nakonec vypracovala nový pravděpodobný autoportrét umělce. Pro větší přesvědčivost byl obraz vytištěný 3D technologií, aby textura povrchu vyvolávala dojem autentické malby.

Obr. 7. Obvious: Next Rembrandt, 2016.

Z umělecko-historického hlediska takový postup samozřejmě zcela přehlíží přibližně padesátiletý vývoj Rembrandtovy umělecké osobnosti a nečasově průměruje jeho malířské sklony. Chová se jako zkušený falzifikátor, který ve své práci může právě těžit z již vypracovaného, dovršeného a existujícího tvůrčího přístupu. Nejedná se zde o konstruování svobodné tvůrčí umělé inteligence, ale o omezenou, napodobující rutinu.

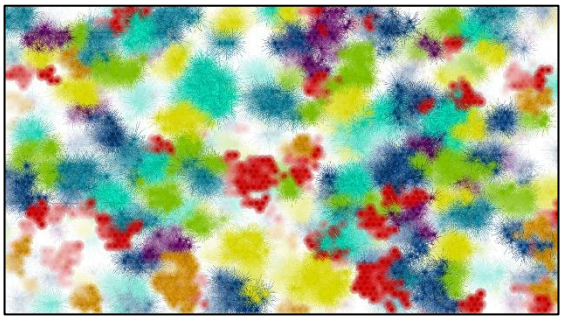
3.3 Obvious, Portrét Edmonda de Belamy

Dva roky na to vytvořila skupina Obvious jiný obraz, který tentokrát nepracuje s autorským stylem, ale se stylem určité historické epochy – obraz má evokovat malbu z 18. století. Tentokrát umělá neuronová síť GAN (*Generative Adversarial Network*) devatenáctiletého Američana Robbieho Barrata s algoritmem pro strojové učení zpracovala datový soubor 15 000 portrétů ze 14. až 20. století. Učení posilovalo právě výběr charakteristik preferovaného období.
Výsledek působí skicovitým akvarelovým dojmem, vyvolává dojem nejistého rozostření. Nelze si nevzpomenout na impresionistický princip zvýšeného zapojení představivosti diváka. Jako poněkud povrchní ornament působí umístění použitého algoritmu v tradičním místě pro podpis umělce.[2]

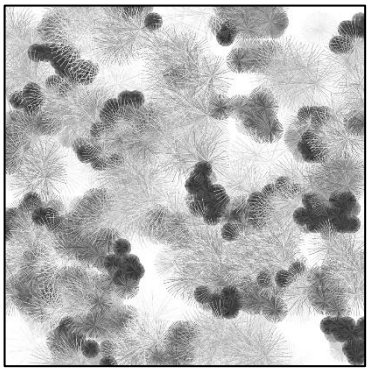
Obr. 8. Pierre Fautrel, spoluzakladatel skupiny Obvious s obrazem Portrét Edmonda de Belamyho, 2018, digitální tisk ne plátně, 70 x 70 cm.

4 Umělý život – otevřená vizuální struktura

Dokonce ještě dříve, než byl umělý život vyhlášen za vědeckou disciplínu (1986), se řada umělců věnovala tvorbě abstraktních struktur, které počítaly s vnitřní vývojovou interakcí. Mezi ně patřil Paul Brown (1973) a Yoichiro Kawaguchi (1982). Záhy je následoval Wiliam Latham (1989), Karl Sims (1991), Christa Sommererová a Laurent Mignonneau (1994-95), John McCormack (1995) a Philip Galanter (1996). Společným rysem jejich práce je spoléhání na emergenci, kdy ze spolupráce prvků řízených jednoduchými pravidly se vynořuje stále komplikovanější celek.
Stačí připomenout slova Bodenové, která o paralele tvorby složité umělecké struktury a růstu biologických organismů nemá pochyb: „Klíčovým rysem biologických organismů je jejich schopnost zkonstruovat sebe samotné. Samo-organizace je spontánní emergence řádu z počátku, který je uspořádán na nižším stupni. Je to matoucí, dokonce kvazi-paradoxní vlastnictví. A není samozřejmé, že by se to mohlo stát s neživými věcmi. Celkem vzato, samo-organizace je kreativním fenoménem (Boden 2016: 112).“ Principy umělého života tak připravují stále se otevírajícího pole možností, které nicméně provázejí všudypřítomná uzpůsobující pravidla. Podobnou cestou se ubírá i moje vlastní umělecká práce.

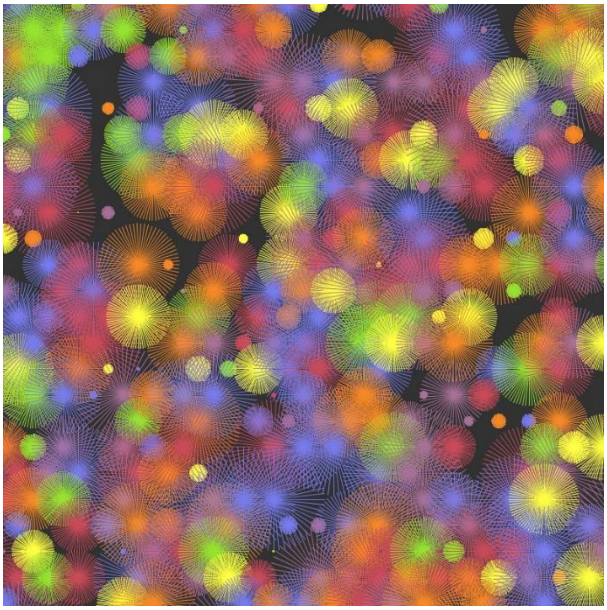


Obr. 9. Aleš Svoboda: Flora gama bw 21-09-18, 2021, digitální tisk, 500 x 900 mm.



Obr. 10. Aleš Svoboda: Flora gama bw 21-11-11, 2021, digitální tisk, 900 x 900 mm.

[2] Obraz se vydražil v aukčním domě Christies v New Yorku za částku 432 500 dolarů.



Obr. 11. Aleš Svoboda: Flora delta, 2022

5 Závěr – perspektivy svobodné tvorby

Pokud příklady AARONa a Portrétu Edmonda de Belamy svědčí o možnosti podílu specifického softwaru na kumulování poznatků a zapojení rozhodovacích procesů při simulaci umělecké tvorby, jednoznačně zároveň poukazují na limity vynucené průběžným lidským dohledem. Proces tvorby může v dílčích aspektech probíhat nepředvídaně, ale v globálním výsledku je očekávatelný, stabilní a setrvalý. Naopak program vystavený na principech umělého života naplňuje požadavek otevřeného systému, který umožňuje stále se obnovující emergenci.

Předvídat umělecké potřeby společenského vývoje bude zřejmě ještě dlouho nemožné. A to především proto, že společenskou úmluvu o potřebnosti daných projevů provází značná dávka libovolnosti, ovšem následně ji kodifikuje vynucené učení. Nicméně naplnění estetických potřeb ve smyslu formalistického poměru složitosti a uspořádání pro požadavky lidského vnímání bude zřejmě možné vyladěním komplexního vývojového chování procesů odpovídajících umělému životu.

Poděkování

Tento příspěvek vznikl za podpory programu COOPERATIO Univerzity Karlovy v rámci dílčího programu Art and Culture Studies.

Literatura

[1] H. Cohen: A Self-Defining Game for One Player: On the Nature of Creativity and the Possibility of Creative Computer Programs, *Leonardo* 1(2002) 59–64. První verze článku byla představena na třetí Konferenci tvořivost a poznání v Loughborough University, U. K. v roce 1999 a publikována ve sborníku z konference (New York: ACM Press, 1999).

[2] Margaret A. Boden: *AI, Its nature and future***,** Oxford University Press, Oxford, 2016

[3] P. Galanter: Generative Art Theory. In *A Companion to Digital Art* (Ch. Paul, ed.), John Wiley & Sons, Inc. Chichester, 2016

[4] L. Manovich: Abstraction and complexity. *Abstraction Now*, Edition Camera Austria, Graz, 2004

[5] Vladimír Mařík, Olga Štěpánková, Jiří Lažanský (eds.): *Umělá inteligence (1)*. Academia, Praha, 1993.

[6] J. L. Stephensen: Towards a Philosophy of Post-creative Practices? – Reading Obvious' "Portrait of Edmond de Belamy", [online]. 2019 [cit. 19. 11. 2021] Dostupné z: https://www.researchgate.net/publication/337891233_Towards_a_Philosophy_of_Post-creative_Practices_-_Reading_Obvious%27_Portrait_of_Edmond_de_Belamy.

COVID-19 pandemic may have changed our attitudes to science, but not our ability to reason scientifically

Jakub Šrol & Vladimíra Čavojová

Institute of Experimental Psychology, Centre for Social and Psychological Sciences, SAS
Dúbravská cesta 9, 841 04 Bratislava
jakub.srol@savba.sk, vladimira.cavojova@savba.sk

Abstract

In this paper we examined the scientific literacy of two representative samples of Slovak population and compared the results between a data collection from 2017 and in 2020 – before and after COVID-19 outbreak. Scientific literacy consists of three facets: knowledge of science facts, scientific reasoning and understanding how science works, which we measured by Scientific literacy scale (SLS; Miller, 1998; National Science Board, 2010), Scientific reasoning scale (SRS; Bašnáková et al., 2021; Drummond & Fischhoff, 2017), and Antiscientific attitudes scale (from CART; Stanovich et al., 2016) respectively. Altogether 1513 adult people (N_{2017} = 1012, N_{2020} = 501) aged between 18 and 86 (*M* = 41.13, *SD* = 15.91) participated in the study. The results obtained during the COVID-19 pandemic showed that, in comparison with the first data collection, antiscientific attitudes were significantly lower, and knowledge of science facts was slightly higher, which could be attributed to general increase of interest in understanding how the new virus works and how it differs from bacteria. On the other hand, the scientific reasoning stayed the same and it raises the concern about the ability of people to interpret and evaluate scientific evidence.

1 Introduction

Outbreak of COVID-19 by the end of 2019 and during the early spring of 2020 affected the whole world waiting for the vaccination against it and many experts had predicted that it would lead to more respect for science and scientists. Two years later we do not need the research to tell us that this was just wishful thinking from the part of the scientists. However, back in the spring 2020 we were among those who set to examine the question of how will the encounter with the coronavirus change the scientific literacy of people in Slovakia. We were in the unique position, because three years prior to the outbreak we had explored the three components of scientific literacy for our validization study (Bašnáková et al., 2021). The motivation for the presented paper was to compare the results obtained in 2017 (pre-pandemic) and in 2020 (during the pandemic) and examine how attitudes towards science, scientific knowledge and ability to reason scientifically changed after the event that caused many people to realize that the only way out of the disaster would be based on the pace in which the scientists will be able to understand the viral threat and come with the cure and/or vaccination.

2 Components of scientific literacy

Scientific literacy is a broad concept, often confused with other related constructs and it is not always clear on what aspect (e.g. intellectual, societal, attitudinal, axiological) of scientific activity researchers focus (Fasce & Picó, 2019). Scientific literacy has at least three dimensions (Fasce & Picó, 2019; Miller, 1983; 1998; 2004): (1) knowledge of scientific theories/science vocabulary/conceptual understanding, (2) the understanding of scientific reasoning/scientific method, and (3) trust in science and its values/understanding of science as an organized endeavor/understanding of the importance of science.

Besides the question what scientific literacy actually is, we need to address the question why it is important and why should we be concerned that it is rather low in general public. Scientific literacy implies that statements that we encounter daily in media should be based on some available and reliable evidence that is plausible explanation of observed data. In another words, functional understanding of science could be defied also in this way: majority of citizens will never do science as their profession. However, every person has to function as a citizen and to be able to make informed decision they need to be scientifically literate (Trefil, 2008). Because also in daily life it is important to understand connection between some assertion (theory, hypothesis) and evidence for the given assertion.

There is evidence that the question of public scientific literacy deserves special attention in Slovakia. First of all,

the PISA testing, which compares high-school children primarily in OECD countries on various tests shows that Slovak children scored below average of OECD in scientific literacy in the years 2006, 2009, 2012, 2015, and 2018 (Miklovičová & Valovič, 2018). We also have a lot of indirect evidence when it comes to adult population that people in Slovakia have a hard time to assess the quality of evidence and thus may be especially susceptible to some types of epistemically suspect beliefs. For example, a report by Globsec (2020) showed that among the 10 surveyed countries in Central and East Europe, Slovakia was the country with the highest susceptibility to conspiracy theories. Indeed, psychological studies conducted in Slovakia during the COVID-19 pandemic showed that various conspiracy and/or pseudoscientific beliefs about COVID-19 were endorsed by between 5 – 35% of the representative samples of the population. Unsurprisingly, pseudoscientific and conspiracy beliefs both in general and specifically related to COVID-19 pandemic were previously shown to be associated with lower scientific literacy and/or scientific reasoning (Allum et al., 2008; Čavojová et al., 2022; Kahan et al., 2012). Based on this direct and indirect evidence of a need for better scientific literacy among the Slovak public, we believe it is important to continuously survey the three aspects of scientific literacy and try to examine the factors that may affect scientific literacy (including exogenous factors such as COVID-19 pandemic).

3 Methods

3.1 Participants

Sample 1 (from 2017) was recruited through a market research agency to be representative of the Slovak general population with respect to age, gender, education and geographic location, who filled out the questionnaire online. Overall 1012 people (510 men and 502 women) with mean age 39.2 (*SD* = 15.6) years participated; 8.8% of participants had only elementary education (9 years of schooling), 31.7% of participants had vocational education (12yrs), 39.4% had full secondary education (13-14yrs), 5.4% of participants had a Bachelor´s degree, 13.7% of participants had a Master's degree and 0.9% of participants obtained a post-graduate degree.

Sample 2 (from 2020) was recruited through a market research agency to be representative of the Slovak general population with respect to age and gender, who filled out the questionnaire online. In total, we collected the data of 501 participants (241 men, 260 women) who were between 18 and 85 years old (M = 45.05, SD = 15.92). Most of the participants reported having high school diploma education (73.5 %), a smaller part had some college/university education (17.8 %), and the rest finished elementary education or high school education without a diploma (8.8 %).

3.1.1 Materials

The *Scientific knowledge* (SK) was a 9-item true/false questionnaire on basic scientific facts, such as "*Antibiotics kills viruses as well as bacteria*", based on National Science Indicators (National Science Board, 2010; Miller, 1998). We used a composite score as an index of public comprehension of science (Allum et al., 2008; Kahan et al., 2012).

Scientific reasoning. We used a 6-item adaptation of the Scientific Reasoning Scale (Bašnáková et al., 2021), which is adapted from Drummond and Fischhoff (2017). Items presented short scenarios contained the following validity threats: causation vs correlation, confounding variables, construct validity, control group, ecological validity, random assignment to conditions and participants indicated true/false answers to the final claims made in these scenarios. Higher number of correctly answered items reflected better scientific reasoning ability.

Anti-Scientific Attitudes subscale from CART developed by Stanovich et al. (2016) was used to measure anti-scientific sentiments, as the public attitudes about science are not only a matter of understanding how science works but also of trust in scientist and regulatory authorities (Allum et al., 2008). Participants had to indicate their agreement with 13 items on a 6-point scale ranging from 1 (*strongly disagree*) to 6 (*strongly agree*), with statements such as "I don't place great value on 'scientific facts', because scientific facts can be used to prove almost anything". A higher score indicated a stronger anti-scientific attitude, and thus lower trust in science and understanding of the importance of science.

4 Results & Discussion

To examine how the general public's scientific literacy changed after the outbreak of the COVID-19 pandemic we performed the t-test for two independent samples and we compared their score in scientific knowledge, scientific reasoning and attitudes toward science (Table 1).

Table 1. T-test for independent samples comparing samples from 2017 and 2020 in three components of scientific literacy

	Sample 1 (2017), N = 1012		Sample 2 (2020), N = 501				
	M	*SD*	*M*	*SD*	*t*	*p*	*d*
scientific knowledge	3.01	1.20	3.43	1.16	-6.42	< .001	0.35
scientific reasoning	4.16	1.46	4.14	1.40	0.20	0.85	-0.01
anti-scientific attitudes	3.34	0.58	3.17	0.70	4.73	< .001	-0.27

Note. The table shows descriptives for the three components of scientific literacy in two samples and the results of an independent t-test for their comparisons along with Cohen's *d* as a measure of effect size.

The results showed that antiscientific attitudes were significantly lower, and knowledge of science facts was slightly higher in the dataset from 2020 compared to the data collected three years earlier. The higher knowledge of science facts in the sample from 2020 could be attributed to general increase of interest in understanding how the new virus works and how it differs from bacteria. When we analyzed differences in individual items of Scientific knowledge, the number of people who correctly responded to question "Antibiotics kill viruses and bacteria" was higher: 46 % in 2017 sample and 75 % in sample 2020 ($p < .001$). Differences in other items ranged from 1 to 4% and were not significant. Although both differences were highly significant, they were of small to moderate effect size. On the other hand, the scientific reasoning was almost exactly the same across the two samples and it raises the concern about the ability of people to interpret and evaluate scientific evidence. This is important, as the studies on scientific reasoning show that it is a strong predictor of having less misconceptions, unfounded, and pseudoscientific beliefs – both generic, and specifically related to COVID-19 (e.g., Čavojová et al., n.d., 2020). Moreover, although the percentage of people in the category of lowest trust in science decreased from 2017 to 2020 by 7.4 %, there is still 25.1 % of people in this category, while there are only 16.8 % people in the category of the highest trust in science (Figure 1).

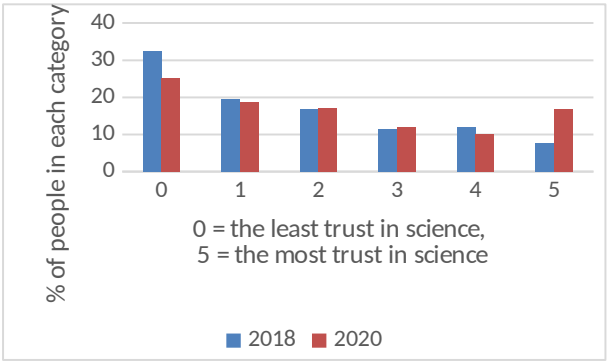


Fig. 1 Comparison of anti-scientific attitudes between 2017 and 2020

While during the time of our second data collection (April 2020) these results seemed at least cautiously promising and many of us harbored the hope that decrease in anti-scientific attitudes could be beneficial side effect of the pandemic, as the pandemic unrolled it became increasingly clear that unless we are able to address the actual ability of people to evaluate evidence (scientific reasoning), the distrust of science would prevail at the end of the day. Moreover, it is rather unwise for society to rely on external events that would change people´s thinking instead of investing to better education of scientific thinking in schools.

Importantly, we realize that the comparison of the three facets of scientific literacy between the representative sample collected in 2017 and 2020 does not give us a strong evidence for the effect of COVID-19 pandemic per se on scientific literacy. Of course, many other factors could have changed between these two data collections which could likewise affected the comparisons of scientific literacy. However, it should be mentioned that the level of scientific reasoning surveyed in the two samples was practically identical. This again gives us some indirect evidence for the comparability between the two samples. From among the three facets of scientific literacy, scientific reasoning is without the doubt the most resistant to any change. While attitudes and knowledge of facts are relatively malleable factors (which may be precisely why they could respond even to exogenous influences such as COVID-19 pandemic), increase in scientific reasoning requires long-term education efforts.

5 Conclusion

In this paper we examined how attitudes towards science, scientific knowledge and ability to reason scientifically changed after the COVID-19 outbreak in Slovakia. Our data showed significant but small positive effects on the two components of scientific literacy: scientific knowledge and attitudes toward science. However, the

most important component – scientific reasoning – remained unchanged. We conclude that external events, even negative ones, can have some beneficial effect, such as learning more about viruses (which increased the score of scientific knowledge) and realizing the importance of science for daily life, however, to make any lasting changes, we need to address the ability of people to reason scientifically and to evaluate evidence.

Funding

The study was supported by the Slovak Research and Development Agency as part of the research project APVV-20-0335 and by the scientific grant agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic as part of the project VEGA 2/0053/21.

References

Allum, N., Sturgis, P., Tabourazi, D., & Brunton-Smith, I. (2008). Science knowledge and attitudes across cultures : a meta-analysis. *Public Understanding of Science*, *17*, 35–54. https://doi.org/11.1077/0963662506070159

Bašnáková, J., Čavojová, V., & Šrol, J. (2021). Does concrete content help people to reason scientifically? *Science & Education*, *30*(4), 809–826. https://doi.org/https://doi.org/10.1007/s11191-021-00207-0

Čavojová, V., Šrol, J., & Ballová Mikušková, E. (2022). How scientific reasoning correlates with health-related beliefs and behaviors during the COVID-19 pandemic? *Journal of Health Psychology*, *27*(3), 534–547. https://doi.org/10.1177/1359105320962266

Čavojová, V., Šrol, J., & Jurkovič, M. (2020). Why should we try to think like scientists? Scientific reasoning and susceptibility to epistemically suspect beliefs and cognitive biases. *Applied Cognitive Psychology*, *34*(1), 85–95.

Čavojová, V., Šrol, J., & Mikušková, E. B. (n.d.). With the little help of science understanding: Examining the direct and indirect role of scientific reasoning and trust in science in normative health behaviour during pandemic. *Preprint.* https://doi.org/10.31234/OSF.IO/XAHDJ

Drummond, C., & Fischhoff, B. (2017). Development and Validation of the Scientific Reasoning Scale. *Journal of Behavioral Decision Making*, *30*(1), 26–38. https://doi.org/10.1002/bdm.1906

Fasce, A., & Picó, A. (2019). Science as a Vaccine. *Science & Education*, 1–17. https://doi.org/10.1007/s11191-018-00022-0

Globsec. (2020). *Voices of Central and Eastern Europe: Perceptions of democracy & governance in 10 EU countries - GLOBSEC*. https://www.globsec.org/publications/voices-of-central-and-eastern-europe/

Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L. L., Braman, D., & Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature Climate Change*, *2*(10), 732–735. https://doi.org/10.1038/nclimate1547

Miklovičová, J., & Valovič, J. (2019). *Národná správa PISA 2018*. NÚCEM.

Miller, J. D. (1983). Scientific Literacy: A Conceptual and Empirical Review. *Daedalus*, *112*(2), 29–48. https://doi.org/10.2307/20024852

Miller, J. D. (1998). The measurement of civic scientific literacy. *Public Understanding of Science*, *7*, 203–223. https://doi.org/10.1088/0963-6625/7/3/001

Miller, J. D. (2004). Public Understanding of, and Attitudes toward, Scientific Research: What We Know and What We Need to Know. *Public Understanding of Science*, *13*(3), 273–294. https://doi.org/10.1177/0963662504044908

National Science Board. (2010). Science and Engineering Indicators 2010. In *Transactions of the American Society of Civil Engineers*. National Science Foundation. https://doi.org/10.1080/03602458208079655

Stanovich, K. E., West, R. F., & Toplak, M. E. (2016). *The Rationality Quotient: Toward a test of rational thinking*. MIT Press.

Trefil, J. (2008). *Why science?* Teachers College Press.

Etické aspekty neurorobotických simulácií

Martin Takáč[1,2], Alistair Knott[1,3], Mark Sagar[1,4]

[1] Soul Machines Ltd, Auckland, Nový Zéland
[2] Centrum pre kognitívnu vedu FMFI, Univerzita Komenského v Bratislave
[3] Department of Computer Science, University of Otago, Dunedin, Nový Zéland
[4] Auckland Bioengineering Institute, University of Auckland, Nový Zéland
{martin.takac,alistair.knott,mark.sagar}@soulmachines.com

Abstrakt

V príspevku analyzujeme etické aspekty interagovania s realistickými avatarmi – simulátormi inteligentného správania v ľudskej podobe. Zameriame sa najmä na „ubližujúce správanie“ zo strany užívateľa: čo to znamená a kedy je eticky sporné, buď z dôvodu dopadu na užívateľa alebo na samotný simulátor. Zatiaľ čo dopady na užívateľa sú najmä otázkou empirického skúmania, morálny status simulátora je v doméne filozofie mysle. V práci predstavíme morálny behaviorizmus Johna Danahera a budeme s ním polemizovať: pre morálny status simulátora je podstatný spôsob/implementácia jeho vtelenia. Naše tézy budeme exemplifikovať na prípade BabyX – hyperrealistického simulátora malého dieťaťa (cca 1,5-2r) vyvinutého firmou Soul Machines. Príspevok je skrátenou slovenskou verziou článku Knott, A., Sagar, M., Takac, M.: The ethics of interaction with neurorobotic agents: a case study with BabyX. AI Ethics, 2021.

1 Úvod

V tomto príspevku sa zameriame na systémy umelej inteligencie, ktoré majú podobu ľudí, či už stelesnených v robotickom hardvéri alebo simulovaných na obrazovke príp. vo virtuálnej realite v podobe tzv. avatarov. Napriek tomu, že takéto simulácie nie sú dokonalé, dosahujú vysoký stupeň realistickosti,[1] čo otvára otázku etického statusu simulovanej osoby (Sparrow, 2004).

V produkcii humanoidov sa uplatňuje buď *inžiniersky* prístup, ktorého hlavným cieľom je vybudovať umelo-inteligentné agenty schopné spolupracovať s ľuďmi, alebo *vedecký*, ktorý študuje ľudskú kogníciu budovaním simulovaného mozgu v simulovanom tele. Takéto simulácie budeme ďalej nazývať *neurorobotickými*. Etické otázky spojené s prvým prístupom zahŕňajú najmä bezpečnosť produktu, nebezpečenstvo zneužitia a širší dopad na spoločnosť. U druhého prístupu je namieste otázka, aké sú implikácie toho, že produkt sa snaží reprodukovať funkcionalitu ľudského či iného biologického mozgu a tela. V článku sa pokúsime prispieť k tejto debate analýzou konkrétneho neurorobotického systému BabyX vyvíjaného spoločnosťou Soul Machines. BabyX je virtuálna simulácia približne 18-mesačného dieťaťa (pozri Obr. 1). Obsahuje vysoko realistický grafický komponent (počítačovo animovaný 3D model tváre a tela s množstvom správaní) napojený na kognitívnu architektúru inšpirovanú ľudským mozgom. Architektúra pozostáva z viacerých umelých neurónových sietí a vzájomne spriahnutých dynamických systémov. Z vedeckého hľadiska je BabyX platforma pre neurorobotický výskum – umožňuje nám implementovať stelesnené modely mechanizmov ľudskej kognície a testovať ich pozorovaním, či simulované správanie virtuálneho dieťaťa je podobné správaniu skutočných detí. Zatiaľ čo naše kognitívne modely sú ešte stále pomerne jednoduché, BabyX je vizuálne vcelku presvedčivá simulácia skutočného dieťaťa:[2] vidí a počuje používateľa prostredníctvom kamery a mikrófonu, vníma zdieľané simulované prostredie, v ktorom môže manipulovať s objektami, môže sa učiť slová a činnosti. Prejavuje aj emocionálne správanie v reakcii na udalosti, ktoré vníma: môže sa usmievať, smiať, plakať, rozčúliť sa alebo prejaviť frustráciu.

V nasledujúcej kapitole predstavíme základný etický rámec a rozanalyzujeme, čo budeme myslieť pod zlým zaobchádzaním s (umelým) dieťaťom. Ďalšie dve kapitoly rozoberajú etické aspekty podľa dopadov na používateľa a na samotné simulované dieťa. Kapitola 4 tiež detailnejšie predstavuje tie aspekty systému BabyX, ktoré sú relevantné pre etickú problematiku, vrátane jeho biologicky inšpirovaného modelu emócií. Kapitola 5 je zhrnutím.

[1] Viď napr. https://robots.ieee.org/robots/geminoiddk, http://www.geminoid.jp, https://www.soulmachines.com

[2] Viď napr. https://www.soulmachines.com/icdl2021-demo

2 Morálne subjekty a objekty

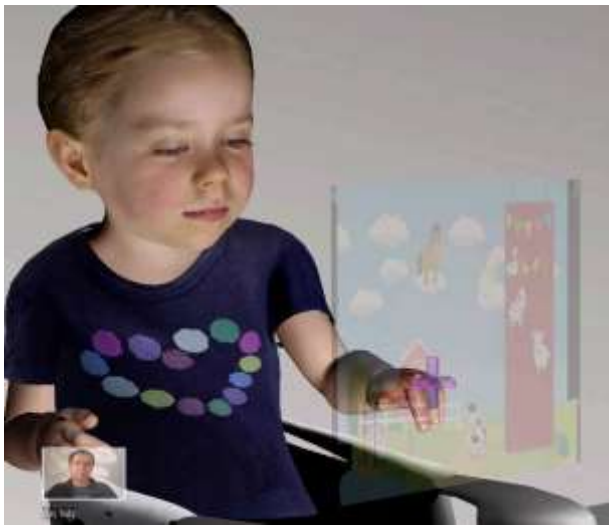
V morálno-etických diskusiách sa obvykle rozlišuje status *objektu morálky* (*moral patiency*) a status *subjektu morálky* (*moral agency)*. Morálnym objektom je entita, voči ktorej majú iní (ľudia) morálne záväzky. Subjekt morálky má sám morálne záväzky voči iným (Gunkel, 2012). Pre účely tohto článku sa v ďalšom uvažovaní obmedzíme na otázku, či má neurorobotický humanoid/avatar status morálneho objektu, inými slovami, či sa na správanie používateľov voči nemu vzťahujú nejaké morálno-etické obmedzenia. Neurorobotický systém BabyX je obzvlášť relevantný pre kladenie si takýchto otázok, pretože modeluje zraniteľného člena spoločnosti – dieťa. Ak si však chceme klásť otázku, či je etické zle zaobchádzať s virtuálnym dieťaťom, musíme spresniť, čo myslíme pod pojmom „zlé zaobchádzanie“ aj v kontexte bežných výchovných interakcií reálnych rodičov a detí.

2.1 Zlé zaobchádzanie s deťmi

Nie každá situácia, v ktorej dospelí vyvolávajú alebo umožňujú u detí negatívne emócie, je zlým zaobchádzaním resp. ubližovaním. V bežnom živote deti zažívajú rôzne emócie, ktoré sú len čiastočne pod kontrolou dohliadajúcich dospelých. Je dokonca veľmi dôležité, aby bábätká za určitých okolností zažívali negatívne emócie, napríklad keď sa učia metódou pokusov a omylov. Rodičovská úloha formovať a vychovávať si niekedy vyžaduje, aby v dieťati vyvolali negatívne emócie, a to pokarhaním rôzneho druhu. Rodičia majú tiež nechať deti naučiť sa určité veci „samé“, t. j. poskytnúť im pri učení autonómiu. Rodič, ktorý u dieťaťa vyvoláva negatívne emócie ako súčasť odôvodneného učenia, alebo ktorý nezabráni negatívnej emócii tým, že umožní dieťaťu konať samostatne v kontrolovanom bezpečnom prostredí, nie je vinný zo zlého zaobchádzania. Sme presvedčení, že takisto interakcie s virtuálnym dieťaťom BabyX, ktoré spadajú do týchto kategórií, sú úplne prijateľné: jednoducho simulujú bežný život dieťaťa. BabyX sa môže napr. rozrušiť, ak sa mu nepodarí splniť úlohu, o ktorú sa pokúša. Môže tiež rozpoznať „karhavé“ výroky používateľa na základe zvukových profilov v oblasti intenzity a farby hlasu. Používateľ ho môže takýmito výrokmi odradiť od činnosti, ktorú vykonáva.

V kontexte vedeckých experimentov sú obvykle považované za prijateľné aj interakcie, ktoré bábätkám experimentálne navodia určité (veľmi malé!) množstvo stresu za účelom získania vedeckého poznania. Príkladom je experimentálna paradigma „kamenná tvár“ (stillface, Weinberg *a kol.,* 2008), pri ktorej sa dospelý krátky čas pozerá na dieťa bez akejkoľvek viditeľnej reakcie. Toto neobvyklé rodičovské správanie spôsobuje deťom stres, ale detské reakcie poskytujú vhľad do roly rodičovských výrazov tváre vo vnímaní a prežívaní sveta (Adamson & Frick, 2003). Preto ani obdobné experimenty s virtuálnym dieťaťom nepovažujeme za eticky sporné.

Po vylúčení horeuvedených prípadov sa zameriame na „ozajstné“ zlé zaobchádzanie, ktoré by sa v prípade skutočných detí považovalo za neetické. Ale je neetické aj ak sa deje voči virtuálnemu dieťaťu? Argumentáciu rozčleníme podľa toho, či je založená na dopadoch zlého zaobchádzania s avatarom na používateľa, alebo na (potenciálnych) dopadoch na avatara samotného.



Obr. 1. Používateľ a avatar zdieľajú virtuálne prostredie, v ktorom môžu manipulovať s objektami.

3 Etika podľa účinku na používateľa

Pre účely argumentácie v tejto kapitole predpokladajme, že BabyX nie je morálnym objektom – je to „len počítačový program.“ Aj v tomto prípade má zmysel uvažovať o etickej signifikancii užívateľovho správania voči BabyX – kvôli dopadom na samotného užívateľa. Kľúčové je, že interakcia s BabyX sa v mnohých aspektoch podobá na interakciu s reálnym dieťaťom. Pokiaľ zlé zaobchádzanie s avatarom z užívateľovej perspektívy pripomína zlé zaobchádzanie s reálnym dieťaťom, je to z etického hľadiska problém?

3.1 Dopady zlého zaobchádzania

Existujúce publikácie venujúce sa zlému zaobchádzaniu s avatarmi sa sústreďujú na sex roboty (Danaher, 2017;

Sparrow, 2017), šikanovanie (Keijsers & Bartneck, 2018), či zabíjanie postáv v počítačových hrách (Luck, 2009). Najčastejším argumentom je, že ak užívateľ ubližuje avatarovi, ovplyvní to jeho budúce správanie voči reálnym ľuďom (Danaher, 2020). Hypotéza, že násilné zaobchádzanie s avatarmi v simuláciách alebo videohrách sa prenáša na agresívne alebo antisociálne návyky voči ľuďom v reálnom živote, bola skúmaná v mnohých empirických experimentoch. Výsledky však nie sú jednoznačné: dokonca aj na úrovni veľkých metaanalýz existujú štúdie, v ktorých sa našli dôkazy o prenose (pozri napr. Anderson, *a kol.*, 2010; Greitemeyer, 2019; Greitemeyer & Mügge, 2014), ale aj štúdie, ktoré nachádzajú len minimálne dôkazy (napr. Ferguson, 2015; Drummond *a kol*., 2020).

Mnohé štúdie však ukazujú, že to, nakoľko sa agresivita prenesie do reálnych interakcií, závisí od stupňa grafickej realistickosti zobrazenia (Barlett & Rodeheffer, 2009), stupňa ponorenia hráča do hry (Kim & Sundar, 2013; Persky & Blascovich, 2007) a realistickosti správania hrových postáv (Zendle *a kol.*, 2018). BabyX skóruje vysoko vo všetkých troch aspektoch: grafické zobrazenie dieťaťa BabyX je veľmi realistické, správanie dieťaťa je vierohodne simulované, a to na úrovni jednotlivých gest, ako aj na úrovni väčších jednotiek správania. Užívateľ je v grafickom rozhraní fyzicky veľmi blízko k dieťaťu, v pozícii podobnej pozícii rodiča alebo opatrovateľa.

Existuje však aj opačná možnosť, a to, že zlé zaobchádzanie s avatarom môže viesť k pozitívnemu účinku: používatelia, ktorí majú sklon k týraniu, by si mohli tento sklon uspokojiť na avataroch a vyhnúť sa tak ubližovaniu skutočným ľuďom. Danaher (2017) tvrdí, že možnosť terapeutického využitia avatarov týmto spôsobom „by sa mala aktívne a starostlivo skúmať“.

3.2 Dopady pozitívnych interakcií

Prirodzené a neubližujúce interakcie s BabyX môžu mať vysoko pozitívne dopady: môžu napr. slúžiť na výučbu princípov rodičovskej starostlivosti, ako simulátor umožňujúci zažiť si interakciu s bábätkom pre budúcich rodičov alebo v kurzoch rodičovstva pre tínedžerov. Dokonca aj pokiaľ dieťa prejavuje bolesť alebo utrpenie, vieme si predstaviť simulátor, v ktorom sa medici učia ako utrpenie zmierňovať.

Aj pozitívne interakcie môžu však mať negatívny dopad na používateľa. Dôležitou možnou škodou, ktorú treba zvážiť, je, že používateľ sa príliš citovo zaangažuje. Tento scenár je pravdepodobný najmä v prípade používateľov, ktorí nedávno stratili dieťa alebo nemôžu mať vlastné deti: sú emocionálne zraniteľní a existuje riziko, že sa pripútajú k niečomu, čo im nemôže city skutočne opätovať (Bryson, 2010).

3.3 Správanie používateľa z hľadiska etiky cností

Odvodzovanie etického statusu zaobchádzania s avatarmi z dopadov na používateľa patrí do etiky utilitarizmu. Z hľadiska iných paradigiem, napr. normatívnej etiky alebo etiky cností (Boddington, 2017) treba posudzovať používateľovo správanie samo osebe , nezávisle od účinku na kohokoľvek. Etika cností sa nezaoberá správnym konaním, ale skôr správnym charakterom. Sparrow (2021) tvrdí, že znásilniť alebo mučiť avatara je vždy nemorálne, pretože to implikuje niečo o charaktere aktéra: aký charakter má človek, ktorý mučí a znásilňuje?

4 Etika podľa účinku na avatara

V tejto kapitole sa zameriame na to, či môže mať zlé zaobchádzanie s virtuálnym dieťaťom eticky relevantný negatívny účinok na neho samotné (bez ohľadu na účinok na používateľa alebo iných ľudí). Najčastejšie sa status objektu morálnych obligácií odvodzuje od schopnosti cítiť (angl. *sentience*): pokiaľ entita dokáže cítiť, resp. trpieť, je nemorálne jej ubližovať. Otázka schopnosti umelých systémov cítiť, resp. mať mentálne stavy, spadá do filozofie mysle, má však etické dôsledky. Za východisko nášho ďalšieho uvažovania si vyberieme etickú koncepciu Johna Danahera.

4.1 Danaherov etický behaviorizmus

Danaher (2020) nazýva svoju koncepciu morálnych záväzkov voči avatarom etickým behaviorizmom, pretože ich odvodzuje od pozorovateľného správania: ak je správanie avatarov dostatočne podobné správaniu ľudí, potom to samo osebe stačí na to, aby im bol priznaný morálny status. V skutočnosti je jeho argument ešte širší: ak je správanie avatara dostatočne blízke akémukoľvek agentovi, ktorému priznávame morálny status, potom by sme mali aj avatarovi priznať rovnaký morálny status.

Danaher hovorí, že jeho tvrdenia o mentálnych stavoch avatarov nie sú *metafyzické*, ale *normatívne* a *epistemické*. Síce pripúšťa, že konečný dôvod na poskytnutie morálneho statusu ľuďom (a zvieratám) je to, že sú schopní cítiť, teda že majú určitú formu vedomia, ale dôkazy o tom máme vždy iba nepriame, prostredníctvom ich správania, čiže akéhosi etického Turingovho testu. Ak je správanie robota „približne performatívne ekvivalentné“ správaniu človeka, mali by sme robotovi priznať rovnaký morálny status ako majú ľudia. (A ak sú približne performatívne ekvivalentné s niektorými nižšími živočíchmi, ktorým priznávame znížený morálny status,

mali by sme robotovi priznať rovnaký morálny status ako má tento živočích.)
Danaherova definícia pozorovateľného správania je však výrazne širšia ako u behavioristov začiatku 20. storočia, pretože zahŕňa aj stavy mozgu, resp. mozgovú aktivitu, nielen vonkajšie pohybové reakcie tela na podnety. Kvôli prehľadnosti budeme rozlišovať dve verzie etického behaviorizmu: *úzku*, podľa ktorej vonkajšie fyzické správanie agenta je dostatočné na rozhodnutie, či máme voči nemu etické povinnosti, a *širokú*, podľa ktorej sa rozhodujeme na základe fyzického správania a pozorovateľných vnútorných (mozgových alebo počítačových) stavov. Široká verzia etického behaviorizmu umožňuje vyjadriť funkcionalistický opis pocitov a iných eticky relevantných duševných stavov na základe vnútorných mechanizmov, ktoré generujú agentovo správanie. Sem spadá napríklad Putnamova klasická funkcionalistická definícia mentálneho stavu „bolesť“ (Putnam, 1967). Podľa neho agent schopný pociťovať bolesť musí mať určitú *funkčnú organizáciu*, minimálne so *senzormi* na detekciu určitých podnetov, mechanizmami na priraďovanie *hodnoty* podnetom (aj prostredníctvom naučených asociácií) a mechanizmami na vytváranie *správania* na základe týchto hodnôt.
Napriek tomu, že Danaher teoreticky pripúšťa širokú verziu etického behaviorizmu, je skeptický voči jej možnosti uspieť. Špeciálne varuje pred „biologickým mysterianizmom“, teda pripisovaniu špeciálneho statusu biologickým organizmom. Rovnako je však skeptický voči Putnamovskému funkcionalistickému popisovaniu mozgových mechanizmov pomocou *algoritmov*: vieme tak málo o tom, ako stavy mozgu súvisia s morálne relevantnými metafyzickými stavmi typu vedomé vnímanie, že je lepšie založiť epistemické dôkazy o morálnych stavoch agenta priamo na pozorovateľnom správaní bez akéhokoľvek odkazovania na stavy mozgu. S touto pozíciou chceme polemizovať.

4.2 Vnútorná organizácia BabyX

Začneme opisom vlastností, ktoré robia z BabyX zaujímavého kandidáta na etické úvahy. V prvom rade, BabyX nie je len neurálny model, ale je avatar, teda má okrem simulovaného mozgu aj graficky realistické renderovanie tváre a tela. Je vnorený v prostredí a interaguje s ľuďmi (na rozdiel od algoritmov, ktoré spracúvajú dátové množiny). Takých systémov existuje veľa (napr. Chevalier-Boisvert a kol., 2019; Oudeyer, 2017), BabyX je však výnimočné stupňom realistickosti zobrazenia aj prirodzenosťou interakcií. Potiaľ relevancia pre Danaherov argument pozorovateľného správania, poďme však dovnútra.

4.2.1 Stelesnený kognitívny model

BabyX implementuje stelesnený/vtelený model ľudskej kognície. Paradigma vtelenej kognície predpokladá, že štruktúra kognitívneho systému je silne ovplyvnená štruktúrou senzomotorického aparátu (Ballard a kol., 1997; Clark, 1997). Napríklad to, že oko má foveu, ktorá vníma predmety postupne a nie všetky naraz, má dôležité dôsledky pre architektúru kognitívneho systému. Rovnako aj skutočnosť, že motorické pohyby rúk sú zvyčajne riadené vizuálnymi fixáciami tak, že siahnutie rukou po predmete si zvyčajne vyžaduje najprv vizuálne zameranie pozornosti. V systéme BabyX simulujeme procesy ako vizuálna pozornosť a vizuomotorická koordinácia, ktoré posielajú vnemy a reaferentné kópie motorických signálov do simulovaného mozgu s podobnými časovými charakteristikami, aké sa nachádzajú v skutočnom ľudskom mozgu. Simulujeme telo BabyX, pretože si myslíme, že nám to pomôže navrhnúť model mozgu s plauzibilnejším rozhraním k vonkajšiemu svetu. Ale skutočnosť, že BabyX má telo, môže mať tiež etický význam. V našom modeli môžeme napríklad simulovať hmatové mechanizmy, prostredníctvom ktorých sa na tele registrujú príjemné alebo bolestivé podnety, a motorické mechanizmy, ktoré na takéto podnety reagujú, napr. prostredníctvom reflexov spätného rázu alebo zľaknutia. Intuitívne cítime odlišnosť týchto mechanizmov od abstraktnejšej odmeny a trestu v algoritmoch učenia posilňovaním.

4.2.2 Biologicky inšpirovaný model emócií

Azda najrelevantnejší pre etickú diskusiu je vtelený model emócií systému BabyX inšpirovaný prácami Pankseppa (1998) a Damasia (2010). Evolučne najstaršie emocionálne obvody ľudského mozgu, nachádzajúce sa u všetkých cicavcov, prebiehajú cez mozgový kmeň a hypotalamus a sú reflexné – prepájajú zmyslové alebo interoceptívne podnety priamo na fyzické správanie (Panksepp, 1998). Tieto obvody zabezpečujú homeostázu: udržiavajú agenta sýteho, zdravého, konkurencieschopného, a mimo nebezpečenstva. V BabyX implementujeme reflexné obvody pre „priblíženie / záujem“, „radosť“, „strach“, „hnev“, „stres“ a „zaskočenie“. Tento súbor je mierne odlišný od Pankseppovho, ale vytvára podnety a reakcie relevantné pre detského avatara. Reflexnosť je implementovaná pomocou pravidiel „ak-potom“, napríklad reflex „priblíženie/záujem“ je spúšťaný ľudskou rečou s určitou farbou a výškou tónu. Prejavuje sa spustením viečok a úsmevom.
Tieto nízkoúrovňové emočné obvody tvoria základ pre sofistikovanejšiu emočnú nadstavbu. Behaviorálne reakcie

vyvolané týmito reflexnými obvodmi zahŕňajú signály pre agentovo telo prostredníctvom uvoľňovania rôznych neurochemických látok v hypotalame. Napríklad „priblíženie/záujem“ spúšťa uvoľňovanie dopamínu a oxytocínu, „radosť“ vyvoláva uvoľňovanie dopamínu; „strach“ a „hnev“ vyvolávajú uvoľňovanie noradrenalínu a kortizolu. Aktivita v týchto okruhoch je kľúčová pri definovaní cieľového stavu agenta. Napríklad dopamínový okruh nielenže definuje jednu zo základných emócií agenta, ale aj riadi učenie agenta operačným podmieňovaním, takže agent je motivovaný robiť veci, ktoré vedú k určitým emóciám.

Aktivita v šiestich základných reflexných emocionálnych okruhoch aktivuje vektor 8 neurochemických koncentrácií, ktoré označujeme ako neurochemický stav agenta. Tento stav moduluje mnohé aspekty správania agenta. Napríklad kortizol zvyšuje srdcovú a dychovú frekvenciu BabyX, zatiaľ čo oxytocín tieto frekvencie znižuje; norepenefrín zvyšuje parameter zrýchlenia motorických pohybov BabyX, takže sú náhlejšie a trhavejšie. To znamená, že emocionálne správanie BabyX je emergentným efektom komplexného súboru mozgových mechanizmov. Okrem toho sa emocionálne správania agenta navzájom inhibujú prostredníctvom okruhov výberu činností v bazálnych gangliách, čo vedie k ďalším emergentným efektom. Táto vzájomná inhibícia spolu s kontinuálne sa meniacim neurochemickým stavom môže viesť k náhlym nespojitým zmenám v pozorovateľnom správaní, aké vidíme u malých detí i u zvierat.

Doteraz opísané emočné okruhy sú z veľkej časti subkortikálne. U cicavcov existujú aj vysokoúrovňové emočné okruhy zahŕňajúce mozgovú kôru. V mozgovej kôre sa reprezentujú objekty, ľudia, udalosti a situácie. Tieto reprezentácie sú asociované s emočnými stavmi pomocou vstupov z podkôrových okruhov a neurochemického stavového vektora. Zároveň však emočný stav spätne ovplyvňujú, čím vzniká dynamický systém so spätnoväzobnými slučkami. Agentov súčasný emočný stav ovplyvňuje budúce stavy, takže prechádza trajektóriou v stavovom priestore. V stavovom priestore existujú atraktory zodpovedajúce vyšším emóciám, ktoré sú u ľudí vyjadriteľné v jazyku.

Model kortikálnych emócií výrazne rozširuje priestor emocionálnych stavov a pridáva ďalšie kognitívne dimenzie k neurochemickému priestoru definovanému podkôrovým systémom. Napríklad emocionálny stav „trucovania“ modelujeme ako subkortikálny pocit „hnevu“, spojený s tendenciou zdržať sa konania. „Nostalgiu“ modelujeme ako podkôrový pocit „radosti“ v spojení s kognitívnou operáciou spomínania si na udalosti z epizodickej pamäte (viď nižšie). Stav „zmätku“ modelujeme ako podkôrový pocit „úzkosti“ spojený s nízkym stupňom dôvery v nejaký kognitívny úsudok, napríklad predikovaný následný stav alebo udalosť (vyjadrený ako rozdelenie pravdepodobnosti nad možnými udalosťami, podmienené postupnosťou nedávnych udalostí).

4.2.3 Model epizodickej pamäte

Kognitívny model BabyX zahŕňa model epizodickej pamäte (jeho staršia verzia je opísaná v Takáč & Knott, 2016). BabyX si dokáže zapamätať udalosti a stavy, ktoré zažíva, a dokáže si z pamäti neskôr vybaviť sekvenčne štruktúrované série udalostí a stavov. Na základe spomienok dokáže znovu rozpoznať opakujúce sa situácie, ako aj predpovedať nasledujúce udalosti a stavy. Takisto sa učí emocionálne asociácie udalostí, to znamená, že znovurozpoznanie alebo predikcia vyvolá emóciu, ktorá sprevádzala udalosť pri jej prvom prežívaní. Emócie ovplyvňujú aj zapamätávanie udalostí: udalosti so silným emočným nábojom sa zapamätávajú s vyššou hodnotou sily učenia (podobne je to aj s udalosťami, ktoré sú neočakávané, keď nesúlad pozorovania s predikciou vyvolá emóciu prekvapenia). Agenty s pamäťou sú bližšie k statusu objektu morálky ako agenty bez pamäte, keďže môžu predvídať negatívne udalosti alebo si spomínať na traumatické zážitky. Okrem toho, podľa Damasia (2010) je epizodická pamäť nevyhnutná na to, aby sa vyvinulo sebauvedomenie (pozri aj Tulving, 2002), ktoré je samo osebe dôležitým kritériom morálneho statusu.

4.2.4 Vyvíjajúca sa kognícia

BabyX je vyvíjajúcou sa kogníciou v dvoch zmysloch slova: na jednej strane obsahuje algoritmy učenia a zrenia inšpirované detskou vývinovou psychológiou, na druhej strane je dlhodobo vyvíjaným a postupne zlepšovaným softvérovým produktom. Z tohto hľadiska je vhodné zamýšľať sa nad etickými otázkami nielen vzhľadom na jeho súčasnú podobu, ale premýšľať aj o potenciálnych dôsledkoch jeho možnej budúcej podoby.

4.3 Je vnútorná organizácia dôležitá?

Napriek všetkým vyššie opísaným vlastnostiam nechceme tvrdiť, že BabyX (alebo iný neurorobotický systém disponujúci podobnými vlastnosťami) má status morálneho objektu. Chceme skôr posunúť uvažovanie o etických otázkach z polaritného buď-alebo, kde kognitívny systém buď je, alebo nie je morálnym objektom, ku pohľadu kontinua, kde na jednom konci spektra sú neživé objekty (kameň, kalkulačka) a na druhom cítiace bytosti (zvieratá a človek). Niekde na

tomto kontinuu sú umiestnené neurorobotické systémy (a zrejme sa budú v budúcnosti po ňom posúvať), preto má zmysel klásť si otázky, ktoré vlastnosti dávajú umelým systémom vyšší status z hľadiska morálnych obligácií.
Túto časť zakončíme jednoduchým hypotetickým príkladom: Predstavme si na jednej strane avatara, ktorého vyjadrovanie emócií riadi vyššie opísaný biologicky inšpirovaný stelesnený model emócií. Predpokladajme, že stlačenie klávesu spôsobí simulovaný bolestivý podnet, ktorý vyvolá aktivitu v receptoroch bolesti. Tie následne spôsobia zmeny hodnôt neurotransmiterov (napr. zvýšenie hodnoty reprezentujúcej stresový hormón kortizol). To vyvolá kaskádu zmien v rôznych častiach simulovaného mozgu a tela s emergentným efektom. Dynamické systémy jednotlivých komponentov tela prepojené na počítačovú grafiku spôsobia, že avatar sebou šklbne, nervy inervujúce svaly na tvári spôsobia bolestivý výraz tváre a avatar vykríkne. Teraz si predstavme iného avatara, ktorý má rovnakú vizuálnu podobu, ale nemá simulovaný mozog a stlačením zmieneného klávesu sa len spustí prednahratá počítačová animácia, v ktorej avatarovým telom šklbne, nadobudne bolestivý výraz tváre a vykríkne. Z hľadiska pozorovateľného správania v zmysle úzkej verzie etického behaviorizmu sú oba systémy nerozlíšiteľné. Z hľadiska širokej verzie má však prvý systém k statusu morálneho objektu bližšie ako druhý, ktorého správanie bolo iba „predstieraním/falzifikátom“.

5 Záver

V práci sme sa pokúsili na konkrétnom príklade neurorobotickej simulácie virtuálneho dieťaťa BabyX predstaviť rámec umožňujúci klásť si otázky morálneho statusu takýchto systémov a našich etických obligácií. Časť argumentácie sme postavili na účinkoch interakcií s vizuálne realistickými humanoidnými neurorobotickými systémami na ľudí samotných. V druhej časti článku sme skúmali, aké vlastnosti neurorobotického systému ho posúvajú smerom k statusu morálneho objektu. Keďže vývoj takýchto systémov sa nezastaví, zvažovanie etických dôsledkov ich budovania bude naberať na dôležitosti.

Literatúra

Adamson, L., Frick, J. (2003). Research with the face-to-face still-face paradigm: a review. *Infancy* 4, 451–473.

Anderson, C., Shibuya, A., Ihori, N., Swing, E.L., Bushman, B.J., Sakamoto, A., Rothstein, H.R., Saleem, M. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in Eastern and Western countries. *Psychol. Bull.* 136(2), 151–173.

Ballard, D., Hayhoe, M., Pook, P., Rao, R. (1997). Deictic codes for the embodiment of cognition. *Behav. Brain Sci.* 20(4), 723–767.

Barlett, C., Rodeheffer, C. (2009). Effects of realism on extended violent and nonviolent video game play on aggressive thoughts, feelings, and physiological arousal. *Aggress. Behav.* 35, 213–224.

Boddington, P. (2017). *Towards a Code of Ethics for Artificial Intelligence*. Springer.

Bryson, J. (2010). Robots Should Be Slaves. In: Y. Wilks (ed.): *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*. John Benjamins.

Chevalier-Boisvert, M., Bahdanau, D., Lahlou, S., Willems, L., Saharia, C., Nguyen, T.-H., Bengio, Y. (2019). BabyAI: a platform to study the sample efficiency of grounded language learning. arXiv:1810.08272.

Clark, A. (1997). *Being there: putting brain, body and world together again*. MIT Press, Cambridge.

Damasio, A. (2010). *Self comes to mind: constructing the conscious brain*. Vintage, London.

Danaher, J. (2017). Robotic rape and robotic child sexual abuse: should they be criminalised? *Crim. Law Philos.* 11, 71–95.

Danaher, J. (2020). Welcoming robots into the moral circle: a defence of ethical behaviourism. *Sci. Eng. Ethics* 26, 2023–2049.

Drummond, A., Sauer, J., Ferguson, C. (2020) Do longitudinal studies support long-term relationships between aggressive game play and youth aggressive behaviour? A meta-analytic examination. *R. Soc. Open Sci.* **7**, 200373.

Ferguson, C. (2015). Do angry birds make for angry children? A metaanalysis of video game influences on children's and adolescents' aggression, mental health, prosocial behavior, and academic performance. *Perspect. Psychol. Sci.* 10(5), 646–666.

Greitemeyer, T. (2019). The contagious impact of playing violent video games on aggression: Longitudinal evidence. *Aggress. Behav.* 45, 635–642.

Greitemeyer, T., Mügge, D. (2014). Video games do affect social outcomes: a meta-analytic review of the effects of violent and prosocial video game play. *Pers. Soc. Psychol. Bull.* 40(5), 578–589.

Gunkel, D.J. (2012): *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. MIT Press.

Keijsers, M., Bartneck, C. (2018). Mindless robots get bullied. In: *Proceedings of HRI'18,* March 5th–8, 2018, Chicago, IL, USA.

Kim, K.-J., Sundar, S. (2013). Can interface features affect aggression resulting from violent video game play? An examination of realistic controller and large screen size. *Cyberpsychol. Behav. Soc. Netw.* 16(5), 329–334.

Luck, M. (2009). The Gamer's dilemma. *Ethics Inf. Technol.* 11(1), 31–36.

Oudeyer, P.-Y. (2017). What do we learn about development from baby robots? *Wiley Interdiscip Rev Cogn Sci.* 8(1-2).

Panksepp, J. (1998). *Affective neuroscience: the foundations of human and animal emotions*. Oxford University Press, New York.

Persky, S., Blascovich, J. (2007). Immersive virtual environments versus traditional platforms: effects of violent and nonviolent video game play. *Media Psychol.* 10(1), 135–156.

Putnam, H. (1967). The nature of mental states. In: Putnam, H. (ed.) *Mind, Language, and Reality*, 429–440. Cambridge University Press, Cambridge.

Sparrow, R. (2004). The Turing triage test. *Ethics Inf. Technol*. 6, 203–213.

Sparrow, R. (2017). Robots, rape, and representation. *Int. J. Soc. Robot*. 9, 465–477.

Sparrow, R. (2021). Virtue and Vice in Our Relationships with Robots: Is There an Asymmetry and How Might it be Explained? *International Journal of Social Robotics* 13, 23-29.

Tulving, E. (2002). Episodic memory: from mind to brain. *Annu. Rev. Psychol.* 53, 1–25.

Weinberg, M. K., Beeghly, M., Olson, K. L., & Tronick, E. (2008). A Still-face Paradigm for Young Children: 2½ Year-olds' Reactions to Maternal Unavailability during the Still-face. *The Journal of Developmental Processes*, 3(1), 4–22.

Zendle, D., Kudenko, D., Cairns, P. (2018). Behavioural realism and the activation of aggressive concepts in violent video games. *Entertain. Comput.* 24, 21–29.

Embodied ideas

Silvia Tomašková

Katedra filozofie, Filozofická fakulta, Univerzita Konštantína Filozofa v Nitre, Hodžova 1, 949 01 Nitra

silvia@libris.sk

Abstract

Paper aims at reconsidering the „4 E“ (embodied, embedded, enactive, extended) model of cognition for the study of human mind. Importance of the embodiment thesis is pointed out especially for understanding the relationship between mind, thought and language. Contrary to traditional approaches (disembodied mind, transcendental reason), concepts and categorization are considered as deeply embedded in the world. Argumentation is supported by research in cognitive linguistics (Gibbs, Lakoff-Johnson), cognitive psychology (Rosch), philosophy of language (Kövecses) and philosophy (Metzinger). Author claims that mutual interdependence between perception, language and knowledge changes radically our epistemological perspective on *attunement* with others and the surrounding world.

1 Introduction

The meaning and use of the term „embodiment“ has undergone a revision in the development of both cognitive science and philosophy. After the decline of popular analogy mind/software – brain/hardware together with representational-computationist theory of mind, since 1970´s, second generation of cognitive scientists emphasized the need of a new approach in the study of cognition. Concept of „embodied cognition“ has been introduced and worked out into a „4 E“ (embodied, embedded, enactive, extended) model of cognition for the study of human mind. According to the model, states of cognition (language, reason, feelings, images, ideas etc.) are considered as deeply dependent or strongly influenced by the *integral* activity of the brain, body and an agent in the surrounding reality. More precisely, all components of the „4 E“ model are considered as playing a significant causal role in cognitive processing.

In the following text I will point out partial, though crucial implications of the proposed model for human conceptualisation and knowledge. Sketched approach results from the application of three interrelated facts: 1) developmental nature of language acquisition, 2) unconscious character of cognitive operations and 3) metaphorical nature of language and thought. Cognitive linguists thus reopened central philosophical questions such as Who am I? Where does meaning come from? How do we conceptualize What are the mechanisms of conceptualisation? What is thinking?

2 No(body) - no mind

Why is it so common to feel that our concepts reflect the world “as it is”, that our categories of mind fit the categories of the world? One reason is that we have been “thrown into a world” and with the help of complex evolutionary mechanisms adapted to it quite well. Besides survival needs we also strived for understanding the surrounding reality and searched for the meaning of it all. We have also evolved to acquire language and to categorize. The process of language development gave rise to an important class of categories that optimally fit our bodily experiences of entities and certain important differences in the natural environment - called basic-level categories. The basic level, is not just about objects (ball), but also about social concepts (families, clubs) social actions (arguing) and basic emotions (happiness, anger). Basic-level categories are the source of our most stable knowledge, and the technological capacity to extend them allows us to extend this knowledge.

Spatial concepts, such as “front”, “back”, “up”, and “down”, provide clear examples in which embodied experience exists. These concepts are articulated in terms of our body's position in, and movement through space and make sense of space for us. We use spatial-relations concepts (near, far) unconsciously, and we impose them via our perceptual and conceptual systems. Not surprisingly, for the most part categorization is not a product of conscious reasoning. Importance of unconscious cognition, the processing of perception, memory, learning, thought, and language without our awareness has been highlighted by introducing the concept of *cognitive unconscious.*

Claims about embodied character of our conceptualization and the structure of language are supported by

experimental and theoretical research (Grady, 1999; Naranayan, 2000). Behavioral and neural evidence has shown that the process of language comprehension activates motor simulations and involves motor systems. Experimental research demonstrates how acquisition of words from early infancy is based on physical sensations, on our feeling bodies. theory of conflation in the course of learning. For young children, subjective (nonsensorimotor) experiences and judgments, on the one hand, and sensorimotor experiences, on the other, are so regularly conflated in experience, that for a time children do not distinguish between the two when they occur together (Johnson, 1997). For example, for an infant, the subjective experience of affection is typically correlated with the sensory experience of warmth, the warmth of being held. During the period of conflation, associations are automatically built up between the two domains. Later, during a period of differentiation, children are then able to separate out the domains, but the cross-domain associations persist. These persisting associations are the mappings of conceptual metaphor that will lead the same infant, later in life, to speak of "a warm smile," "a big problem," and "a close friend." Associations made during the period of conflation are realized neurally in simultaneous activations that result in permanent neural connections being made across the neural networks that define conceptual domains. These connections form the anatomical basis of source-to-target activations that constitute metaphorical entailments. Because of the way neural connections are formed during the period of conflation, we all naturally think using hundreds of primary metaphors (*Time is money., She is an early bird.)*.

3 Rethinking metaphors

We use metaphors automatically and unconsciously simply by functioning in the most ordinary of ways in the everyday world from our earliest years. This applies, especially, when we speak about our inner experience such as following : „love is a journey“, „their marriage became a nightmare“, „he is out of his mind“, „she´s her own worst enemy“, „if I were you, I´d hate me“ „„You are the light in my life“, „She is my better half“, „He is wasting my time“, „We have arrived at the crucial point in the argument“, „I can´t follow you“, „I see what you´re saying“. Thus, states of mind are vastly conceptualized in bodily terms: grasping ideas, reaching conclusions, being unclear etc. In their thought provoking book *Philosophy in the Flesh,* G. Lakoff and M. Johnson (1999) presented a novel approach on the nature of mind and thought. Inspired by the work of L. Wittgenstein (1953) on the character of philosophical language and categorization (concept of *family resemblance*) they have worked out in detail how unconscious processes shape and structure all conscious thought. Embodiment in their own terms has been identified strongly with bodies (sensorimotor experience) and neuronal activity of the brain. This claim followed the research of the psychologist E. Rosch and her colleagues on natural categories.

3.1 Prototypes, container schemas, basic metaphors

According to Rosch, human categories are typically conceptualized in more than one way, in terms of what are called prototypes (Rosch, 1981). She has criticized the classical theory on the nature of concepts (Aristotle, Plato) based on an idea that concepts can be expressed in terms of their defining properties, which are the necessary and sufficient attributes that items must have to be instances of the concept. Rosch, in contrast, proposed a theory, in which instances of a natural concept are defined by their resemblance to a prototype. A prototype is understood as a best or most typical example of the concept, sharing the maximum number of features or attributes with other instances and a minimum number with instances of other concepts. Thus a prototype consists of characteristic features rather than defining properties, and according to this interpretation concepts have indistinct boundaries and may be represented by fuzzy sets. Most people, for example, regard robins or pigeons as good examples of birds category, so pigeons and robins are more prototypical of a bird than f.e. chicken. Although everyone knows that chicken is a bird, for some reason it has a less privileged status than other birds.

Furthermore, each prototype is a neural structure that permits us to do some sort of inferential or imaginative task relative to a category. Typical-case prototypes are used in drawing inferences about category members in the absence of any special contextual information (f.e. typical husband). Ideal-case prototypes allow us to evaluate category members relative to some conceptual standard (f.e. ideal husband).

The richness and embodied character of metaphors in our use of language is elaborated in the *conceptual theory of metaphor* by Lakoff and Johnson. Mechanisms of conceptual metaphor work, according to authors, on the basis of using the „logic of physical“ to describe the inner „mental realm“. In order to illustrate their approach in more detail they adopted the concept of *container schema* as a useful tool for analysing the mind. Via metaphor the mind is conceptualized in terms of a container image, which is given an inside, a boundary and an outside. Internaly *felt* ideas and concepts expressed in language refer to the things in the external (physical) world.

Container schema has a gestalt structure, that is to say, that the parts make no sense without the whole. There is no inside without a boundary and an outside, no outside without a boundary and an inside, and no boundary without sides. The structure is topological in the sense that the boundary can be made larger, smaller, or distorted and still remain the boundary of a container schema. Even if a container schema is conceptual, it can be physically instantiated, either as a concrete object, like a room or a cup, or as bounded region in space, like a basketball court or a football field. A physical boundary can impose forceful and visual constraints: It can protect the container's contents, restrict their motion, and render them inaccessible to vision.

For Lakoff, the mind has been conceptualized in bodily terms due to *mapping* across conceptual domains. The first domain incloses the target (tenor) – the subject to which attributes are ascribed and the second domain incloses the vehicle - the object whose attributes are borrowed or transfered. This process can be ilustrated by the metaphor *Well-Functioning Mind is a Healthy Body* as following: 1. domain: well-functioning mind is a target; 2. domain: healthy body is a vehicel. Ideas are food, acquiring idea is eating, helpful ideas are nutritious foods, disturbing ideas are disgusting foods, fully comprehending is digesting and communicating is feeding. Ideas conceptualized as an appetite for food, for learning, raw facts are not suitable because they are not digestible. Digestion is the full „mental processing“ required for understanding. The metaphor *Well-Functioning Mind is a Healthy Body* presents criteria for acceptability of an idea – it has to smell good, be cooked properly etc.

4 Perspectives

The conceptual theory of metaphor together with „4 E“ model shed light on the way we perceive the world through our senses. Together they intend to replace a traditional picture of cognitive processing based on an image of a so called „input-output“ model suggesting a static picture of an agent on one side and surrounding world on the other. This perspective can be illustrated by a number of philosophical theories on perception in philosophy of mind which concentrate is either on the side of subject (internalism) or on the side of object reality (externalism). According to an „integrated embodied dynamic systems“ model, an agent, her/his body and surrounding environment form an integrated whole Fig. 1. (Thelen 2000).

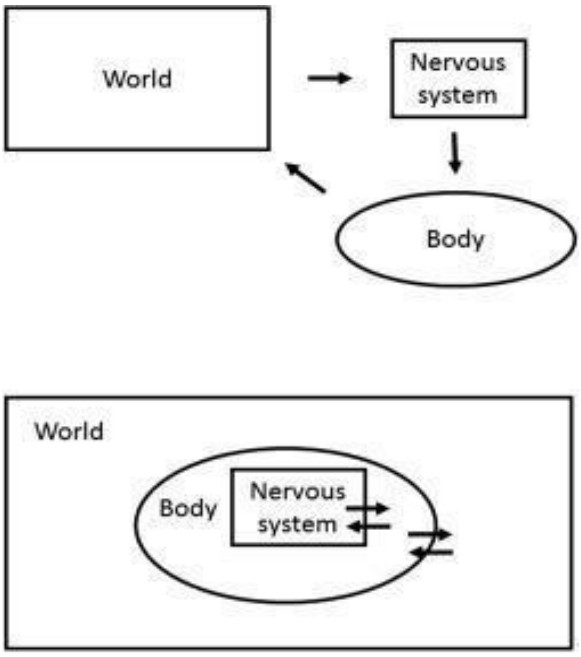


Fig. 1. Thelen's (2000) schematic overview of the contrast between an "input–output" model of human cognitive processing (top panel) and an integrated embodied dynamic systems model (bottom panel).

Ultimately, seemingly paradoxical consequence of the above mentioned model is concerned with our experience of the world as such. The existence of outside world – objective reality is indubitably real. But in moving through this world we constantly apply unconscious filter mechanisms and unknowingly construct our own world. The construction process is largely invisible, we see only what our reality tunnel allows us to see and most of us are completely unaware of this fact (Metzinger, 2010,9).

5 Summary

According to the conceptual theory of metaphor, human concepts are not just reflections of an external reality, but that they are crucially shaped by our bodies and brains, especially by our sensorimotor system. Importance and implications of the „4E“ model are of course much more complex and significant than indicated above. Here are at least few outcomes we have to reconsider: a) our ideas, thoughts, desires are embodied, b) when talking about mental states and events we talk about (mostly unconscious) processes, c) rationality of men and free agency is constrained by our bodies embedded in reality, d) the nature of an asymmetry between consciously felt experience and theoretical explanation is primarily epistemological, e) having direct access to causes of our behavior, to the nature of our „mind´s I and subjective

feelings is an illusion, f) metaphorical thought is the principal tool of philosophical insight.

The shift in understanding of mind and reason brought by cognitive linguists entails a corresponding shift in our understanding of what we are as human beings. Therefore, the proposed model requires rethinking of a traditional model of a person, in which mind and reason have been considered as mainly conscious, literal and unconstrained by the body. What we now know about the mind from contemporary experimental findings and theoretical research is at odds with many classical philosophical views of what a person is. In accordance with this proposal we should be sceptical about any account which argues for the existence of a *direct conscious access* to experience itself and to most of our thought.

Metaphorical language *shows vividly* vagueness of this picture and somehow *hides* the fleshy nature of conscious states „from our own sight". Finally, no matter how non-intuitively it may sound, our ideas, feelings and most intimate thoughts are, similarly as seeing the sun rise or sunset, natural phenomena. The felt duality of our own identity together with experiencing the world as independent from our minds is a result of a capacity of our own consciousness to make reality appear within itself.

References

[1] R. Gibbs, H. Colston: *Interpreting figurative meaning.* Cambridge University Press, Cambridge, 2012.

[2] J. Grady: *The Conduit Metaphor Revisited: A Reassessment of Metaphors for Communication*. In: J. P. Koenig, ed., Discourse and Cognition: Bridging the Gap. Stanford: CSLI/Cambridge, 1998.

[3] C. Johnson: The Acquisition of the "What's X Doing Y?" Construction. In E. Hughes, M.
Hughes, and A. Greenhill, eds., *Proceedings of the Twenty-First Annual Boston University*
Conference on Language Development 2: 343-353. Somerville, Mass.: Cascadilla Press, 1997.

[4] Z. Kövecses: *Language,Mind and Culture*. Oxford, Oxford University Press, 2010.

[5] G. Lakoff, M. Johnson: *Metaphors we live by.* Chicago University Press, Chicago, 1980.

[6] G. Lakoff, M. Johnson: *Philosophy in the Flesh. The Embodied Mind and Its Challenge to Western Thought*. Basic Books, New York, 1999.

[7] T. Metzinger: *The Ego Tunnel: The Science of Mind and the Myth of the Self*. Basic Books, New York, 2010 .

[8] S. Naranayan: Talking the Talk Is Like Walking the Walk: A Computational Model of Verbal Aspect. In M. G. Shafto and P. Langley, eds*., Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*. Mahwah, N.J.: Erlbaum, 1997.

[9] E. Rosch: Prototype Classification and Logical Classification: The Two Systems. In E. Scholnick, ed., *New Trends in Cognitive Representation: Challenges to Piaget's Theory*, Hillsdale, N.J.: Erlbaum, 1981: 73-86.

[10] E. Thelen: Grounded in the World: Developmental Origins of the Embodied Mind. *Infancy 1*, (2000): 3–28.

[11] L. Wittgenstein: *Philosophical Investigations*. New York: Macmillan, 1953.

Přehled obecných přístupů k vyhodnocování inteligence umělých systémů

Ondřej Vadinský[0000−0002−0910−3140]

Katedra informačního a znalostního inženýrství VŠE v Praze
Náměstí Winstona Churchilla 4, 130 67 Praha 3, ČR
E-mail: ondrej.vadinsky@vse.cz

Abstrakt

Cílem obecné umělé inteligence je vytvořit počítačové systémy, které budou schopné řešit mnoho různých a nepředvídaných úloh. Proto jsou potřeba vhodné obecné metody vyhodnocování inteligence umělých systémů. Tento přehledový článek hledá takové metody mezi přístupy vycházejícími z algoritmické teorie informace. Za tímto účelem je provedena rešerše literatury, představeny a srovnány existující přístupy. Výhodou těchto přístupů je vysoká míra formalizace, pevné teoretické základy a nízká míra antropocentričnosti. Limitem je pak vysoká výpočetní náročnost testů plynoucí z jejich obecnosti, ale i to, že některé pokročilejší teoretické návrhy dosud nebyly implementovány do podoby prakticky proveditelných testů.

1 Úvod

Ač otázka *„Jak poznat a vyhodnotit, zda je umělý systém inteligentní?“* stála již na samém počátku oboru umělá inteligence (Turing, 1950), nebyla dosud uspokojivě vyřešena. Sama umělá inteligence tak má několik šířeji přijímaných avšak navzájem ne zcela kompatibilních definic (Wang, 2019). Následkem toho výzkum v oboru často sklouzává k řešení dílčích problémů a konkrétních úloh, což sice přináší zajímavé technologie, nicméně, nakolik nás to přibližuje k vytvoření skutečně inteligentních systémů, zůstává nejasné.

Oblast *obecné umělé inteligence* (Goertzel a Pennachin, 2007) usiluje o vytvoření takových umělých systémů, které dokáží řešit široké spektrum úloh v mnoha různých kontextech a to bez dalších zásahů svých tvůrců. Jde tak vlastně o návrat k původním idejím disciplíny tradičně také známým pod jménem *silná umělá inteligence* (Searle, 1980) či nověji také označovaným jako *umělá inteligence lidské úrovně* (Minsky a spol., 2004). Právě pro takové pojetí umělé inteligence je problematika solidně ukotveného vyhodnocování inteligence klíčovou.

Od průkopnické práce Hernandez-Orallo (2000) lze sledovat rozvoj snahy založit vyhodnocování inteligence umělých systémů na *algoritmické teorii informace* (Li a Vitányi, 2008). Výstupy lze rozdělit do několika kategorií: Sekce 2 představí *univerzální inteligenci* (Legg a Hutter, 2007b) jako příklad formálních definic; Sekce 3 shrne *kdykoliv přerušitelný test inteligence* (Hernández-Orallo a Dowe, 2010) jako ukázku formálních návrhů testů; a konečně Sekce 4 popíše *test algoritmického IQ* (Legg a Veness, 2013) jakožto zástupce prakticky proveditelných testů.[1] Kromě umělých systémů lze však zaměření testů a definic rozšířit na biologické systémy a disciplínu zobecnit do podoby *univerzální psychometrie* (Hernández-Orallo, 2017). Nabízí se tak nová perspektiva na zbývající otevřené problémy i další otázky, jak ukáže Sekce 5.

2 Definice univerzální inteligence

Inteligence je poměrně obecný a obtížně uchopitelný pojem, proto Legg a Hutter (2007a) prozkoumali řadu definic, teorií a testů zaměřených na inteligenci u lidí i zvířat. Zobecněním tohoto přehledu dospěli k následující neformální definici inteligence: *„Inteligence měří schopnost agenta dosahovat cílů v mnoha různých prostředích.“* Formální vyjádření uvedené pracovní definice zachycené v rovnici (1) Legg a Hutter (2007b) nazvali *univerzální inteligencí*.

Legg a Hutter (2007b) uvažují interakci agenta π s prostředím μ probíhající po krocích, ve kterých agent zasílá akce a_i a od prostředí dostává odměny r_i a pozorování o_i. Definice nijak neomezuje uvažované agenty, aby však umožnila testování na počítačích předpokládá pouze taková prostředí, která lze popsat Turingovsky vyčíslitelnou pravděpodobnostní měrou.

$$\Upsilon(\pi) := \sum_{\mu \in E} 2^{-K(\mu)} V_\mu^\pi, \text{ kde } \quad V_\mu^\pi := \mathbb{E}\left(\sum_{i=1}^{\infty} r_i\right) \leq 1, \tag{1}$$

kde *univerzální inteligence* Υ agenta π je dána jeho schopností dosahovat cílů, kterou v konkrétním prostředí μ popisuje hodnotová funkce V_μ^π jakožto očekávanou sumu všech budoucích odměn. Definice pak kombinuje výsledky agenta z množiny všech uvažovaných prostředí E pomocí *algoritmické pravděpodobnosti* $2^{-K(\mu)}$ založené na *Kolmogorovově složitosti* (Kolmogorov, 1963). Následkem toho

[1] Podrobnější rozbor obecných přístupů k vyhodnocování umělých systémů podává Vadinský (2018b).

schopnost agenta dosahovat cílů v různých prostředích přispívá k celkové míře inteligence různou měrou: V souladu s *Occamovou břitvou* má úspěch agenta ve složitých prostředích menší váhu než v jednoduchých prostředích.

Univerzální inteligenci (Legg a Hutter, 2007b) lze také chápat jako zobecnění *C-testu* (Hernandez-Orallo, 2000) ze statických úkolů na dynamická interaktivní prostředí. Ačkoliv je představená definice velmi obecná a má řadu žádoucích vlastností, lze si povšimnout i několika praktických limitací. Především využití *Kolmogorovovy složitosti* ji činí nevyčíslitelnou, což je dále umocněno i tím, že definice uvažuje nekonečnou množinu všech prostředí jakož i nekonečnou sekvenci interakcí agenta s prostředím. Každý praktický test odvozený z této definice tak bude nutně jen její aproximací.

Definice *univerzální inteligence* (Legg a Hutter, 2007b) není závislá na konkrétní kultuře, není vlastně vůbec antropocentrická, nicméně právě kvůli *Kolmogorovově složitosti* je závislá na zvoleném *referenčním Turingově stroji*. Hibbard (2009) ukázal, že to může mít závažné následky. Výsledná míra je totiž předpojatá k relativně malé množině prostředí popsaných krátkým programy, jež tak dominují celkovému hodnocení agenta. Změna referenčního Turingova stroje pak může prostředí spadající do této rozhodující množiny značně změnit a vést tak i k zásadním rozdílům v celkové míře inteligence agenta. Určitou ochranu před tímto problémem poskytuje arbitrární omezení přípustné minimální délky programů popisujících přípustná prostředí.

Hibbard (2009) také dokázal, že míry inteligence musí dávat prostředím různé váhy, aby se tak vyhnuly *No Free Lunch teorému*. Tuto podmínku *univerzální inteligence* splňuje právě díky použití *algoritmické pravděpodobnosti*.

Dalších limitů definice *univerzální inteligence* si všímá Goertzel (2010). Podnětnou je zejména jeho připomínka, že v reálných situacích neočekáváme od agentů skutečnou univerzálnost jako spíše dostatečně širokou (tedy vhodně předpojatou) obecnost. Ostatně ani člověk není zřejmě schopen vést si dobře ve všech možných prostředích (úlohách), ale spíše jen v sice rozsáhlé ale omezené množině prostředí (úloh).

3 Návrh kdykoliv přerušitelného testu inteligence

Hernández-Orallo a Dowe (2010) navrhli *kdykoliv přerušitelný test inteligence* jako test zaměřený na současné i budoucí, umělé i biologické agenty, přičemž by tento test měl zvládnout vyhodnotit jak agenty s libovolně vysokou (či nízkou) inteligencí, tak i agenty interagující se světem v libovolných časových měřítkách (krátkých i dlouhých). Návrh testu počítá s tím, že lze test kdykoliv přerušit, přičemž vydá tak přesnou aproximaci výsledku, jak mu dostupný čas na testování umožní. Návrh testu spojuje *univerzální inteligenci* (Legg a Hutter, 2007b), s dřívější prací autorů na *C-testu* (Hernandez-Orallo, 2000) a *Turingově testu rozšířeném o indukci a kompresi* (Dowe a Hájek, 1998).

Aby byl navržený test prakticky proveditelný, vypořádali se Hernández-Orallo a Dowe (2010) se třemi aspekty nevyčíslitelnosti *univerzální inteligence* následujícím způsobem:

- K aproximaci množiny všech prostředí využívají konečný vzorek. To však vznáší požadavek na *diskriminační sílu* prostředí v tomto omezeném vzorku, aby do něj vybíraná prostředí co nejvíce přispěla k vyhodnocení inteligence testovaného agenta. Hernández-Orallo a Dowe navrhují uvažovat pouze taková prostředí, která jsou *citlivá vůči odměně*, tedy ta, kde výběr akce agenta vždy ovlivní jeho budoucí odměny.

- Pro aproximaci nekonečné interakční sekvence uvažují konečný počet interakcí. To si žádá vhodný způsob agregace odměn do celkového skóre. Hernández-Orallo a Dowe navrhují zprůměrovat odměny počtem uskutečněných interakcí. To pak vyžaduje, aby prostředí byla *vyvážená*, tedy poskytovala odměny z intervalu $[-1; +1]$, což způsobí, že náhodné chování agenta povede k průměrné odměně okolo 0.

- Jako způsob aproximace *Kolmogorovovy složitosti* používají složitostní funkci $Kt^{\max}$ inspirovanou Levinovu Kt složitostí Levin (1973). Tato funkce je založena na horním odhadu výpočetního času potřebného pro jednu interakci a je dále omezena celkovým počtem interakcí. Funkce tak kromě využití jako distribuce prostředí zachovávající *Occamovu břitvu* také umožňuje vynutit časový limit výpočtu odměn a pozorování při interakci agenta s prostředím.

Zásadním přínosem navrženého testu je také to, že zahrnuje *fyzický čas* (Hernández-Orallo a Dowe, 2010). U prostředí je k tomu využitá zmíněná složitostní funkce $Kt^{\max}$. U agenta pak jde o stanovení celkového časového limitu testu a zahrnutí reakčních časů agenta do jeho celkového skóre. Následkem toho počet interakcí agenta s prostředím není fixován, nýbrž závisí na časovém limitu a reakčních časech agenta. Aby agent nemohl snadno podvádět, zahrnuje výpočet skóre také zpoždění mezi akcemi agenta.

Hernández-Orallo a Dowe (2010) dále navrhují, aby testovací procedura adaptovala složitost prostředí a časové limity na inteligenci i časové měřítko agenta. Test začíná s nízkou složitostí prostředí a krátkými časovými limity. Pokud agent nestihne včas reagovat, časový limit se zvyšuje. Pokud agent dosáhne přiměřeně vysokého skóre, dojde ke zvýšení složitosti prostředí.

Naopak pokud agent dosahuje přiměřeně nízkého skóre, dojde opět ke snížení složitosti. Vhodným vyvážením těchto mechanizmů pak lze předejít podvádění agenta.

Insa-Cabrera a spol. (2011) představili prototypovou implementaci *kdykoliv přerušitelného testu inteligence* a využili ji v jednoduchých experimentech s několika lidskými a umělými subjekty. Velmi zajímavou myšlenkou je využití *druhově specifických rozhraní* před stejným testem. Vytvořený prototyp je však příliš zjednodušenou variantou jinak velmi zajímavého návrhu.

4 Test algoritmického IQ

Legg a Veness (2013) vytvořili prakticky proveditelný test, který aproximuje *univerzální inteligenci* (Legg a Hutter, 2007b) a zahrnuje některé myšlenky z návrhu *kdykoliv přerušitelného testu inteligence* Hernández-Orallo a Dowe (2010). Legg a Veness převedli *univerzální inteligenci* popsanou rovnicí (1) do podoby uvedené v rovnici (2), kterou označují jako *algoritmické IQ*[2].

Aby překonali nevyčíslitelnost původní definice, *test algoritmického IQ (AIQ test)* (Legg a Veness, 2013) používá konečnou *délku epizody* o k krocích a konečný vzorek N *programů prostředí* p_i, které popisují prostředí. Způsob výběru programů do vzorku zachovává ideu *Occamovy břitvy* a vyhýbá se *Kolmogorovově složitosti* tím, že vychází ze Solomonoffovy *univerzální distribuce*: $M_{\mathcal{U}}(x) := \sum_{p:\mathcal{U}(p)=x*} 2^{-l(p)}$ (Solomonoff, 1964a,b). Použití této pravděpodobnostní distribuce umožňuje jednoduché generování programů do vzorku postupným přidáváním instrukcí, avšak jedno stejné prostředí může být popsáno vícero různými programy.

$$\hat{\Upsilon}(\pi) := \frac{1}{N}\sum_{i=1}^{N} \hat{V}_{p_i}^{\pi}, \text{ kde} \quad \hat{V}_{p_i}^{\pi} := \frac{1}{k}\sum_{i=1}^{k} r_i, \qquad (2)$$

AIQ odhad univerzální inteligence $\hat{\Upsilon}$ agenta π je dán jeho schopností dosahovat cílů popsaných empirickou hodnotovou funkcí $\hat{V}_{p_i}^{\pi}$ jako průměrná odměna dosažená agentem za k interakcí s programem prostředí p_i z konečného vzorku N programů.

Stejně jako v případě *univerzální inteligence* ovlivňuje volba *referenčního stroje* (jazyka programů prostředí) třídy programů, u kterých je pravděpodobné, že budou obsaženy ve vzorku (Legg a Veness, 2013). Aby byl tento problém minimalizován, používá *AIQ test* poměrně jednoduchý *BF referenční stroj* (Müller, 1993). *BF* je nízkoúrovňový jazyk zahrnují pouze 10 instrukcí, které úzce souvisí s ovládáním Turingova stroje, avšak programy mohou být nedeterministické (Legg a Veness, 2011).

[2]Legg a Veness (2013) označení IQ zvolili pro jeho asociaci s inteligencí, uvedená míra však není kvocientem.

Test splňuje požadavek na *vyvážená prostředí* (Hernández-Orallo a Dowe, 2010), přidělované odměny jsou totiž normalizovány do intervalu $[-100, +100]$, který také limituje dosažitelné AIQ skóre agenta. Způsob, jakým test funguje, zajišťuje, že náhodně se chovající agent dosáhne AIQ blízké 0 (Legg a Veness, 2013, 2011).

Limit na dobu trvání výpočtu jedné interakce zajišťuje, že jsou z testu vyřazeny programy, které by běžely příliš dlouho, nebo by vůbec neskončily. Výpočet programu prostředí je také ukončen ve chvíli, kdy se program pokouší zapsat více než nastavený počet pozorování. Zastoupení neinteragujících programů ve vzorku je dále umenšeno vyřazením programů bez instrukce ke čtení nebo zápisu a také těch programů, které vracejí konstantní odměny (Legg a Veness, 2013, 2011). Požadavek na vyloučení prostředí *bez diskriminační síly* (Hernández-Orallo a Dowe, 2010) je tak částečně splněn.

Legg a Veness (2013, 2011) použili řadu technik redukujících rozptyl a zrychlujících proces konstrukce odhadu AIQ. Implementovaný test je dostupný jako Open Source a lze jej nastavit několika způsoby: Volba počtu programů prostředí ve vzorku ovlivňuje přesnost odhadu AIQ. Nastavená délka epizody umožňuje zvýšit čas, který je agentu dostupný pro učení. Lze také změnit počet symbolů použitých BF strojem a počet vydaných pozorování, čímž je možné zvýšit komplexitu prostoru interakcí. Součástí zveřejněného testu je pak několik jednoduchých agentů využívajících principy posilovaného učení.

Představený *test algoritmického IQ* však není prost nedostatků. Jejich podrobnou analýzu předložil Vadinský (2018c), který také poskytl sadu návrhů, jak zjištěné nedostatky odstranit, a to včetně implementace vybraných z nich. Zejména jde o:

- Zvýšení výpočetní efektivity testovací procedury, které zároveň umožňuje zkoumat vývoj skóre agenta během testu a získat tak přesnější představu o jeho „rychlosti učení“.
- Implementaci limitu minimální délky programů prostředí, což umožňuje snížit závislost testu na zvoleném referenčním stroji, jak navrhl Hibbard (2009).

Analýzou programů prostředí používaných v *AIQ testu* se pak zabýval Vadinský (2018d). Náhodné vzorkování programů prostředí v původním testu Legga a Venesse způsobuje:

- častý výskyt *zbytečného kódu*, který programy činí nepřehlednými a zdánlivě komplexními;
- a pak zejména nezanedbatelné zastoupení programů prostředí *bez diskriminační síly*.

Implementovaná vylepšení testu pak tyto problémy do značné míry odstraňují. Rozšířený test algoritmického

IQ (Vadinský, 2018a) tak zůstává nadějným prakticky proveditelným přístupem k obecnému vyhodnocování inteligence umělých systémů.

5 Otevřené problémy a další otázky

Tato sekce uvede několik příkladů otevřených problémů, kterým přístupy k vyhodnocování inteligence umělých systémů čelí. Nutno podotknout, že nejde o úplný výčet.

Sekce 5.1 se zabývá alternativním pohledem na vymezení *obtížnosti prostředí*. Sekce 5.2 představuje snahu o vypracování podrobnější *teorie úloh*. Sekce 5.3 věnuje pozornost *omezeným zdrojům*, se kterými se musí inteligentní systém vyrovnat. Sekce 5.4 naznačuje možnost problematiku vyhodnocování inteligence umělých systémů uchopit v širším kontextu vznikající disciplíny *univerzální psychometrie*.

5.1 Obtížnost prostředí

Klíčovou myšlenkou představených přístupů vyhodnocujících inteligenci umělých systémů je způsob agregace výsledků agentů v jednotlivých prostředích zohledňující složitost těchto prostředí. Složitost se zde používá k hodnocení obtížnosti prostředí, tedy vlastně k hodnocení obtížnosti problému, který má agent v prostředí řešit.

Nový pohled na definici *obtížnosti prostředí* přináší Hernández-Orallo (2015). Namísto dosavadního ztotožnění obtížnosti prostředí se složitostí jeho popisu (neboli se zadáním problému) navrhuje zaměřit se na složitost politik úspěšných v daném prostředí (neboli možných řešení problému). Tedy obtížný problém není ani tak problém, který má složité zadání, ale problém, který je složité řešit. Přesněji, namísto složitosti nejjednoduššího zadání problému, jde o složitost nejjednoduššího řešení problému.

Hernández-Orallo (2015) zkoumal různá pojetí obtížnostních funkcí a přiklonil se k verzi inspirované Levinovu složitostí (Levin, 1973), která zohledňuje jak délku řešení, tak i počet potřebných výpočetních kroků. To ve výsledném testu umožňuje zohlednit časový aspekt.

Kromě toho, že navrhovaná změna přesněji postihuje definovaný koncept, vedou také nutné změny ve výpočtu celkového skóre agenta ke snížení závislosti na referenčním stroji (Hernández-Orallo, 2015). Nejprve se totiž používá uniformní, případně mírně klesající distribuce obtížností. Následně se pro každou obtížnost generuje množina řešení, která mají stejnou váhu. Teprve pak se pro každé řešení generují úlohy podle *univerzální distribuce* (Solomonoff, 1964a,b). Závislost na volbě referenčního stroje tak již není multiplikativní, ale pouze aditivní.

5.2 Teorie úloh

Úlohy či prostředí hrají ústřední roli ve všech představených přístupech k vyhodnocování inteligence. Přesto se jejich formalizace soustředí zejména na způsob interakce s agentem a jejich celkové vlastnosti jako je složitost či obtížnost, což zřejmě pramení ze snahy o co nejobecnější vyhodnocení inteligence agenta a tedy co nejmenší restrikce na použité úlohy. Thórisson a spol. (2015, 2016) se pokusili formulovat *teorii úloh*, která by umožnila hlouběji porozumět úlohám a zejména způsobům, jak je vytvářet a porovnávat. Jedním ze zajímavých přínosů by tak byla schopnost vygenerovat sadu variant nějaké obecněji pojaté třídy úloh stále splňující určitá vlastnosti, což by umožnilo vyhodnotit agenta dostatečně obecně a současně přiměřeně specificky.

Práce na *teorii úloh* je stále v počátcích. Thórisson a spol. (2015, 2016) stanovili požadavky, které musí taková teorie splňovat. Východiskem je zde konečný cíl obecné umělé inteligence, tedy sestrojení agentů použitelných v reálném světě. Zásadním požadavkem na teorii úloh je tak její ukotvení ve fyzice, což umožní modelovat fyzické úlohy včetně souvisejících realistických aspektů jako je spotřeba energie a plynutí času.

Thórisson a spol. (2015, 2016) navrhli počáteční verzi formalizmu, který by měl usnadnit analýzu takto pojatých úloh. Prostředí lze popsat množinou proměnných s navázanými obory hodnot. Mezi proměnnými mohou platit nějaké neměnné vztahy reprezentující různé zákonitosti (np. fyzikální). Vývoj hodnot proměnných v čase od nějakého počátečního stavu pak zajišťují zadané dynamické funkce. Agent může vnímat prostředí skrze senzory, které mu zprostředkují (potenciálně zašuměný) přístup k (některým) proměnným. Jednání v prostředí provádí agent pomocí aktuátorů, které ovlivňují hodnoty (opět potenciálně nepřesně) jiných proměnných. Úloha v takto pojatém prostředí je pak zadána pomocí cílového stavu s množinou žádoucích vlastností a stavu selhání s množinou nežádoucích vlastností.

5.3 Omezené zdroje

Otázka omezených zdrojů, se kterými se musí inteligentní systémy při řešení problémů vypořádat, je některými od problematiky vyhodnocování inteligence oddělována. Například Legg a Hutter (2007b) považují zahrnutí omezených zdrojů do definice inteligence buď za nadbytečné (pokud jsou omezené zdroje neoddělitelnou součástí reality), nebo za nesprávné (pokud je s omezenými zdroji svázána jen naše současná úroveň poznání).

Jiní nicméně tuto otázku považují za důležitou. Goertzel (2010) tak například navrhuje definici *efektivní pragmatické inteligence*, která by ve výsledném skóre zohlednila zdroje užité agentem. Hernández-

Orallo a Dowe (2010) do svého návrhu *kdykoliv přerušitelného testu inteligence* zahrnují koncept fyzického času, což umožňuje některé aspekty omezených zdrojů zohlednit. Ukotvení *teorie úloh* ve fyzice pak se zohledněním časových a energetických limitů při řešení úloh vyloženě počítá (Thórisson a spol., 2015, 2016).

Ještě dále jde Wang (2019), který svou definici inteligence spojuje s *předpokladem nedostatečných znalostí a zdrojů*. Ty tvoří obvyklé pracovní podmínky inteligentního systému, za kterých se má být schopen adaptovat na své prostředí, respektive za nichž začíná řešit nový úkol.

Předpoklad nedostatečných znalostí a zdrojů (Wang, 2019) klade na inteligentní systém tři základní požadavky:

1. Systém musí být konečný.
2. Systém musí být otevřený vůči novým úlohám.
3. Systém musí pracovat v reálném čase.

Dále Wang (2019) svou definici inteligence spojuje s *adaptací*, kterou pojímá následujícím způsobem:

- Nejde o evoluční (druhovou) adaptaci, ale o adaptaci jedince závisející na zkušenosti.
- Oproti klasickému pojetí strojového učení, jde o celoživotní, kumulativní proces s vícero cíli a otevřeným koncem.
- Kromě přizpůsobení systému samotného zahrnuje také možnost přizpůsobit si prostředí, ve kterém systém pracuje.
- Důraz je kladen na záměr nikoliv na výsledek. Vyhodnocení tedy probíhá vůči minulým, nikoliv budoucím, zkušenostem. V případě razantních změn v prostředí, tak nemusí adaptivní chování vést k větší úspěšnosti.

Adaptace je tak s předpokladem omezených znalostí a zdrojů těsně provázaná.

Zajímavým důsledkem tohoto pojetí inteligence tak je, že opakem inteligence není neschopnost vyřešit žádný problém, ale neměnnost schopnosti, což Wang (2019) ztotožňuje s pojmem výpočtu v počítačové vědě. Wang tedy inteligenci pojímá jako fundamentálně odlišnou od výpočtu a jeho přístup tak vlastně již jde za hranici rodiny přístupů ukotvených v algoritmické teorii informace.

5.4 Univerzální psychometrie

Hernández-Orallo (2017) vymezuje *univerzální psychometrii* jako oblast výzkumu zabývající se měřením *obecných rysů chování*, zejména kognitivních schopností a osobnostních charakteristik, *libovolných interaktivních systémů*[3], tedy biologických, umělých či hybridních, individuálních či kolektivních. Hlavní aspirací univerzální psychometrie je navrhnout takové měřící nástroje, které budou fungovat v situaci, kdy máme k dispozici pouze chování dosud neznámého systému. Takové nástroje pak nutně musí být univerzální, tedy nepředpojaté vůči uvedeným typům subjektů a schopné měřit jejich charakteristiky na jedné škále.

Inspirován převládajícím komputacionalistickým paradigmatem kognitivní vědy, které vnímá inteligenci a kognici jako zpracování informací, viz např. (Rescorla, 2017), usiluje Hernández-Orallo (2017) o ukotvení *univerzální psychometrie* do *algoritmické teorie informace*. Univerzální psychometrie tak vlastně zobecňuje a rozšiřuje dosud představené přístupy k vyhodnocování inteligence umělých systémů. Navíc díky integraci a reinterpretaci dosavadních poznatků o vyhodnocování kognitivních schopností odlišných typů systémů přináší novou perspektivu pro zodpovídání starých obtížných otázek.

6 Závěr

Tento příspěvek představil obecné přístupy k vyhodnocování inteligence umělých systémů. Uvedené přístupy vycházejí z *algoritmické teorie informace* (Li a Vitányi, 2008) a usilují o vysokou míru formalizace, ukotvení v pevných teoriích, nízkou míru antropocentričnosti a kulturních a jiných závislostí. Tím se podstatně liší od přístupů jako je *Turingův test* (Turing, 1950), které bývají vymezeny poměrně vágně a hodnocení úspěšnosti v nich závisí na lidském úsudku.

Od průkopnické práce Hernandez-Orallo (2000) lze ve vývoji oblasti spatřovat několik důležitých milníků. Definice *univerzální inteligence* (Legg a Hutter, 2007b) ověřuje schopnost agenta uspět v mnoha různých prostředích. Výsledky agenta pak agreguje pomocí funkce, která klade důraz na jednoduchá prostředí. Jde však o definici ideálního pojmu upozaďující praktičnost, v důsledku čehož je definovaná míra inteligence nevyčíslitelná. Významným problémem je závislost na referenčním Turingově stroji, jehož volba může zásadně změnit, vůči jakým prostředím je agent primárně hodnocen, jak ukázal Hibbard (2009). Tento problém však lze usměrnit, inspirativním je zejména návrh, který podává Hernández-Orallo (2015). Klíčovou prací zabývající se otázkami převodu definice *univerzální inteligence* na prakticky proveditelný test je (Hernández-Orallo a Dowe, 2010), která podává návrh *kdykoliv přerušitelného testu inteligence*. Praktické testování agentů s sebou přináší řadu problémů, zejména pak otázku výběru takových prostředí, která co nejvíce přispějí k vyhodnocení inteligence testovaného agenta. Důležitým aspektem návrhu je také

[3] Hernández-Orallo (2017) označuje tuto množinu subjektů jako „machine kingdom“.

zohlednění fyzického času a celkově adaptivní povaha testovací procedury. Prakticky použitelným testem vycházejícím z algoritmické teorie informace je *test algoritmického IQ* (Legg a Veness, 2013), který aproximuje míru univerzální inteligence agenta a zohledňuje některé z požadavků představených v návrhu kdykoliv přerušitelného testu inteligence (Hernández-Orallo a Dowe, 2010). Test algoritmického IQ nicméně trpí určitými nedostatky, které byly analyzovány a do určité míry odstraněny v (Vadinský, 2018c,d).

Oblast obecného vyhodnocování inteligence umělých systémů stále čelí otevřeným problémům. Příkladem je otázka definice *obtížnosti prostředí*, kterou Hernández-Orallo (2015) nově navrhuje chápat jako složitost nejjednoduššího řešení problému řešeného v prostředí namísto dosud používané složitost prostředí samotného. Dále lze zmínit snahu o vystavění formální *teorie úloh* (Thórisson a spol., 2015, 2016), která by umožnila úlohy lépe pochopit, modelovat a následně generovat ty užitečné pro testování vybraného systému či případu jeho užití. Opakovaně rezonující otázkou je problematika *omezených zdrojů*, kterou například Wang (2019) považuje za klíčovou pro definování inteligence, jelikož podle něj spolu s omezeným znalostmi tvoří výchozí pracovní podmínky všech inteligentních systémů. Oblastí, která by snad mohla přinést odpovědi na všechny možné otevřené otázky je rodící se disciplína *univerzální psychometrie* (Hernández-Orallo, 2017).

Řadu nevyjasněných otázek, z nichž jen některé byly představeny v tomto příspěvku, nelze relativně mladé a malé disciplíně příliš vyčítat. Limitem obecných přístupů k vyhodnocování inteligence umělých systémů, respektive příležitostí pro další práci v této oblasti, je pak také to, že řada zajímavých pokročilých návrhů nebyla dosud implementována do podoby prakticky proveditelných testů. Také míra s jakou jsou umělé systémy testovány obecným přístupy je poměrně nízká.

Reference

Dowe, D. L. a Hájek, A. R. (1998). A non-behavioural, computational extension to the Turing test. Selvaraj, H. a Verma, B. (Eds.), V *Proceedings of ICCIMA'98*, str. 101–106, Singapore. World Scientific.

Goertzel, B. (2010). Toward a formal characterization of real-world general intelligence. Baum, E., Hutter, M. a Kitzelmann, E. (Eds.), V *Proceedings of AGI 2010*, vol. 11 z *Advances in Intelligent Systems Research*, str. 19–24, Amsterdam-Beijing-Paris. Atlantis Press.

Goertzel, B. a Pennachin, C. (Eds.) (2007). *Artificial general intelligence*, vol. 8 z *Cognitive technologies*, Berlin. Springer.

Hernandez-Orallo, J. (2000). Beyond the Turing test. *Journal of Logic, Language and Information*, 9(4):447–466.

Hernández-Orallo, J. (2015). C-tests revisited: Back and forth with complexity. Bieger, J., Goertzel, B. a Potapov, A. (Eds.), V *Proceedings of AGI 2015*, vol. 9205 z *Lecture Notes in Artificial Intelligence*, str. 272–282, Berlin. Springer.

Hernández-Orallo, J. (2017). *The measure of all minds*. Cambridge University Press, Cambridge, 1. vyd.

Hernández-Orallo, J. a Dowe, D. L. (2010). Measuring universal intelligence: Towards an anytime intelligence test. *Artificial Intelligence*, 174(18):1508–1539.

Hibbard, B. (2009). Bias and no free lunch in formal measures of intelligence. *Journal of Artificial General Intelligence*, 1(1):54–61.

Insa-Cabrera, J., Dowe, D. L., España-Cubillo, S., Hernández-Lloreda, M. V. a Hernández-Orallo, J. (2011). Comparing humans and AI agents. Schmidhuber, J., Thórisson, K. R. a Looks, M. (Eds.), V *Proceedings of AGI 2011*, vol. 6830 z *Lecture Notes in Artificial Intelligence*, str. 122–132, Berlin. Springer.

Kolmogorov, A. N. (1963). On tables of random numbers. *Sankhyā: The Indian Journal of Statistics, Series A*, 4(25):369–376.

Legg, S. a Hutter, M. (2007a). A collection of definitions of intelligence. Goertzel, B. a Wang, P. (Eds.), V *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*, vol. 157 z *Frontiers in Artificial Intelligence and Applications*, str. 17–24. IOS Press, Amsterdam.

Legg, S. a Hutter, M. (2007b). Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17(4):391–444.

Legg, S. a Veness, J. (2011). AIQ: Algorithmic intelligence quotient [source codes]. `https://github.com/mathemajician/AIQ`.

Legg, S. a Veness, J. (2013). An approximation of the universal intelligence measure. Dowe, D. L. (Ed.), V *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence*, vol. 7070 z *Lecture Notes in Computer Science*, str. 236–249. Springer, Berlin, Heidelberg.

Levin, L. A. (1973). Universal sequential search problems. *Problems of Information Transmission*, 9(3):265–266.

Li, M. a Vitányi, P. (2008). *An introduction to Kolmogorov complexity and its applications*. Springer-Verlag, New York, 3. vyd.

Minsky, M., Sing, P. a Sloman, A. (2004). The St. Thomas common sense symposium: Designing architectures for human-level intelligence. *AI Magazine*, 25(2):113–124.

Müller, U. (1993). dev/lang/brainfuck-2.lha in aminet. `http://aminet.net/package.php?package=dev/lang/brainfuck-2.lha`.

Rescorla, M. (2017). The computational theory of mind. Zalta, E. N. (Ed.), V *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2017. vyd. `https://plato.stanford.edu/archives/spr2017/entries/computational-mind/`.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, (3):417–457.

Solomonoff, R. J. (1964a). A formal theory of inductive inference, part 1. *Information and Control*, 7(1):1–22.

Solomonoff, R. J. (1964b). A formal theory of inductive inference, part 2. *Information and Control*, 7(2):224–254.

Thórisson, K. R., Bieger, J., Schiffel, S. a Garrett, D. (2015). Towards flexible task environments for comprehensive evaluation of artificial intelligent systems and automatic learners. Bieger, J., Goertzel, B. a Potapov, A. (Eds.), V *Proceedings of AGI 2015*, vol. 9205 z *Lecture Notes in Artificial Intelligence*, str. 187–196, Berlin. Springer.

Thórisson, K. R., Bieger, J., Thorarensen, T., Sigurdardóttir, J. S. a Steunebrink, B. R. (2016). Why artificial intelligence needs a task theory and what it might look like. Steunebrink, B., Wang, P. a Goertzel, B. (Eds.), V *Proceedings of AGI 2016*, vol. 9782 z *Lecture Notes in Artificial Intelligence*, str. 118–128, New York. Springer.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236):433–460.

Vadinský, O. (2018a). AIQ: Algorithmic intelligence quotient [source codes]. `https://github.com/xvado00/AIQ/archive/v1.3.zip`.

Vadinský, O. (2018b). Přehled přístupů k vyhodnocování inteligence umělých systémů. *Acta Informatica Pragensia*, 7(1):74–103.

Vadinský, O. (2018c). Towards general evaluation of intelligent systems: Lessons learned from reproducing AIQ test results. *Journal of Artificial General Intelligence*, 9(1):1–54.

Vadinský, O. (2018d). Towards general evaluation of intelligent systems: Using semantic analysis to improve environments in the AIQ test. Iklé, M., Franz, A., Rzepka, R. a Goertzel, B. (Eds.), V *Proceedings of AGI 2018*, vol. 10999 z *Lecture Notes in Artificial Intelligence*, str. 248–258, Cham. Springer.

Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence*, 2(10):1–37.

Vlci a smečkový algoritmus ve světě membránových agentů

Daniel Valenta, Lucie Ciencialová, Luděk Cienciala

Slezská univerzita v Opavě, Filozoficko-přírodovědecká fakulta v Opavě,
Ústav informatiky
a
Research Institute of the IT4Innovations Centre of Excellence
Bezručovo náměstí 1150/13, 746 01 Opava
Email: daniel.valenta@fpf.slu.cz, lucie.ciencialova@fpf.slu.cz, ludek.cienciala@fpf.slu.cz

Abstrakt

Membránové systémy, zavedené v roce 1998 Gheorghem Păunem, představují jeden z biologicky motivovaných výpočetních modelů teoretické informatiky Păun (2000). Od jejich uvedení vznikla celá řada různých variant membránových systémů, v závislosti na struktuře a používaných pravidlech. Model P kolonie, jeden z modelů patřících mezi membránové systémy, se skládá z jedno-membránových agentů sdílejících společné prostředí. Agenti mají omezený počet objektů a množinu programů určujících způsob vývoje objektů nebo výměnu objektů s prostředím. Přestože se jedná o jednoduchý model tvořený jedno-membránovými agenty, reprezentací prostředí a chováním agentů připomíná tradiční multiagentní systémy, které využívají algoritmy aplikované v praxi – například pro řešení optimalizačních úloh. Jedním z takových multiagentních systémů je tzv. smečkový algoritmus inspirovaný chováním vlků v přírodě, jejich hierarchií a způsobem lovu. V našem příspěvku představíme jeden z nových úhlů pohledu na chování P kolonií a algoritmů inspirovaných chováním vlků nazývaných smečkové algoritmy. Poukážeme na okolnosti související právě se simulací a porovnáním těchto dvou poměrně odlišných biologicky inspirovaných systémů, a to diskrétně založeného systému a systému využívajícího spojitých funkcí.

1 Úvod

V teoretické informatice se setkáváme s celou řadou výpočetních modelů. V našem článku budeme studovat vlastnosti P kolonií, které představují jeden z biologicky motivovaných výpočetních modelů, spojující dvě oblasti teoretické informatiky, kolonie (více v Kelemen a Kelemenová (1992)) a membránové systémy (např. Păun a spol. (2010)). P kolonie byly zavedeny v roce 2004 v práci Kelemen a spol. (2004).

Do prostředí P kolonie umístíme agenty představující organismy „žijící“ v P koloniích. Každý agent je tvořen pouze jednou membránou ohraničující oblast s objekty. Jako objekt si můžeme představit látky, které organismy mění, přijímají, či vylučují do prostředí. V každém okamžiku jsou všechny objekty uvnitř agenta změněny nebo přesunuty. Počet objektů v každém agentu je pevně dán a je neměnný v průběhu celého výpočtu, je také stejný pro každého agenta. Počet objektů je parametrem P kolonie, který nazýváme kapacitou.

V prostředí se na začátku výpočtu nachází pouze kopie speciálního objektu nazývaného environmentální. Tohoto objektu je dostatečné množství pro výpočet systému. Nemůže se stát, že by ho bylo nedostatek. V přírodě si pod environmentálním objektem můžeme představit vzduch nebo vodu v závislosti na tom, o jakých organismech budeme uvažovat.

Každý agent má svou množinu programů, které jsou tvořeny pravidly. Počet pravidel v programu je shodný s počtem objektů uvnitř agenta. Agent je určen svým stavem (obsaženými objekty) a svou množinou programů. Programy určují činnost agenta, v přeneseném slova smyslu jeho životní projevy.

Přestože agenty můžeme chápat jako nezávislé organismy, mohou svou činnost navzájem ovlivňovat, a to prostřednictvím prostředí, které představuje komunikační kanál. Agent do něj může umístit objekty odpovídající jeho stavu v daném okamžiku a tím ovlivnit činnost ostatních agentů.

V tomto příspěvku se budeme věnovat také dalšímu přírodou motivovanému modelu nazývaného smečkové algoritmy. Daný model je inspirovaný chováním vlků, jejich vzájemnou hierarchií a způsobem zajišťování potravy. Uvedené algoritmy byly publikovány v práci Mirjalili a spol. (2014). Vstupem algoritmu je funkce reprezentující prostředí, v němž jsou umístění náhodně agenti. Změna ohodnocení a pohyb agentů probíhá pomocí fitness funkce. Mezi agenty je hierarchické uspořádání podobně jako u vlků a to Alpha, Beta, Delta a Omega. Pohyb agentů je závislý na pozici třech nejvýše hierarchicky postavených jedinců.

P kolonie i smečkové algoritmy představují multiagentní systémy.

Efektivitu určitého algoritmu z pohledu časové nebo paměťové složitosti posuzujeme ve vztahu ke konkrétnímu výpočetnímu modelu. Použití výpočetních

modelů nám umožňuje objektivně sledovat výkon, efektivitu algoritmů, a to nezávisle na specifických vlastnostech konkrétní implementace. Algoritmus může být velmi efektivní pro určitý výpočetní model, zatímco na jiném modelu může fungovat jen omezeně, pomaleji, nebo nemusí být realizovatelný vůbec.

Ačkoli v praxi nás zajímá zejména uplatnění modelu na soudobém hardware - elektronickém počítači, z hlediska vědy a budoucnosti jsou zajímavé také abstraktní výpočetní modely, pro které nemáme k dispozici skutečný stroj, který by jejich činnost realizoval. V oblasti teoretické informatiky je abstraktních výpočetních modelů celá řada. Takový model může být například daleko „paralelnější“, než může být v současné době ten nejlepší skutečný počítač. Takové modely jsou obvykle výpočetně nebo paměťově velmi náročné a v současné době nedokážeme využít jejich plný praktický potenciál. V mnoha případech je ale možná takzvaná počítačová simulace takového modelu, která umožní sledovat a zkoumat chování daného modelu.

Musíme si uvědomit, že se jedná primárně o výpočetní modely inspirované přírodou, nikoli biologické modely simulující chování určitých organismů. Vzhledem k probíhajícímu výzkumu buněk, buněčných struktur, DNA výpočtů, kvantových výpočtů a nanotechnologií je zde jistá pravděpodobnost jejich implementace na vysoce paralelním neelektrickém (biologickém) substrátu. V syntetické biologii se už v současné době můžeme setkat se syntetizovaným genomem a programovým chováním. V případě naplnění představy o vytvoření masivních nedeterministických „in vivo“ (vnitro buněčných) modelů paralelních procesorů inspirovaných například membránovými výpočty, by to znamenalo významné zvýšení rychlosti zpracovávání informací paralelním způsobem a možnost využívat nedeterministických postupů.

V našem příspěvku se zaměříme na porovnání dvou zcela odlišných biologicky inspirovaných modelů, membránových systémů a smečkových algoritmů. Zaměříme se na možnosti a omezení membránových systémů pro simulaci tradičních přístupů v oblasti přírodou inspirovaných optimalizačních algoritmů, které jsou již využívány při řešení složitých úloh, jako jsou právě smečkové algoritmy.

2 P kolonie a 2D P kolonie

Jak jsme již uvedli, P kolonie jsou jednou z variant výpočetních modelů ze skupiny membránových systémů a inspirují se strukturou a činností živých organismů, kteří spolu žijí ve společném prostředí, jako jsou například mravenci nebo včely. U mravenců mezi nejdůležitější komunikační prostředky patří mravenčí pachy, tzv. feromony. I přes malý mozeček, který mají, jsou schopni vytvořit poměrně složité společenské struktury. Jako druhý příklad jsme uvedli včely. Žihadllový aparát včely tvoří několik žláz. V jejich výměšcích je histamin, fosfolipáza a hyaluronidáza vyvolávající nepříjemné pocity svědění a pálení v okolí vpichu. Tyto žlázy vylučují i feromony, které slouží k upozornění na nebezpečí. Do rány vpichu vstřikují včely i látku, pomocí které se navigují ostatní včely a vpichují další žihadla v blízkosti prvního bodnutí. Kromě jedu tedy včely vpichují i dávku izoamylacetátu. Včely z jednoho úlu se poznají dle pachu, který se skládá z mnoha složek. Pyl a nektar dvou úlů nevoní stejně. Povrch těla včely je uzpůsoben tak, že nabírá vůni úlu a udržuje si ji. Včely značí cestu k nektaru, to je i nápomocno k nalezení cesty zpátky k úlu. Chemické látky tedy slouží k dorozumívání včel a umožňují fungování včelstva. Komunikují prostřednictvím prostředí, ve kterém žijí.

Model P kolonie se skládá z jednoduchých agentů, umístěných do společného prostředí a představují multiagentní systém. Prostředí slouží jako komunikační kanál a také jako skladiště objektů.

Každý agent je tvořen jednou membránou, která ohraničuje oblast s objekty. Počet objektů je shodný pro všechny agenty. Činnost agentů je dána programy. Program je tvořen pravidly a obsahuje tolik pravidel, kolik je objektů uvnitř agenta. Existují tři základní typy pravidel a to přepisující, komunikační a řídící pravidla.

Nyní si uvedeme tvary uvedených pravidel a pokusíme se je alespoň z části interpretovat v přírodě.

Přepisující pravidla mají tvar $a \rightarrow b$. Aplikací pravidla je objekt a přepsán na objekt b. Dané pravidlo umožňuje změnu látky a na látku b uvnitř organismu.

Komunikační pravidla mají tvar $c \leftrightarrow d$. Pomocí tohoto pravidla je objekt c, který je uvnitř agenta, přesunut ven, a objekt d, který je vně, je přesunut dovnitř agenta. Komunikační pravidla umožňují komunikaci organismů – agentů. Jedna látka c je vyloučena do prostředí a další látka d je přijata organismem. Objekt d můžeme klidně i chápat jako zpětnou vazbu při vyloučení objektu c.

Pomocí **řídících pravidel** dáváme agentům možnost výběru mezi dvěma možnostmi. Řídící pravidla mají tvar $\langle a \rightarrow b, c \leftrightarrow d / c' \leftrightarrow d' \rangle$. Pravidlo skládá ze dvou částí, komunikačních pravidel. Pokud je řídící pravidlo aplikováno, má část $c \leftrightarrow d$ vyšší prioritu k provedení než část $c' \leftrightarrow d'$. To znamená, že agent vybere komunikační pravidlo $c \leftrightarrow d$ (pokouší se najít uvnitř objekt c a objekt d v prostředí). Pokud toto pravidlo může být vykonáno, tak je použito. Když první pravidlo nemůže být provedeno, agent použije druhé pravidlo z dané dvojice pravidel – $c' \leftrightarrow d'$. U řídících pravidel organismus zjišťuje, jestli se v prostředí vyskytuje určitý typ látky, a pokud ne, přizpůsobí své chování tomuto nedostatku. Například pokud mravenec nenajde danou látku, ztratí feromonovou stopu, bude ji dále hledat, a pokud ji najde, bude ji sledovat.

Výpočet P kolonie začíná v počáteční konfiguraci, která je definována následujícím způsobem: prostředí a všichni agenti obsahují pouze kopie objektů e, kde e je environmentální objekt. Aplikováním programů mohou agenti měnit svůj obsah a pomocí prostředí mohou ovlivňovat chování ostatních agentů v dalších krocích výpočtu. V přírodě se také často prostředí využívá jako komunikační kanál. Některé organismy vylučují jisté látky do prostředí a tím ovlivňují chování ostatních. V každém kroku výpočtu každý agent nedeterministicky vybere jeden ze svých aplikovatelných programů a vykoná jej. Výpočet končí zastavením, kdy žádný agent nemůže aplikovat žádný ze svých programů. Výsledkem výpočtu je počet určitých objektů v prostředí na konci výpočtu. Z důvodu nedeterminismu v průběhu výpočtu můžeme získat několik výpočtů, které končí zastavením.

P kolonie jsou výpočetně úplné. Zajímavá je otázka, jaký je nejmenší počet agentů a počet programů uvnitř agenta při zachování výpočetní úplnosti. Cílem tohoto článku není výpočetní síla uvedených systémů, proto je nebudeme dále rozvádět. Pro získání informací o výpočetní síle a vlastnostech P kolonií doporučujeme čtenáři článek Ciencialová a spol. (2019). Více informací o membránových systémech čtenář nalezne v knize Păun a spol. (2010).

Pokud rozšíříme původní model P kolonií o dvourozměrné prostředí tak, aby agenti neměli přímý přístup ke všem objektům a museli se po dosažení cíle pohybovat, budeme daný model nazývat 2D P kolonie. Dané rozšíření umožní právě použití pro modelování reálných situací a tím i predikci budoucího stavu reálného prostředí. Například v práci Cienciala a spol. (2014) se autoři věnovali simulaci bleskových povodní pomocí 2D P kolonií. V reálném světě (stejně tak i v kybernetickém světě) se liší koncentrace látek v závislosti na místě a živé organismy nemají možnost vědět, co je „za horizontem“. Tyto úvahy nás inspirovaly k zavedení nového modelu P kolonií, kde jsou agenti umístěni ve 2D kartézské mřížce. Agenty tedy umístíme do této mřížky a „pohled“ agenta omezíme na buňky bezprostředně jej obklopující. Na základě obsahu těchto buněk je ovlivněna poloha agenta v mřížce v dalším kroku výpočtu. Obsah agenta je také limitován dvěma objekty, kolonie má tedy kapacitu 2.

A jak se změní pravidla? Budeme vycházet z původního modelu P kolonií, tak jak jsme je popsali. Program agenta může být tedy sestaven ze dvou typů pravidel přepisujících a komunikačních s tím rozdílem, že komunikační pravidlo bude mít navíc vliv pouze na místo, kde se právě agent nachází.

Přidáme nový typ pravidel pro pohyb agenta ve 2D prostředí. Podmínkou pro pohyb agenta je nalezení konkrétních objektů v konkrétních místech prostředí. To je určeno pomocí matice základních objektů. Agent hledá nejvýše jeden objekt v každé buňce okolí. Pokud takové objekty nalezne, přemístí se o jednu buňku nahoru, dolů, doleva nebo doprava. Řídící pravidlo je nahrazeno pravidlem pohybovým. Pro jednoduchost chování agenta stanovíme ještě jednu podmínku: pokud agent mění polohu, pak nemůže komunikovat s prostředím. To znamená, že pokud program obsahuje pravidlo pro pohyb, tak druhým pravidlem musí být pravidlo přepisovací.

Nyní uvedeme formální definici 2D P kolonie.

Definice: 2D P kolonie je struktura:
$\Pi = (A, e, Env, B_1, \ldots, B_k, f), k \geq 1$, kde

- A je abeceda kolonie, její prvky nazýváme objekty,
- $e \in A$ je základní objekt prostředí 2D P kolonie, který nazýváme environmentální
- Env je dvojice $(m \times n, w_E)$, kde $m \times n, m, n \in N$ je matice o velikost $m \times n$ multimnožin objektů nad množinou $A - \{e\}$, $m, n \in N$ je velikost prostředí a w_E je počáteční obsah prostředí.
- $B_i,\ 1 \leq i \leq k$, jsou agenti, každý agent je struktura $B_i = (O_i, P_i, [r_i, s_i])\ ,\ 0 \leq o \leq m,\ 0 \leq p \leq n$, kde:
  - O_i je multimnožina nad množinou A, určuje počáteční stav (obsah) agenta, $|O_i| = 2$,
  - $P_i = \{p_{i,1}, \ldots, p_{i,l_i}\}, l \geq 1, 1 \leq i \leq k$ je konečná množina programů, kde každý program obsahuje právě 2 pravidla v jedné z následujících forem:
    - $a \rightarrow b$, zvané vývojové pravidlo, $a, b \in A$,
    - $c \leftrightarrow d$, zvané komunikační pravidlo, $c, d \in A$,
    - $[a_{q,r}] \rightarrow s, 0 \leq q, r \leq 2, s \in \{\Leftarrow, \Rightarrow, \Uparrow, \Downarrow\}$, zvané pohybové pravidlo,
  - $[r_i, s_i]$ je počáteční pozice agenta B_i ve 2D prostředí,
    $0 \leq r_i \leq m-1,\ 0 \leq s_i \leq n-1,\ 1 \leq i \leq k$,
- $f \in A$ je konečný objekt kolonie.

Konfigurace 2D P kolonie je dána stavem prostředí, tedy maticí typu $m \times n$ tvořenou multimnožinou objektů z $A - \{e\}$, a stavem všech agentů, tedy páry jejich objektů z abecedy A a jejich souřadnicemi v prostředí. Počáteční konfigurace je dána definicí 2D P kolonie.

Výpočetní krok se skládá ze tří částí. První část spočívá ve stanovení aplikovatelné množiny programů v závislosti na aktuální konfiguraci P kolonie. Existují programy patřící do množin programů všech agentů. Ve druhé části je u každého agenta vybrán jeden program patřící do množiny aplikovatelných programů. Neexistuje kolize mezi komunikačními pravidly různých programů. Ve třetí části jsou vykonány vybrané programy.

Vykonáním programů se změní konfigurace kolonie. Změna konfigurace je založena na změně stavu prostředí, obsahu a umístění agentů.

Výpočet probíhá nedeterministicky a maximálně paralelně a končí zastavením, kdy žádný agent nemá aplikovatelný program.

Výsledek výpočtu je počet kopií finálního objektu umístěného v prostředí na konci výpočtu.
Cílem výzkumu 2D P kolonií je sledovat jejich chování (změny konfigurací, stavů agentů), nikoli výpočetní sílu. Můžeme definovat několik pohledů, jak posoudit dynamiku výpočtu:

- počet pohybů agentů,
- počet „navštívených“ buněk (nebo „nenavštívených“ buněk),
- počet kopií určitého objektu v „domovské“ buňce nebo v celém prostředí.

Navíc můžeme uvažovat, zda daný počet sledujeme po jednotlivých krocích výpočtu nebo až na konci výpočtu.

3 Vlci a smečkový algoritmus

Nyní se na chvíli oprostíme od membránových systémů a zaměříme jednu z metod optimalizačních algoritmů, a to na takzvaný smečkový algoritmus.

Smečkový algoritmus, anglicky zvaný Grey wolf optimization algorithm, je zavedenou metodou pro řešení problémů v oblasti globální optimalizace. Vstupem je multidimenzionální matematická funkce a cílem je nalezení globálního extrému, tedy nejvyšší nebo nejnižší bod této matematické funkce.

U problému nalezení globálního optima se ještě pozastavíme. Ačkoli se tento problém může zdát na první pohled triviální, v případě některých funkcí tomu tak není. Existují funkce velmi rozměrné, které mají mnoho extrémů nebo nejsou diferencovatelné. Taková matematická funkce může představovat velmi složitý problém ze skutečného světa, který potřebujeme vyřešit v určitém časovém úseku, přičemž obecně známé deterministické přístupy a algoritmy selhávají – nemáme dostatek času projít všechna možná řešení.

Potřeba řešit složité problémy v uspokojivém čase vedla k výzkumu nových metod založených na optimalizaci. Řada z nich je odvozena heuristicky a motivuje se přírodou. Výstupem takových algoritmů sice nemusí být naprosto přesné řešení, ale zato rychlé a přitom dostatečně přesné. Smečkový algoritmus je jednou z úspěšných a praxí již ověřených metod optimalizace, který se inspiruje společenským soužitím vlků obecných, zejména jejich způsobem lovu a utvářením hierarchie ve smečce. Byl zaveden v roce 2014 a jeho autory jsou Seyedali Mirjalili, Seyed Mohammad Mirjalili a Andrew Lewis (Mirjalili a spol., 2014).

Na proces lovu vlčí smečkou můžeme analogicky nahlížet jako na proces řešení optimalizačního problému. Chování vlků při lovu a vytváření hierarchie dokážeme matematicky popsat a modelovat v multiagentním systému. Vlci (agenti) se v prostředí (matematická funkce) pohybují za účelem nalézt co nejvydatnější kořist (globální extrém funkce).

V hierarchii vlčí smečky rozlišujeme vlky Alpha, Beta, Delta a Omega. Vlci Alpha jsou v hierarchii postavení nejvýše a obvykle tvoří dominantní pár (samec a samice). Jsou vůdci smečky a ostatní vlci ve smečce je plně respektují. Vlci Beta podporují Alpha pár při rozhodovacích činnostech a poskytují jim zpětnou vazbu. Delta vlci se dále dělí na skauty, strážce a ošetřovatelé, plní rozkazy výše postavených vlků a starají se o běžný chod smečky. Vlci omega jsou takzvaní „obětní beránci“, mají například právo jíst až jako poslední a ostatní vlci si na nich mohou „vylít zlost“. Mají však v hierarchii důležitou roli, a to filtrovat agresi a udržovat tím smečku pohromadě.

Také smečkový algoritmus rozlišuje agenty (vlky) typu Alpha, Beta, Delta a Omega. Pozice v prostředí, kde se agent v danou chvíli nachází, představuje jedno z možných řešení daného problému. Tři nejlepší agenti podle fitness funkce, která kvantitativně vyjadřuje kvalitu řešení každého z nich vzhledem k hledanému extrému funkce, označujeme jako agenty Alpha, Beta a Delta. Předpokládáme, že globální extrém (potrava) se nachází někde mezi nimi, což agenti dále zohledňují při svých pohybech. Agenti Alpha, Beta a Delta jsou si v případě smečkového algoritmu rovnocenní a jejich pozice mají stejný význam.

Lov probíhá v několika fázích. Nejprve vlci pátrají po co nejvydatnější potravě s ohledem na úsilí, které musí k jejímu ulovení vynaložit. Následně kořist pronásledují, snaží se ji zahnat do kouta, nebo oddělit jedince od stáda. V další fázi kořist obklučují z různých stran, aby neměla kam utéct. Jakmile kořist obklíčí, útočí na ni z různých směrů a zaměřují se na její slabá místa. Poté, co se kořist vysílí a přestane se bránit, vlci ji zadáví.

Analogicky k tomuto principu algoritmus plynule přepíná mezi fázemi průzkumu a lovu a používá k tomu dva vektory $\vec{A}$ a $\vec{C}$, jejichž prvky mají náhodné hodnoty v rozmezí od -2 do 2. Počet prvků obou vektorů je roven dimenzi prostředí a každý prvek ovlivňuje směr pohybu vlka v konkrétním rozměru prostředí. Prvek vektoru s hodnotou v rozsahu od -1 do 1 přinutí vlky spíše lovit, zatímco jinak se snaží intenzivně procházet okolí a vyhnout se tak nalezení pouze lokálního optima.

Prvky vektoru $\vec{A}$ oproti vektoru $\vec{C}$ navíc závisí na aktuální iteraci algoritmu – s přibývajícími iteracemi algoritmu se zvyšuje pravděpodobnost, že se hodnoty prvků vektoru přibližují k hodnotě 0, což přinutí vlky (agenty) více lovit a méně prozkoumávat okolí.

Prvky vektoru $\vec{C}$ jsou naopak čistě náhodné. Tento vektor simuluje překážky v prostředí a umožňuje v malé míře prozkoumávat okolí i v posledních iteracích algoritmu, kdy většina agentů volí spíše fázi lovu.

Vstupem algoritmu je matematická funkce, která reprezentuje daný problém, který chceme vyřešit. Tato funkce představuje prostředí, do něhož jsou náhodně

umístění agenti. Jedná se tedy o multiagentní systém. Chování agentů v tomto prostředí určitým způsobem simuluje chování skutečných vlků. Algoritmus pracuje v iteracích, přičemž v každé z nich jsou provedeny následující kroky:

1. Ohodnocení agentů pomocí fitness funkce,
2. Určení sociální hierarchie,
3. Pohyb agentů v prostředí na základě pozic tří nejlepších vlků (agentů) v prostředí podle fitness funkce (v závislosti na hodnotách prvků vektorů $\vec{A}$ a $\vec{C}$ pro konkrétní dimenzi prostředí: ve fázi lovu směrem k potravě, ve fázi průzkumu naopak dále od potravy),
4. Kontrola ukončovačího kritéria.

Ve fázi lovu mají agenti tendenci se k potravě přibližovat z různých směrů a dochází tak k jevu obkličování, podobně jako v případě skutečných vlků.

4 Porovnání obou modelů

Oba modely se inspirují v přírodě, a to společenským soužitím vybraných živočichů, avšak v mnohém se značně liší. Zatímco 2D P kolonie pracují s diskrétním prostředím, smečkový algoritmus se spojitou matematickou funkcí. Agenti 2D P kolonie spolu komunikují výhradně pomocí prostředí, zatímco u smečkového algoritmu agenti při svých pohybech využívají informace o pozicích dalších agentů. Takovou informaci agenti membránových systémů nemají k dispozici. Smečkový algoritmus dále silně pracuje s náhodností a pravděpodobností, což 2D P kolonie umožňují jen v omezené míře, a to nedeterministickým výběrem z množiny pravidel pro jednu a tu samou konfiguraci.

Abychom eliminovali rozdíly obou modelů, zavedli jsme nový model 2D P kolonie, jehož prostředí dokáže uchovávat číselné hodnoty uvnitř objektů matice. Model jsme dále rozšířili o takzvanou tabuli, která agentům umožňuje sdílet například informace o nalezených číselných hodnotách v prostředí.

Definice: 2D P kolonie s tabulí, přizpůsobená pro simulaci smečkového algoritmu, je struktura: $\Pi_{gw} = (A, e, Env, B_1, \ldots, B_k, f), k \geq 1$, kde

- $V = \left\{\boxed{\ }, \boxed{\ }', e, b, m, n, f\right\}$,
- $e \in V$ je základní objekt prostředí,
- Env je trojice $(i \times j, w_e, f\,(x, y))$, kde $i, j \in \mathbb{N}$, $w_E = |a_{r,s}|, a_{r,s} = \varepsilon, 1 \geq r \geq i, 1 \geq s \geq j$,
- $A_1, A_2, \ldots, A_k$ jsou agenti, $A_i = (O_i, P_i, [r_x, r_y])$, kde:
  - $|o_i| = 2$,
  - $P_1 = P_2 = \cdots = P_k$, P_i jsou pravidla definovaná níže,
  - $[r, s]$ jsou počáteční souřadnice,
- BB je tabule, kterou popíšeme v dalším textu,
- f je konečný objekt, $f \in V$.

Všimněme si rozdílů oproti definici 2D P kolonie uvedené v druhé kapitole. Abeceda A zahrnuje speciální box-objekt $\boxed{\ }$, který umožňuje uvnitř sebe uchovávat číselné hodnoty. Prostředí env obsahuje navíc funkci $f(x, y)$, která každému bodu v matici prostředí přiřazuje konkrétní číselnou hodnotu podobně, jako je tomu v případě fitness funkce u smečkového algoritmu. V neposlední řadě zde máme navíc tabuli BB (od anglického slova *blackboard*), která je dostupná kdykoli pro čtení a zápis všem agentům a kterou podrobně popíšeme později.

Další vlastnosti takto upravené 2D P kolonie zůstávají neměnné. Konfigurace je dána stavem prostředí a stavem všech agentů. Výpočetní krok spočívá opět v aplikaci pravidel programů určených pro danou konfiguraci. Výpočet je nedeterministický a maximálně paralelní. Výpočet ukončíme ve chvíli, kdy se již agenti nepohybují a zároveň nedochází k dalšímu (průběžnému) zlepšování hodnot v tabuli – agenti Alpha, Beta a Delta mají vždy k dispozici program, který mohou aplikovat, výpočet tedy může probíhat libovolně dlouho. Výsledkem výpočtu jsou pozice agentů v prostředí a stav tabule na konci výpočtu.

Pravidla pro takto upravenou 2D P kolonii jsme navrhli s cílem simulovat činnost smečkového algoritmu. Počáteční konfigurace agenta je ee a jeho pozice v prostředí je $[r, s]$. V následujících odrážkách uvedeme příklad některých pravidel, úplný přehled pravidel je pak dostupný v publikaci Valenta a spol. (2021).

1. $\left\langle e \rightarrowtail \boxed{x}'; e \rightarrow \mathit{Get}(BB[\mathit{alpha}, \boxed{\ }], \right)\rangle$ $x \in \mathbb{R}$ je číslo umístěné v prostředí na pozici $[r, s]$, *alpha* je odkaz na hodnotu v tabuli vyhrazenou pro zápis alfa agenta do tabule; Tento program umožňuje číst hodnotu z prostředí a hodnotu vlka alfa z tabule.
2. $\left\langle x > y : \boxed{x} \rightarrow \boxed{x}, \boxed{y}' \rightarrow A \right\rangle$ – Tento program agent použije pro porovnání své pozice s hodnotami v tabuli a zařadí se do hierarchie. Symbol A v tomto případě reprezentuje vlka Alpha, kterým se agent stane v případě, je-li jeho hodnota vyšší než ta v tabuli, ale obdobné pravidla máme také pro agenty Beta, Delta a Omega.
3. Je-li jeden z vnitřních objektů agenta ekvivalentní s A, tedy agent se zařadil do hierarchie na pozici Alpha, použije se program:
   $\langle \mathit{Update}(\boxed{x}, BB[\mathit{alpha}]), A \rightarrow a' \rangle$,
   Funkce $Update$ je funkcí tabule, která slouží pro

aktualizaci konkrétní hodnoty v tabuli. První argument funkce je objekt agenta k zapsání do tabule, druhý argument pak určuje pozici v tabuli, kam se má daná hodnota zapsat. Obdobné pravidla používají také agenti Beta a Delta, kteří svou hodnotu zapisují do tabule na pozice *BB[beta]* a *BB[delta]*.

4. Agenti Omega (nejníže postavení v hierarchii) do tabule nezapisují. Namísto toho se pohybují v prostředí směrem o jednu pozici nahoru, dolů, vlevo, nebo vpravo. Agent postupně zkouší jednotlivé směry v náhodném pořadí. Agent zvolí daný směr v případě, pokud se daným pohybem nevzdálí od přibližné pozice potravy, která je vypočtena tabulí na základě pozic agentů Alpha, Beta a Delta a nachází se někde mezi nimi. V opačném případě zvolí jiný směr, který dosud nevyzkoušel.

Podrobný popis všech pravidel systému je k dispozici v publikaci Valenta a spol. (2021).

Tabule je struktura

$$BB_{GWO} = ((fnc, rcv), [\overrightarrow{v1}, \overrightarrow{v2}]),$$

kde dimenze obou vektorů $\overrightarrow{v1}$, $\overrightarrow{v2}$ je $j = \max(7, k)$, $k \geq 1$ je počet agentů.

Vektor $\overrightarrow{v1}$ je vektor s prvky pojmenovanými následovně: *AlphaValue, BetaValue, DeltaValue, AlphaPosition, BetaPosition, DeltaPosition, preyPosition*. Je-li $j > 7$, prvky s indexem větším než 7 jsou již bez jména. Počáteční obsah každé z těchto hodnot je 0.

Vektor $\overrightarrow{v2}$ je vektor s prvky pojmenovanými následovně: $A'_0 sDistanceFromPrey$, $A'_1 sDistanceFromPrey$, ..., $A'_k sDistanceFromPrey$, k je počet agentů.
Tyto prvky vektoru slouží k uchovávání hodnot vyjadřujících vzdálenost jednotlivých agentů od tabule, a jsou vypočteny tabulí po každém použití některé z funkcí tabule provádějících čtení nebo zápis. Počáteční obsah každé z těchto hodnot je 0.

Tabule řeší rozdílný způsob komunikace agentů u obou modelů. Umožňuje agentům kdykoli sdílet své fitness hodnoty s dalšími agenty, což je podstatná informace, kterou potřebují k zařazení se do hierarchie.

Pro komunikaci agentů s tabulí má tabule vestavěné funkce Get pro čtení a $Update$ pro zápis. Při každém použití některé z funkcí agentem je vypočtena vzdálenost agenta od prostředí pomocí takzvaných přijímačů, které fungují na podobném principu, jako navigační systém GPS. Dva přijímače neustále rotují kolem prostředí a naslouchají signálům agentů. Výsledná pozice agenta je vypočtena jako průnik dvou kružnic se středem v přijímačích a poloměrem odpovídajícím době putování signálu od agenta. Dobu putování signálu zaokrouhlujeme, čímž vzniká odchylka, která plní funkci určité náhodnosti, podobně jako je tomu v případě náhodných vektorů u smečkového algoritmu.

5 Využití 2D P kolonie s tabulí pro řešení optimalizačních úloh

Implementaci modelu popsaného v předchozí kapitole jsme popsali v publikaci Valenta a spol. (2021). Implementace umožnila model testovat v porovnání se smečkovým algoritmem a sledovat jeho chování. Testování ukázalo, že lze pomocí modelu 2D P kolonie s tabulí úspěšně simulovat smečkový algoritmus. Má to ale určitá omezení.

Velikost kroku agenta je v případě 2D P kolonie omezena vždy na jednu pozici směrem vlevo, vpravo, nahoru, nebo dolů, zatímco smečkové algoritmy umožňují pohyb agentů téměř bez omezení, a to v libovolném směru až do dvojnásobku vzdálenosti od přibližné pozice od potravy. V případě 2D P kolonie s tabulí se navíc pohybují jen agenti Omega, zatímco s případě smečkového algoritmu se v omezené míře pohybují i Alpha, Beta a Delta agenti.

Velikost kroku agenta se negativně projevuje také v případě rozměrných prostředí, kdy agentům 2D P kolonie s tabulí trvá daleko déle, než se dostanou z jednoho místa na druhé.

Model 2D P kolonie je také méně odolný vůči konvergenci k lokálnímu optimu. Pokud agent nalezne lokální extrém a všechny jeho okolní pozice mají horší (nižší) fitness hodnoty, nemá již důvod měnit svou pozici - agenti totiž při svých pohybech zohledňují jen pozice v jejich bezprostředním okolí. Toto omezení dále souvisí s funkcí, která reprezentuje prostředí (daný problém), a kterou je nutné nejprve vhodným způsobem diskretizovat vzorkováním. Popsaný problém se projevuje, jsou-li ve výsledné matici větší plochy se stejnými hodnotami (nedochází k průběžnému zlepšení nalezených hodnot).

Přes tato úskalí se nám podařilo ukázat, že i takto jednoduchý model, jako 2D P kolonie, dokáže po menších úpravách řešit optimalizační problémy na vhodně zvolených a diskretizovaných funkcích.

Předpokládáme, že uvedený model 2D P kolonie můžeme po menších úpravách využít také pro simulaci dalších optimalizačních algoritmů, jako jsou mravenčí kolonie nebo algoritmus hejna částic. Uplatnění si dokážeme představit také ve skutečném multiagentním systému, kde může daný model sloužit například pro řízení činnosti robotů prohledávajících prostředí.

6 Závěr

V tomto článku jsme si popsali dva zcela odlišné modely teoretické informatiky, a to membránové systémy, z nichž jsme si zvolili 2D P kolonie, a optimalizační algoritmy, z nichž jsme si zvolili smečkový algoritmus. Oba modely jsme porovnali a poukázali na odlišnosti, které znemožňují simulaci smečkového algoritmu pomocí 2D P kolonií. Také jsme představili model 2D

P kolonie s tabulí a možností uchovávat číselné hodnoty v takzvaných box-objektech, který umožní simulovat nejen smečkový algoritmus, ale také další optimalizační algoritmy pracující na podobných principech. To může vést k dalšímu výzkumu v oblasti využití membránových systémů pro řešení optimalizačních úloh.

Poděkování

Tento příspěvek vznikl za podpory Slezské univerzity v Opavě v rámci grantu Student Funding Scheme, projekt SGS/8/2022.

Reference

Cienciala, L., Ciencialová, L. a Langer, M. (2014). Modelling of surface runoff using 2D P colonies. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8340 LNCS:101–116.

Ciencialová, L., Csuhaj-Varjú, E., Cienciala, L. a Sosík, P. (2019). P colonies. *Journal of Membrane Computing*, 1(3):178–197.

Kelemen, J. a Kelemenová, A. (1992). A grammar-theoretic treatment of multiagent systems. *Cybern. Syst.*, 23(6):621–633.

Kelemen, J., Kelemenová, A. a Păun, Gh. (2004). Preview of P colonies: A biochemically inspired computing model. V *Workshop and Tutorial Proceedings. Ninth International Conference on the Simulation and Synthesis of Living Systems (Alife IX)*, str. 82–86, Boston, Massachusetts, USA.

Mirjalili, S., Mirjalili, S. M. a Lewis, A. (2014). Grey wolf optimizer. *Advances in Engineering Software*, 69:46–61.

Păun, Gh., Rozenberg, G. a Salomaa, A. (2010). *The Oxford Handbook of Membrane Computing*. Oxford University Press, Inc., New York, NY, USA.

Păun, Gh. (2000). Computing with membranes. *J. Comput. Syst. Sci.*, 61(1):108–143.

Valenta, D., Langer, M., Ciencialová, L. a Cienciala, L. (2021). On numerical 2d p colonies with the blackboard and the gray wolf algorithm. Freund, R., Ishdorj, T.-O., Rozenberg, G., Salomaa, A. a Zandron, C. (zost.), V *Membrane Computing*, str. 161–177, Cham. Springer International Publishing.

Využití membránového systému pro simulaci komunikace v síti Internetu věcí

Šárka Vavrečková

Ústav informatiky, Slezská univerzita v Opavě
Bezručovo nám. 13, Opava
Email: sarka.vavreckova@fpf.slu.cz

Abstrakt

V počítačových sítích používaných pro propojení zařízení Internetu věcí (IoT) se používá několik komunikačních modelů postavených na principu paralelní komunikace, ať už synchronní nebo asynchronní. Tyto modely mohou být simulovány s využitím membránových systémů. V článku se zaměříme na simulaci jednoho z těchto komunikačních modelů, Publisher-Subscriber, právě pomocí membránového systému.

1 Úvod

Historie membránových výpočtů se píše přibližně od roku 1998, kdy byly představeny Gheorghem Păunem jako paradigma pro paralelní distribuované výpočty. V této oblasti probíhá neustále vývoj a objevují se nové možnosti využití tohoto paradigmatu. Podrobné informace lze najít například v Păun (2002); Păun and Rozenberg (2002); Păun et al. (2010), nebo na webu `http://psystems.eu` [2022-05-12].

Membránové výpočty provádíme nad hierarchickou strukturou membrán, přičemž používáme operace definované pro daný typ membránového systému. Matematický model membránového systému se nazývá P Systém, kde „P“ je odkazem na jeho tvůrce Gheorghe Păuna.

Jedním z trendů současnosti je Internet věcí (Internet of Things, IoT), a tento trend ovlivňuje život čím dál více lidí. IoT zahrnuje nejrůznější typy systémů – od jednoduchých senzorů (které například detekují teplotu, vlhkost, světlo, pohyb, změny vzdálenosti, stoupající hladinu vody třeba kolem pračky, popřípadě RFID čipy), přes různé aktuátory či jiná zařízení reagující na stanovený podnět (například mechanismy pro otevření okna, rozsvícení či zhasnutí světla, různé alarmy), pasivní příjemce dat (což můžou být displeje zobrazující údaje od napojených senzorů) až ke komplexním zařízením kombinujícím více výše uvedených zařízení. Do sítě mohou patřit také různé komunikační body, brány nebo další síťová zařízení, napojit lze i běžná výpočetní zařízení typu mobilů, ze kterých můžeme kontrolovat údaje ze senzorů. S Internetem věcí se setkáváme v domácnosti, ve firmách, obchodech, ale také v průmyslu, zemědělství včetně polí či vinic, ale také tato zařízení mohou být začleněna do infrastruktury měst.

Aby bylo vůbec možné využít potenciál IoT, musí být tato zařízení propojena v síti a musí být stanoven vhodný způsob komunikace mezi nimi. Pro komunikaci mezi IoT zařízeními postupně vyplynulo několik různých konceptů. Vzhledem k tomu, že IoT zařízení bývají často napájená z malé baterie a ne vždy je snadné tuto baterii vyměnit, je třeba volit spíše energeticky úsporný provoz a také energeticky úsporný způsob komunikace.

Pak je tu ještě jedno specifikum: v současnosti je čím dál víc kybernetických útoků mířeno právě na IoT zařízení a přitom data těmito zařízeními odesílaná mohou patřit k bezpečnostně „citlivým“, zvláště pokud jde o údaje o zdraví generované různými čidly monitorujícími fyziologické funkce pacientů. Otázka bezpečnosti ukládání a transferu dat z IoT zařízení je často diskutována, už proto, že například prostřednictvím webu Shodan (`https://www.shodan.io/` [2022-05-13]) je snadné najít přístupové body do domácích či firemních IoT sítí včetně informací o jejich zabezpečení.

Protože zařízení v síti (včetně zařízení IoT) komunikují paralelně, můžeme pro popis této komunikace použít membránový systém, kde lze modelovat tok dat mezi zařízeními taktéž paralelně.

Použití membrán nebo podobných konceptů ve světě IoT, případně při jejich simulaci, není úplnou novinkou. V článku Villari et al. (2016) je představen koncept „osmotic computing“ jako paradigma, jehož hlavním účelem je zvýšení přístupnosti zdrojů a služeb v síti, včetně cloudových služeb. Některé tzv. mikroslužby tradičně poskytované v cloudu (což znamená z velkých datových center) se mohou částečně posunout na hranici sítě (hovoříme o edge computingu), tedy vlastně běží na zařízeních v naší síti pod naší kontrolou. Toto paradigma je motivováno procesy z přírodních věd, kde molekuly rozpouštědla procházejí přes polopropustnou membránu do jiných oblastí v prostředí s vyšší koncentrací rozpouštěné látky (tedy osmóza). Tento koncept je dále rozvíjen v článku Sharma et al. (2017), který upřesňuje, jak mohou zejména mikroslužby migrovat mezi cloudem a naší sítí, a zaměřuje se více na svět IoT.

Autoři nazývají v Villari et al. (2017) výše popsané paradigma „osmosis“ a staví na něm dynamickou správu zdrojů a služeb v IoT (používají pojem MELs, což znamená MicroElements), přičemž si všímají především bezpečnosti komunikace v IoT síti. V článku

Datta and Bonnet (2018) je ukázáno využití MELs při zapojení různých zařízení, především chytrých automobilů.

V předchozích zmíněných článcích MELs procházejí mezi zařízeními pomocí tzv. softwarově definovaných membrán (Software-Defined Membranes, SDM). Princip SDM je velmi dobře popsán například v Villari et al. (2017). Ovšem SDM je představováno jako speciální rozhraní přidané k zařízení, nejde přímo o komunikaci mezi (původními) IoT zařízeními. Naproti tomu v tomto článku reprezentujeme přímo IoT zařízení pomocí membrán a jejich komunikaci popisujeme pomocí evolučních pravidel. Je sice přidán pseudokód, ale jeho úkolem je spíše transformovat data generovaná IoT zařízeními na objekty v membránách a naopak z objektů v membránách vytvořit data, se kterými dokáže pracovat přijímající zařízení. Samotná komunikace je téměř plně v režii membránového systému. Koncept je také rozvinut v Vavreckova (2021).

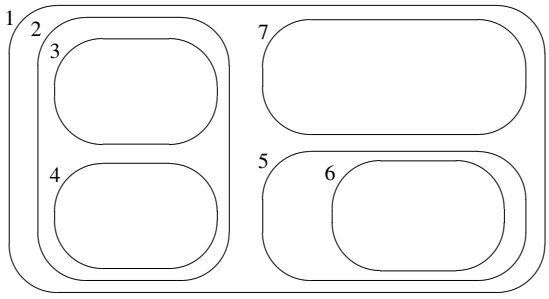
2 Prerekvizity

2.1 Membránové systémy

Dále budeme předpokládat, že čtenář ovládá základy teorie formálních jazyků a membránových výpočtů. Pokud ne, doporučujeme například zdroje Hopcroft and Ullman (1979) a Păun et al. (2010).

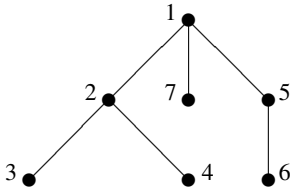
V membránových systémech je základem hierarchická struktura membrán inspirovaná strukturou biologické buňky. Membrána dále může obsahovat objekty, případně také vnořené membrány. S objekty lze pracovat s využitím pravidel, v některých membránových systémech existují také pravidla pracující se samotnými membránami.

Jedna z membrán je hlavní (tzv. „skin membrane“), ta obsahuje všechny ostatní membrány. Ostatní membrány mají vždy jednu rodičovskou membránu, ve které jsou obsaženy.



Obr. 1: Příklad membránové struktury

Strukturu membrán můžeme reprezentovat různě. Existují „graficky orientované“ způsoby zápisu, jak vidíme na obrázku 1 (pomocí Vennova diagramu) nebo na obrázku 2 (ukazuje vnoření membrán pomocí stromu – nadřízený uzel znamená nadřízenou membránu). Jiný způsob zápisu by byl pomocí vnořených závorek opatřených návěštími, podle našeho příkladu $[\,[\,[\,]_3\,[\,]_4\,]_2\,[\,]_5\,[\,[\,]_7\,]_6\,]_1$.



Obr. 2: Stromová struktura

Jednotlivé membrány jsou označeny návěštími, v našem případě (na obou obrázcích) jsou jako návěští použita čísla.

Každá membrána má svůj region, což je oblast membránou ohraničená. V regionu membrány jsou objekty a vnořené membrány (které opět mohou mít své objekty a vnořené membrány). V zobrazení formou stromu region membrány odpovídá podstromu uzlu označeného návěštím dané membrány.

Definice 1 (Păun and Rozenberg (2002), Busi (2007)) *Nechť H je množina návěští. Potom P Systém stupně m, $m \geq 1$, je*

$$\Pi = (V, \mu, w_1, \ldots, w_m, R_1, \ldots, R_m)$$

kde:

(i) *V je neprázdná abeceda, její prvky nazýváme objekty,*

(ii) *μ je membránová struktura skládající se z m membrán, membrány jsou označeny návěštími z H,*

(iii) *w_i, $1 \leq i \leq m$, jsou řetězce reprezentující multimnožiny nad abecedou V (tj. objekty) nacházející se v regionu i-té membrány v μ,*

(iv) *R_i, $1 \leq i \leq m$, jsou konečné množiny evolučních pravidel příslušejících k i-té membráně v μ; evoluční pravidlo je uspořádaná dvojice (u, v), také lze zapsat jako $u \to v$, kde*

- *u je řetězec objektů nad V,*
- *$v = v'$ nebo $v = v'\delta$, kde v' je řetězec nad $\{a_{here}, a_{out}, a_{in_j} \mid a \in V,\ 1 \leq j \leq m\}$, dále δ je speciální symbol $\notin V$ značící operaci rozpuštění membrány.*

Objekty mohou být přenášeny přes hranice membrán pomocí evolučních pravidel podle určení cíle out *(do rodičovské/nadřízené membrány) nebo* in *(do podřízené membrány určené indexem), nebo mohou zůstat v původní membráně při určení cíle* here.

Podrobnosti a příklady můžeme najít například ve zdrojích Păun and Rozenberg (2002) a Busi (2007).

2.2 Internet věcí

Existuje mnoho definic Internetu věcí, ale žádná není plně popisná – tento pojem je totiž velmi těžké pevně uchopit, záleží na konkrétním využití, podmínkách,... Ve zdroji Kassab and Darabkh (2020) můžeme najít dokonce pět definic převzatých z různých zdrojů, a navíc několik částečných definic. Pro naše účely můžeme z nich zkomponovat a dále používat například tuto definici:

Definice 2 (Kassab and Darabkh (2020)) *Internet věcí (IoT) je síť různých typů „chytrých" objektů (věcí) a zařízení. Tyto věci mohou být připojeny do Internetu a navzájem komunikují s co nejmenší potřebou zásahu člověka. Mají zabudovány funkce jako je snímání, analyzování, zpracování a základní správa sebe samotného, to vše založené na komunikačních protokolech. Tyto „chytré" věci by měly být vybaveny jednoznačnou identifikací.*

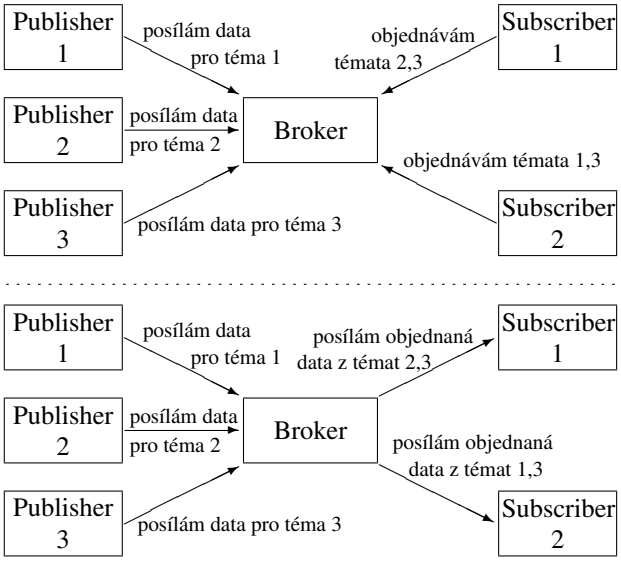
Komunikace v IoT obvykle probíhá způsobem klient-server, tedy serverové zařízení poskytuje informace a klientské zařízení si tyto informace může vyžádat a přijmout. Pro IoT existuje několik komunikačních modelů (vzorů) založených na komunikaci klient-server, nejběžnější jsou modely Request-Response a Publisher-Subscriber.

Model Request-Response se používá spíše v tradičních počítačových sítích. Klient pošle žádost o data konkrétnímu serveru, server odpoví požadovanými daty. Problém tohoto modelu je, že klient musí vědět, komu žádost poslat, potřebuje tedy seznam adres serverů, od kterých buď pravidelně nebo na podnět od jiného zařízení či uživatele bude potřebovat data.

Model Publisher-Subscriber je pro IoT poněkud praktičtější. Počítá s prostředníkem, který řeší problém nastíněný u prvního popsaného modelu. Máme tři typy komponent: *publisher* vytváří data, *subscriber* odebírá a zpracovává data, a *kontroler* (v implementacích obvykle nazvaný *broker*) je centrální řídicí prvek IoT sítě, který zprostředkovává komunikaci mezi oběma předchozími typy komponent. V tomto článku se budeme držet terminologie používané pro níže jmenovaný protokol MQTT.

Publisher posílá data brokerovi, ne přímo subscriberům, vlastně vůbec nepotřebuje vědět, že v síti je někdo jiný než broker. Různé druhy dat posílaných v IoT síti jsou zařazeny do kategorií, které se obvykle nazývají *témata*. Témata jsou v hierarchické struktuře, která umožňuje jednoduše určit celou skupinu témat tak, že označíme nadřízené téma ve struktuře (například pokud chceme dostávat informace o teplotě ze všech dostupných teplotních čidel, nebudeme postupně vyjmenovávat témata pro jednotlivá čidla, ale použijeme nadřízené téma „všechna teplotní čidla"). Publisher tedy brokerovi posílá data, u kterých určil, do kterého tématu příslušejí. Subscriber si (předem) u brokera objednal odběr dat patřících do určitého tématu (případně více témat), tedy opět nespecifikuje zdroj dat (publishera), pouze téma. Broker pak subscriberovi posílá veškerá data příslušející do objednaných témat.

Nejrůznější senzory jsou typickým příkladem publisherů (tedy producentů dat), aktuátory a displeje či mobilní telefony jsou příklady subscriberů (konzumentů dat).



Obr. 3: Ukázka modelu Publisher-Subscriber: nejdřív registrace k odběru témat (od konkrétních publisherů), pak plnění objednávek

Na obrázku 3 vidíme příklad sítě se třemi publishery, dvěma subscribery a jedním brokerem. Na tomto obrázku každý subscriber odebírá data ze dvou různých zdrojů (od dvou různých publisherů).

Pro IoT sítě existuje poměrně hodně různých protokolů vyšších vrstev, můžeme jmenovat například MQTT, CoAP, XMPP. Také lze použít jednoduše HTTP známý z běžných počítačových sítí. Většina ze jmenovaných protokolů umí komunikovat pomocí obou modelů – Request-Response i Publisher-Subscriber, nicméně jak bylo výše poznamenáno, druhý jmenovaný je pro svět IoT obvykle vhodnější.

Další informace o modelech a protokolech pro IoT sítě najdeme například ve zdroji Dizdarević et al. (2019).

3 Určení membránové struktury

IoT systém může být buď napojený na Internet (přes vhodnou bránu, která obsahuje implementaci jak IoT protokolů, tak i běžných síťových protokolů), nebo může být od Internetu oddělen. Druhá možnost je samozřejmě lepší z bezpečnostních důvodů, bohužel poněkud snižuje dostupnost daného systému. Pro zjednodušení budeme předpokládat druhou možnost, téma

článku se týká pouze modelování komunikace uvnitř IoT systému.

Systém bude tedy reprezentován membránovou strukturou a vhodnými evolučními pravidly, a abychom to vše dokázali propojit s reálnými zařízeními, budeme potřebovat jakési překladové rozhraní, které bude sloužit k transformaci dat ze zařízení (senzoru) na objekty membránového systému a naopak transformaci objektů na data, se kterými si poradí zařízení-příjemce.

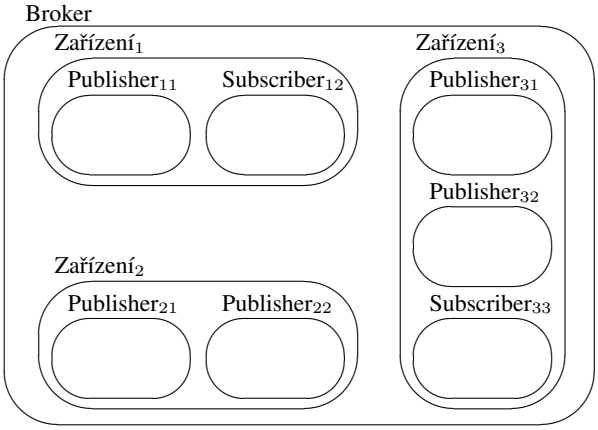
Objekty zpracovávané membránami potřebují také sémantickou informaci, jako jsou například specifická data ze senzorů, určení odesílatele apod., na což běžně pojímané objekty membránového systému nestačí (používáme konečnou množinu objektů, ale data mohou být obecně jakákoliv, jejich sémantickou informaci nelze předem omezit konečnou množinou). Například teplotní senzor může generovat čísla, která všechna jsou sice teplota, ale její hodnota může být velmi různá. K dané informaci se nám také bude hodit označení konkrétního teplotního senzoru, který byl zdrojem daného údaje, protože tím broker určí „téma“.

4 Simulace modelu Publisher-Subscriber

4.1 Navržení membránové struktury

K vytvoření membránové struktury můžeme přistoupit několika různými způsoby, při každém z nich bychom použili jiná evoluční pravidla. Protože komunikace prochází vždy přes brokera, můžeme buď jednotlivá zařízení IoT sítě včetně brokera reprezentovat membránami vnořenými do hlavní membrány, nebo broker může plnit roli hlavní membrány. V obou případech budou objekty přenášeny přes brokerovu membránu.

Zde se věnujeme druhému způsobu, kdy broker bude reprezentován hlavní membránou. Proto struktura membrán bude mít tři úrovně: uvnitř brokerovy membrány je druhá úroveň (přímí potomci hlavní membrány), což jsou IoT zařízení, a třetí úroveň (uvnitř membrán IoT zařízení) jsou komponenty obsažené uvnitř zařízení fungující jako publisher nebo subscriber. Každé IoT zařízení může obsahovat jakýkoliv počet publisherů a/nebo subscriberů, ovšem vždy nejméně jednu z těchto komponent.



Obr. 4: Příklad membránové struktury pro model Publisher-Subscriber

Na obrázku 4 je ukázka takto vytvořené membránové struktury s jedním brokerem a třemi IoT zařízeními. První zařízení obsahuje jednoho publishera (senzor) a jednoho subscribera (aktuátor), v druhém zařízení máme dva publishery (to může být například meteorologická stanice se senzory pro teplotu a vlhkost). Třetí zařízení obsahuje dva publishery a jednoho subscribera, to může být například mobilní telefon vybavený gyroskopem a akcelerometrem, subscriber je pak displej telefonu.

Jak vidíme, k datům můžeme dodávat dva identifikátory pro odesílatele (ID odesílajícího zařízení a ID konkrétního publishera v zařízení) a dva pro příjemce (ID přijímajícího zařízení a ID konkrétního subscribera v zařízení). Ne nutně všechny budou na každém úseku transportu objektu mezi membránami, například na cestě od publishera k brokerovi ještě neznáme identifikátory pro příjemce.

4.2 Evoluční pravidla a objekty

Nejdřív navrhneme objekty, které budou transportovány mezi membránami. Jak bylo výše uvedeno, potřebujeme do objektu dodat určité sémantické informace:

- přenášená data (teplota, vlhkost, vzdálenost, pohyb, snímek z kamery apod.),
- směr, resp. úsek cesty, což je informace pro brokera (od publishera nebo k subscriberovi),
- identifikátory na straně publishera (device ID, component ID),
- identifikátory na straně subscribera (device ID, component ID).

Pro objekty budeme používat následující zápis:

- $p(\langle \mathit{data} \rangle, \mathit{pub\_deviceID}, \mathit{pub\_componentID})$
  pro úsek cesty od publishera k brokerovi (nepoužijeme identifikátory subscribera),
- $s(\langle \mathit{data} \rangle, \mathit{pub\_deviceID}, \mathit{pub\_componentID}, \mathit{sub\_deviceID}, \mathit{sub\_componentID})$
  pro úsek cesty od brokera k subscriberovi.

Použili jsme různé symboly p a s, protože jinak by se komplikovaněji odlišovalo, na kterém úseku cesty vzhledem k brokerovi se objekt právě nachází.

Postupně zkonstruujeme všechna potřebná evoluční pravidla. Pro všechny publishery Pub_{ij}, kde i je identifikátor zařízení a j je ID komponenty, vytvoříme pravidlo:

$$p(\langle \mathit{data} \rangle, i, j) \rightarrow p(\langle \mathit{data} \rangle, i, j)_{out}$$

Tento symbol samozřejmě musí „nějak“ vzniknout, tomu se budeme věnovat v kódu publishera v následující podkapitole. Předpokládejme tedy, že v membráně

komponenty se objeví symbol nacházející se na levé straně pravidla. Podle pravidla je tento symbol (beze změny parametrů) přenesen do rodičovské membrány určující zařízení i.

Dále potřebujeme, aby objekt opustil také membránu zařízení a přesunul se do membrány brokera. Pro každé zařízení Dev_i a pro všechny komponenty s indexem j uvnitř daného zařízení (tj. Pub_{ij}):

$$p(\langle data \rangle, i, j) \rightarrow p(\langle data \rangle, i, j)_{out}$$

Teď je tedy objekt v hlavní membráně. Úkolem brokera je teď změnit typ objektu (místo p teď bude s) a vytvořit tolik kopií objektu, kolik subscriberů bude chtít tento objekt odebrat (je třeba vytvořit vlastní kopii pro každého subscribera, který odebírá téma, do něhož objekt patří). Ke každé vytvořené kopii je třeba vedle informací o publisherovi přidat i informaci určující daný cíl, subscribera (to znamená, že každá kopie se bude právě v tomto údaji lišit, takže pojem „kopie“ není zcela přesný). Rozkopírování objektu by se sice také dalo řešit pravidlem (což je rozvedeno v následující podkapitole), ale zde to necháme na pseudokódu uvedeném v následující podkapitole.

Předpokládejme tedy, že transformace a rozkopírování jsou za námi, následuje další úkol a další pravidla – přenos objektů z hlavní membrány do membrán cílových zařízení. Pro všechna zařízení Dev_i a jejich komponenty s indexy j, a také pro všechny možné publishery Pub_{kl}:

$$s(\langle data \rangle, k, l, i, j) \rightarrow s(\langle data \rangle, k, l, i, j)_{in_{\text{Dev}_i}}$$

Poslední část cesty je transport příslušného objektu do cílové komponenty v zařízení, tedy do membrány třetí úrovně. Pro všechna zařízení Dev_i, jejich komponenty v roli subscriberů s indexy j a všechny potenciální publishery Pub_{kl}:

$$s(\langle data \rangle, k, l, i, j) \rightarrow s(\langle data \rangle, k, l, i, j)_{in_{\text{Sub}_{ij}}}$$

4.3 Přídavný kód

Dříve uvedená evoluční pravidla řeší pouze přenos objektů mezi membránami, ale neřeší přímou interakci s příslušnými komponentami. Tedy je potřeba přidat tyto funkce:

- publisher potřebuje generovat data a vytvořit z nich objekty,
- subscriber musí poslat brokerovi objednávku k určitému tématu (pro naše účely stačí, aby objednávka určovala konkrétní publishery),
- broker potřebuje transformovat příchozí publikované objekty p na odpovídající objednané objekty s včetně rozkopírování pro všechny subscribery, kteří si objekt objednali,
- subscriber musí být schopen příchozí objekt převést do formy dat a zpracovat.

První úsek kódu (algoritmus 1) určuje vlastnosti objektů, formát brokera, zařízení a jejich komponent.

Algoritmus 1: Vlastnosti objektu, komponenty, zařízení a brokera

object:
objType, *// from_publisher (p) | to_subscriber (s)*
data,
// identifikátory publishera a subscribera:
pub_devID, pub_compID,
sub_devID, sub_compID;

component:
deviceID, compID,
compType; *// publisher | subscriber*

publisher (child of: **component**);
subscriber (child of: **component**);

device:
deviceID,
components [] = list of **component**;

subscription:
pub_devID, pub_compID,
sub_devID, sub_compID;

broker:
devices [] = list of **device**,
subscriptions [] = list of **subscription**;

U některých položek je určen typ „list of“. Je to pouze pseudokód, ve kterém předpokládáme, že tyto položky jsou seznamy či pole objektů, přičemž existuje způsob (jako třeba členské funkce, metody), jak přidat nový prvek do seznamu, získat určitý prvek apod.

Podle algoritmu 1 má objekt svůj typ, data a pak identifikátory zdroje (publishera) a cíle (subscribera). Komponenta je identifikována číslem zařízení a číslem komponenty v rámci zařízení. Komponenta může být buď publisher nebo subscriber. Zařízení má kromě svého identifikátoru také seznam obsažených komponent. Pro účely evidence objednávek jsme definovali také typ pro záznam položky registrace objednávek, který využije broker, kromě těchto záznamů si broker vede také seznam zařízení.

V algoritmu 2 na následující straně jsou uvedeny funkce pro brokera. První funkce popisuje běžné chování brokera, což znamená smyčku, ve které se kontroluje, zda se v membráně brokera (tj. hlavní membráně) neobjevil objekt přicházející od některého publishera. Pokud se takový objekt objeví, je nahrazen řetězcem objektů určených pro jednotlivé subscribery – broker projde seznam objednávek, zjistí, kdo se registroval k odběru objektů z daného zdroje, a pro každého objednatele vytvoří vlastní kopii objektu. Každý takto vytvořený objekt bude kromě informací z originálu nést také informaci o cíli (subscriberovi).

Algoritmus 2: Broker – funkce

```
function broker.Start()
begin
    while true do if ((self.presentObjectInMembrane(&object)) and (object.objType = from_publisher)) then
        // objekt, který obdržím od publishera, rozkopíruji pro všechny subscribery, kteří ho objednali:
        self.removeObjectFromMembrane(&object);    // vyjmu objekt z membrány, uložím si jeho údaje
        resObjects = "";    // vytvořím multimnožinu objektů
        object.objType = to_subscriber;
        for (i = 1 to self.devices.count) do
            for (j = 1 to self.devices[i].components.count) do
                // pokud mám objednávku na tento objekt od určitého subscribera, vytvořím pro něj objekt:
                if (self.subscriptions.find(object.pub_devID, object.pub_compID, i, j)) then
                    object.sub_devID = i;
                    object.sub_compID = j;
                    resObjects.add(object);    // vytvořený objekt přidám do multimnožiny
                end
        self.exportToMembrane(resObjects);
    end
end

function broker.OrderSubscription(pub_devID, pub_compID, sub_devID, sub_compID)
begin
    if ((self.devices[pub_devID].components[pub_compID].compType == = publisher) and
      (self.devices[sub_devID].components[sub_compID] = = subscriber)) then
        self.subscriptions.add(pub_devID, pub_compID, sub_devID, sub_compID);
end
```

Tento problém by se ovšem dal řešit pomocí evolučního pravidla, které by odstranilo původní objekt od publishera a nahradilo jej řetězcem či multimnožinou objektů pro jednotlivé subscribery:

$$p(\langle data\rangle, i, j) \rightarrow \bigcup_{k,l} s(\langle data\rangle, i, j, k, l)_{here}$$

(postupně pro všechny možné publishery), kde i, j určuje publishera (zdroj) a k, l určuje subscribera (cíl), přičemž cíle by byly podle seznamu registrací k odběru od daného publishera. Nicméně, tento postup by nebyl moc optimální: jakákoliv změna v registracích k odběru by znamenala změnu těchto pravidel, což v membránovém systému není ideální způsob činnosti (nicméně existují typy membránových systémů, ve kterých je možné dynamicky měnit sadu pravidel).

U brokera máme ještě druhou funkci – OrderSubscription(), která je volána subscriberem (nebo alternativně zařízením obsahujícím subscribera) v případě, že se chce registrovat k odběru objektů/dat od určitého publishera. Broker zkontroluje, zda údaje na objednávce jsou v pořádku, a přidá záznam do seznamu objednávek. K tomu by také mohla existovat funkce s opačným významem – odhlášení z odběru.

Nyní se zaměřme na další algoritmus (3), pro publishera. První uvedená funkce provede inicializaci nově vytvořeného publishera, včetně nastavení potřebných parametrů.

Algoritmus 3: Publisher – funkce

```
// Inicializační funkce volaná hostitelským
   zařízením, je vytvořen nový publisher
   a zařazen do pole komponent:
function publisher.Start(deviceNum,
  compNum,...)
begin
    self.compType = publisher;
    self.deviceID = deviceNum;
    self.compID = compNum;
    // Pokud má publisher vytvářet data
       v pravidelných intervalech, nastavíme
       délku intervalu, případně další parametry.
end

// Tuto funkci spouštíme vždy, když jsou
   vygenerována data a je třeba z nich vytvořit
   objekt:
function publisher.Produce()
begin
    new object;
    object.objType = from_publisher;
    object.data = GetDataFromSensor();
    object.pub_devID = self.deviceID;
    object.pub_compID = self.compID;
    self.exportToMembrane(object);
end
```

Publisher může vytvářet data v pravidelných intervalech (například teplotní čidlo posílající zjištěnou teplotu každých 10 sekund), nebo na vyžádání (uživatel stiskne tlačítko, případně je to reakce na činnost jiné komponenty nebo vnějšího stavu – například světelné čidlo detekuje přímý sluneční svit a vyšle signál, který si objednal motorek stahující rolety, atd.). Druhá funkce (Produce()) je tedy volána vždy, když publisher vyprodukoval data: jejím účelem je převést dat na objekt a vložit do membrány, kde je tento objekt zpracován příslušným evolučním pravidlem.

Algoritmus 4: Subscriber – funkce

```
// Inicializační funkce volaná hostitelským
   zařízením, je vytvořen nový publisher
   a zařazen do pole komponent:
function subscriber.Start(deviceNum,
 compNum,...)
begin
    self.compType = subscriber;
    self.deviceID = deviceNum;
    self.compID = compNum;
    // atd. další parametry podle potřeby.
    // Hned při startu subscribera se spustí
       smyčka, ve které čekáme na příchozí
       objekt, abychom ho mohli vyzvednout
       z membrány a zpracovat:
    while true do
        if ((ObjectFound(&object)) and
          (object.objType == to_subscriber))
          then
            self.removeObjectFromMembrane(object);
            self.ProcessData(object.data,
              object.pub_devID,
              object.pub_compID);
        end
    end
end

// Subscriber odešle brokerovi registraci
   k odběru dat/objektů od určitého publishera:
function subscriber.Subscribe(devID, compID)
begin
    broker.OrderSubscription(devID, compID,
      self.deviceID, self.compID);
end
```

V algoritmu 4 je kód pro subscribera. Po inicializaci komponenty je přímo spuštěna smyčka, která při zjištění příchozího objektu v membráně subscribera tento objekt transformuje na data a zpracuje.

Následuje funkce, pomocí které subscriber využije funkci z kódu brokera a registruje se k odběru z daného zdroje.

5 Diskuse

Už v předchozích sekcích je uvedeno, že některé postupy lze implementovat jinak. Může být jinak uspořádaná membránová struktura, případně určité postupy mohou být provedeny buď evolučními pravidly nebo přídavným kódem...

V úvodní části byla diskutována možnost, že brokera umístíme na stejnou úroveň jako jiná zařízení. To by znamenalo, že v membránové struktuře bude membrána brokera uvnitř hlavní membrány. Pak by bylo třeba pozměnit evoluční pravidla: symboly „p“ by z hlavní membrány musely být transportovány do membrány brokera, tam by proběhla transformace na symboly „s“, které by následně byly přeneseny přes hlavní membránu do membrán jednotlivých zařízení.

Takto upravené řešení má jednu podstatnou výhodu: také broker by mohl obsahovat své komponenty typu publisher a subscriber, tedy by mohlo jít o komplexnější zařízení. Takové zařízení si můžeme představit jako centrální řídicí jednotku opatřenou displejem (v roli subscribera) a dotykovou vrstvou na obrazovce (případně tlačítky) pro řízení celé IoT sítě (v roli publishera).

Výše naznačená úprava by znamenala zvýšení komplexnosti systému. Pokud bychom naopak chtěli snížit komplexnost systému (například využitím menšího počtu membrán), pak přichází v úvahu dvouúrovňová struktura: v hlavní membráně (broker) bychom měli přímo membrány reprezentující subscribery a publishery bez vazby na konkrétní zařízení, v němž se nacházejí. Výhodou by byla kratší cesta objektů při transportu mezi publisherem a subscriberem, za nevýhodu můžeme považovat horší přehlednost takového systému.

Další potenciální úprava by mohla být v rozšíření komunikace směrem ven – můžeme být napojeni na cloud, který by mohl plnit roli brokera, nebo můžeme jen přidávat do systému zařízení připojující se do IoT sítě přes Internet nebo klasickou počítačovou síť (broker by zůstal ve vnitřní síti). To by odpovídalo tomu, s čím se můžeme u IoT setkávat v realitě, ovšem to předpokládá důkladné zabezpečení napojení IoT sítě „ven“ pro případ, že náš broker by byl nalezen pomocí dříve zmíněného vyhledávače Shodan.

Pokud bychom nepoužili model Publisher-Subscriber a zůstali u tradičního Request-Response, pak bychom nepoužili brokera, nicméně hlavní membrána by přesto plnila roli „pošťáka“. Každý cíl dat by buď na nějaký podnět nebo v pravidelných intervalech posílal žádost (request) o data konkrétnímu zdroji dat, načež by zdroj dat odpověděl (response) datovou jednotkou s požadovanými daty. To by ovšem znamenalo téměř dvojnásobné zatížení IoT sítě provozem – v modelu Publisher-Subscriber totiž místo opakujících se žádostí o data máme jednoduše objednávku (registraci) evidovanou u brokera.

Poděkování

Tento příspěvek je financován ze Strukturálních a investičních fondů Evropské unie OP VVV, z projektu „Zvýšení kvality vzdělávání na Slezské univerzitě v Opavě ve vazbě na potřeby Moravskoslezského kraje“, CZ.02.2.69/0.0/0.0/18_058/0010238.

Reference

Busi, N. (2007). Causality in membrane systems. *Membrane Computing*, pages 160–171.

Datta, S. K. and Bonnet, C. (2018). Next-generation, data centric and end-to-end iot architecture based on microservices. In *IEEE International Conference on Consumer Electronics – Asia (ICCE-Asia)*, pages 206–212.

Dizdarević, J., Carpio, F., Jukan, A., and Masip-Bruin, X. (2019). A survey of communication protocols for internet of things and related challenges of fog and cloud computing integration. *Association for Computing Machinery*, 51(6).

Hopcroft, J. E. and Ullman, J. D. (1979). *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley.

Kassab, W. and Darabkh, K. (2020). A–z survey of internet of things: Architectures, protocols, applications, recent advances, future directions and recommendations. *Journal of Network and Computer Applications*, 163.

Păun, G. (2002). *Membrane Computing: An Introduction*. Springer, Heidelberg.

Păun, G. and Rozenberg, G. (2002). A guide to membrane computing. *Theor. Comp. Science*, 287(1):73–100.

Păun, G., Rozenberg, G., and Salomaa, A. (2010). *The Oxford Handbook of Membrane Computing*. Oxford University Press, New York.

Sharma, V., Srinivasan, K., Jayakody, D. N. K., Rana, O., and Kumar, R. (2017). Managing service-heterogeneity using osmotic computing. In *International Conference on Communication, Management and Information Technology (ICCMIT 2017)*, Warsaw, Poland.

Vavreckova, S. (2021). Modeling communication in internet of things network using membranes. In *CEUR Proceedings of the 21st Conference Information Technologies - Applications and Theory (ITAT 2021)*, pages 195–201.

Villari, M., Fazio, M., Dustdar, S., Rana, O., and Ranjan, R. (2016). Osmotic computing: A new paradigm for edge/cloud integration. *IEEE Cloud Computing*, 3(6):76–83.

Villari, M., Fazio, M., Dustdar, S., and Rana, O. F. (2017). Software defined membrane: Policy-driven edge and internet of things security. *IEEE Cloud Computing*, 4(4):92–99.

SEKCE:

ROZŠÍŘENÝ ABSTRAKT

Skúmanie vzdialeností adverzariálnych vstupov k jednotlivým triedam v hlbokých neurónových sieťach

Iveta Bečková, Štefan Pócoš, Igor Farkaš

Fakulta Matematiky, Fyziky a Informatiky
Univerzita Komenského v Bratislave
{iveta.beckova,stefan.pocos,igor.farkas}@fmph.uniba.sk

Abstrakt

Hlboké neurónové siete dosiahujú mimoriadnu úspešnosť v rôznorodých úlohách. Avšak, sú zraniteľné adverzariálnymi vstupmi (AV). V našej práci skúmame vnútorné reprezentácie AV analyzovaním ich aktivácií na skrytých vrstvách natrénovaného klasifikátora. Navrhujeme dve metódy, ktoré sa dajú použiť na porovnanie vzdialeností k triedovo špecifickým varietam, bez ohľadu na meniacu sa dimenzionalitu vrstiev naprieč sieťou. Pomocou týchto metód sme zistili, že niektoré AV neopúšťajú proximitu variety správnej triedy. Následne sme projektovali aktivácie našich dát do 2D priestoru pomocou metódy UMAP, čím sme ukázali, že aktivácie AV sú prepletené s aktiváciami z testovacej množiny. Prepletenie sme potvrdili aj numericky, pomocou metódy soft nearest neighbour loss (SNNL).

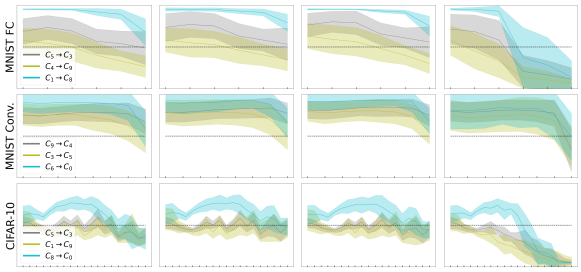
1 Úvod

Ako prví poukázali na problém adverzariálnych vstupov (AV) Szegedy a spol. (2014). AV sú vstupy do modelov strojového učenia, modifikované útočníkom za účelom prinútiť model, aby spravil chybu. AV predstavujú vážny problém, s dopadom najmä na aplikácie, kde je bezpečnosť kritická. V našej práci analyzujeme 4 typy AV vygenerovaných pre MNIST a CIFAR-10 datasety, pričom ich sila bola obmedzená v 4 rôznych L_p-normách ($p = 0, 1, 2, \infty$). Okrem toho analyzujeme aj dva typy falošne pozitívnych vstupov (Goodfellow a spol., 2015), teda nezmyselných obrázkov nepatriacich do žiadnej z tried, ktoré sú aj tak klasifikované ako jedna z tried s veľmi vysokou ($> 95\%$) konfidenciou.

Dôležitou vlastnosťou AV je ich blízkosť k varietam originálnych dát, keďže v útokoch obmedzujeme iba veľkosť perturbácie a nie jej smer. Vrámci našich analýz navrhujeme a testujeme dve metódy skúmania skrytých reprezentácií AV.

2 Podiel najbližších susedov

Na pozorovanie vývoja nesprávne klasifikovaných adverzariálnych vstupov využijeme myšlienku hľadania najbližšieho suseda v priestore aktivácií (Papernot a McDaniel, 2018):
(1) Pre danú sieť a zvolený útok si vyberieme podmožinu AV ($Adv_{C_o \to C_p}$), ktorá pozostáva z AV vygenerovaných z obrázkov prislúchajúcich k triede C_o, pričom sú nesprávne klasifikované do triedy C_p.
(2) Pre každé $\mathbf{x} \in Adv_{C_o \to C_p}$ nájdeme k najbližších susedov v priestore aktivácií. V tomto prípade uvažujeme iba o aktiváciách bodov z trénovacej množiny, ktoré patria do triedy C_o alebo C_p.
(3) Vypočítame pomer k_o/k, pričom k_o je počet aktivácií obrázkov z triedy C_o.
(4) Vizualizujeme priemerný pomer a jeho vývoj na jednotlivých vrstvách siete (Obr. 1).



Obr. 1: Vývoj pomeru k_o/k vnútri rôznych sietí. V mriežke grafov stĺpce reprezentujú individuálne útoky (zľava doprava L_0, L_1, L_2 a L_∞) a riadky zodpovedajú natrénovaným sieťam.

3 Projektovaná vzdialenosť k varietam

V druhej metóde počítame vzdialenosti aktivácií k varietam originálnych dát. Na výpočet vzdialenosti $\mathbf{x}$ k variete ju aproximujeme pomocou konvexného obalu k najbližších susedov ($\mathbf{x}_i$). Projekcia na varietu sa dá potom vyjadriť ako problém konvexnej optimalizácie:

$$\min_{\alpha_1,\ldots,\alpha_k} \left\| \left(\sum_{i=1}^{k} \alpha_i \mathbf{x}_i \right) - \mathbf{x} \right\|_2,$$

kde $\sum_{i=1}^{k} \alpha_i = 1, \quad \alpha_i \geq 0, \; i \in \{1, \ldots, k\}$.
Metóda funguje nasledovne:
(1) Vypočítame projekciu na varietu aktivácií celej trénovacej množiny, zapamätáme si indexy k najbližších

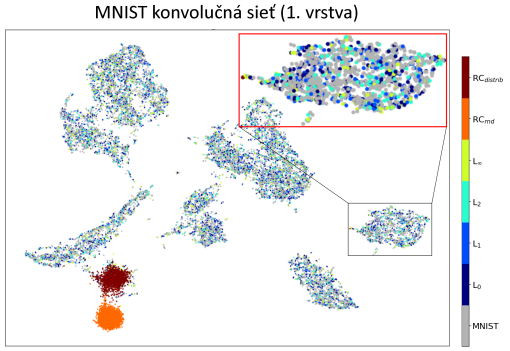
susedov a k nim prislúchajúce koeficienty α_i.
(2) Zapamätané indexy a koeficienty použijeme na výpočet korešpondujúcej konvexnej kombinácie vo vstupnom priestore (niekoľko ukážok takýchto projekcií z vrstiev naprieč sieťou je na Obr.2).
(3) Konvexnú kombináciu vo vstupnom priestore projektujeme na triedovo špecifické variety, ktoré prislúchajú triedam C_o and C_p.

Obr. 2: Projekcie aktivácií vybraných AV z jednotlivých vrstiev siete. Môžeme pozorovať postupný prechod z triedy 3 do nesprávne predikovanej triedy 5.

4 UMAP projekcie

Na vizualizáciu aktivácií na skrytých vrstvách a analýzu geometrických vlastností AV sme použili metódu UMAP (McInnes a spol., 2018), založenú na redukcii dimenzionality.

Ukážka vizualizácie aktivácií je na Obr.3, kde sme projektovali všetky typy vygenerovaných dát (4 typy AV, falošne pozitívne vstupy a MNIST dataset) pomocou metódy UMAP do 2D priestoru. Môžeme si všimnúť, že falošne pozitívne vstupy sú dobre separovné od zvyšku dát, zatiaľ čo AV a dáta z testovacej množiny sú veľmi prepletené.

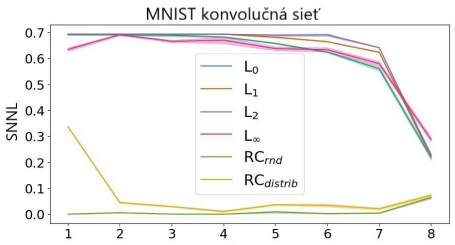


Obr. 3: Projekcia aktivácií rôznych vstupov do 2D priestoru.

5 Soft nearest neighbour loss

Jednotlivé variety sa v neurónových sieťach zvyknú najprv prepletať a potom rýchlo rozvinúť na konci siete. Avšak prepletenie AV zatiaľ nebolo skúmané. Preto počítame ich prepletenie s testovacou množinou pomocou metódy SNNL (Frosst a spol., 2019).

Veľmi prepletené dáta dosahujú vysoké hodnoty SNNL, ktorá klesá so stúpajúcou separáciou bodov z rôznych tried.



Obr. 4: Vývin prepletenia testovacích dát a rôznych typov AV naprieč sieťou, vyhodnotený pomocou SNNL.

Obe metódy - SNNL (výsledky sú zobrazené na Obr.4) a UMAP - poskytujú podobné interpretácie: AV sú prepletené s originálnymi dátami, avšak ich prepletenie v sieti postupne klesá. Na druhej strane, falošne pozitívne vstupy nie sú prepletené a ich prepletenie v sieti výrazne nestúpa.

Poďakovanie

Tento výskum bol čiastočne podporený projektom TAILOR č. 952215, v rámci výskumného a inovačného programu Horizon 2020.

Literatúra

Frosst, N., Papernot, N. a Hinton, G. (2019). Analyzing and improving representations with the soft nearest neighbor loss. V *International Conference on Machine Learning*.

Goodfellow, I. (2018). Defense against the dark arts: An overview of adversarial example security research and future research directions. arXiv:1806.04169.

Goodfellow, I., Shlens, J. a Szegedy, C. (2015). Explaining and harnessing adversarial examples. V *International Conference on Learning Representations*.

McInnes, L., Healy, J., Saul, N. a Großberger, L. (2018). UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*.

Papernot, N. a McDaniel, P. (2018). Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. arXiv:1803.04765.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. a Fergus, R. (2014). Intriguing properties of neural networks. V *International Conference on Learning Representations*.

Contextual Plasticity in Sound Localization vs. Source Separation in Real and Virtual Environments

Stanislava Linková, Gabriela Andrejková, Norbert Kopčo

Institute of Computer Science, Faculty of Science, P. J. Šafárik University in Košice
Jesenná 5, 04001 Košice
stanislava.linkova@student.upjs.sk, gabriela.andrejkova@upjs.sk, norbert.kopco@upjs.sk

Abstract

Contextual plasticity (CP) is a sound localization after-effect that is observed as bias in localization of stimuli and is occurring on the time scale of seconds to minutes. The inclusion of the stimulus adaptor in the sequence of target stimuli provides the possibility to analyze the bias from several points of view (position of the distractor, speed and size of adaptation). In this paper, we present the results of two experiments and evaluate them in the context of two existing models of spatial hearing adaptation. The first model predicts that fatigue due to extended activation reduces responses in spatial channels. The second model suggests that adaptation of spatial representation aims to improve source separation.

1 Introduction

Contextual plasticity (CP) is a localization aftereffect occurring on the time scale of seconds to minutes. It has been observed as a bias in horizontal sound localization of click target stimuli presented alone, when interleaved with contextual adaptor-target trials in which the adaptor was at a fixed location while the target location varied, (Kopčo, et al., 2007). The observed bias is always away from the contextual adaptor location, even though the adaptor is not present on the experimental trials. Here we present results of two experiments, which examined whether this phenomenon is dependent on engagement of the subject in an active localization task on the contextual trials and whether CP is also observed in virtual environments, both reverberant and anechoic. A detailed description of the experiments is available in (Linková, 2022) and (Piková, 2018).

In previous studies, two candidate mechanisms have been proposed to explain adaptation phenomena similar to CP: 1) fatigue due to extended activation reduces responses in spatial channels near adaptor location, proposed by Carlile (Carlile, et al., 2001), and 2) spatial representation adapts to improve source separation at the cost of introducing localization biases, proposed by Lingner (Lingner, et al., 2018). The Carlile's mechanism (Carlile, et al., 2001) predicts that location discrimination performance after adaptation would be worse for targets near adaptor (vs. far from adaptor), while the Lingner (Lingner, et al., 2018) mechanism suggests it would be better for targets near adaptor.

Here, we evaluate these opposing predictions for three bias-independent localization measures: stimulus-response correlation, response standard deviation and information transfer rate (ITR).

2 Methods

In the two experiments, the target stimulus was a 2-ms noise burst (click), while the adaptor was a click train consisting of 12 such clicks. Six target locations were used, ±33, ±22, ±11° in Exp. 1 and ±30, ±20, ±10° in Exp. 2 (Fig. 1). Adaptor locations were fixed across runs at 0, ±45, or ±90° in Exp. 1 and 0 or ±50° in Exp. 2. In addition, base-line runs contained no adaptors. Subjects (8 resp. 9 normal-hearing subjects in Exp. 1 resp. Exp.2) responded using a numerical keypad while seated with their heads supported by a headrest. Exp. 1 was performed in a real reverberant environment (RREn), Exp. 2 in virtual anechoic (VAEn) and reverberant (VREn) environments, simulated by using non-individualized HRTFs and BRIRs.

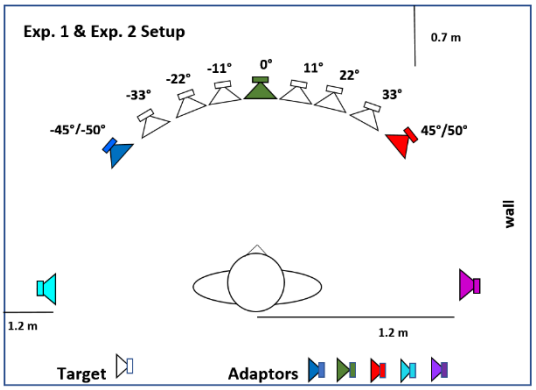


Fig. 1. Setup of experiments Exp. 1 and Exp. 2

3 Results

3.1 Bias in Responses (Fig. 2)

- depends on Adaptor location (smallest for Frontal Adaptor)
- depends on the environment (smallest for RREn and largest for VAEn)

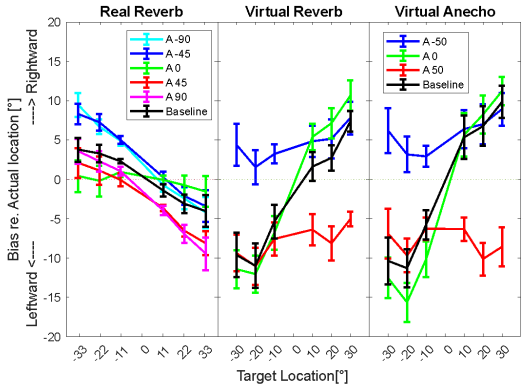


Fig. 2. Bias in responses re. actual target location averaged across time.

3.2 Information Transfer Rate (Fig. 3)

We used the information transfer rate to evaluate contextual plasticity because it is robust to linearity violation. ITR was computed for a triplet of target locations assuming left-right symmetry in the data.
Fig. 3 shows ITR is higher for contralateral adaptors than for ipsilateral ones in all environments. On the other hand, no consistent trend is observed for the information transmission for baseline vs. the frontal adapter.

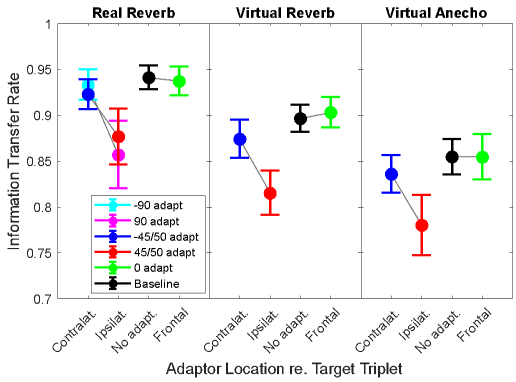


Fig. 3 Information transfer rate values for target triplets.

The results for ITR confirm our hypothesis of consistency with the model of Carlile (Carlile, 2001), because a lower value of information transmission for ipsilateral adapters indicates worse localization in the vicinity of the adapter.

3.3 Correlations and Variances in Responses

Pearson's correlation coefficient, r (for details see (Linková, et al., 2022)):

- better in RREn than in VREn, and in VREn than VAEn ($p < 0.05$)
- better for targets far (Contralateral) than near (Ipsi-lateral) the lateral adaptor ($p < 0.0001$)
- better with frontal adaptor than in no-adaptor baseline ($p < 0.005$)

Correlations are always consistent with Carlile's model (Carlile, et al., 2001)**.**

Standard deviation *re*. baseline:

- increases for targets near Adaptor in RREn ($p < 0.05$)
- no significant effect in VREn
- trend for effect in VAEn, such that standard deviation increases near Adaptor and decreases further away ($p = 0.09$)

Variances tended to increase near the adaptor location in Exp. 1. Results are more consistent with Carlile et al. model (Carlile, et al., 2001)**.**

4 Conclusion and discussion

Pearson's correlation coefficient and information transfer rate provide approximately the same evaluations of results for contextual plasticity. The results for the standard deviation also confirm the hypothesis, but some aspects of the results in virtual environment also show consistency for the Lingner's model (Lingner, et al., 2018). In a real environment, the results show consistency only for Carlile's model (Carlile, et al., 2001). So, these results suggest that listeners' behavior can differ between the real and virtual environments, as the Lingner's model is motivated by data from virtual anechoic environment, while Carlile's model is based on real anechoic environment data.

Acknowledgement

Work is supported by VEGA 1/0350/22 and APVV DS-FR-19-0025.

References

Carlile, S., 2001. Systematic distortions of auditory space perception following prolonged exposure to broadband noise. *J. Acoust. Soc. Am.,* Zväzok 110, pp. 416-424.

Carlile, S., Hyams, S. & Delaney, S., 2001. Systematic distortions of auditory space perception following prolonged exposure to broadband noise. *J. Acoust. Soc. Am.,* Volume 110, pp. 416-424.

Kopčo, N., Best, V. & Shinn-Cunningham, B., 2007. Sound localization with preceding distractor. *Journal of the Acoustical Society of America, 121,* pp. 420-432.

Lingner, A., Pecka, M., Leibold, C. & Grothe, B., 2018. A novel concept for dynamic adjustment of auditory space. *Scientific Reports,* Issue 8:8335, pp. 1-12.

Linková, S., 2022. Modelovanie kontextuálnej plasticity v reálnom a virtuálnom prostredí. *Diploma Thesis, P. J. Šafárik University in Košice.*

Linková, S., Andrejková, G, & Kopčo, N., 2022. Contextual Plasticity in Sound Localization vs. Sound Separation. *VIRTUAL 45th Annual MidWinter Meeting of the Association for Research in Otolaryngology, February 5-9,.*

Piková, V., 2018. Mechanizmy kontextuálnej plasticity v lokalizácii zvukov. *Bachalor Thesis, P. J. Šafárik University in Košice.*

A model of the reference frame of the ventriloquism aftereffect based on head-centered, eye-centered and distance-dependent signals

Ing. Peter Lokša, PhD., doc. Ing. Norbert Kopčo, PhD.

Faculty of Science, Safarik University in Košice
Jesenná 5, 04001 Košice
peter.loksa@upjs.sk, norbert.kopco@upjs.sk

1 Introduction

The ventriloquism effect is observed as a shift of the perceived location of the sound towards the synchronously presented visual adaptor. The ventriloquism aftereffect (VAE), observed as a shift in the perceived locations of sounds after audio-visual stimulation (Recanzone, 1998; Woods and Recanzone, 2004; Bertelson et al., 2006), requires reference frame alignment since hearing and vision encode space in different frames (head-centered vs. eye-centered). Previous experimental studies observed inconsistent results: a mixture of head-centered and eye-centered frames for the VAE induced in the central region (Kopco et al., 2009) vs. a predominantly head-centered frame for the VAE induced in the periphery (Kopco et al., 2019). In the current study, a computational model is introduced to examine these inconsistencies, assuming that there is a fixed relationship between the VAE and the ventriloquism effect.

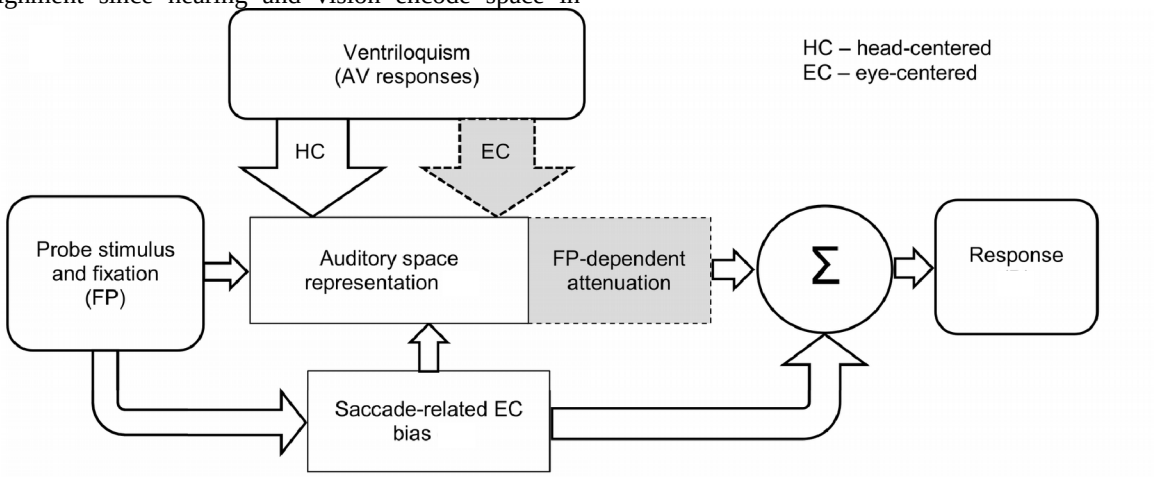


Fig. 1: Block diagram of the model. The model predicts the response bias as a function of the probe stimulus location, with additional input parameters of the fixation position, training locations, and the observed ventriloquism effect at the training locations (rounded blocks). Two mechanisms determine the response (square blocks): 1) saccade-related bias is always present and it is not influenced by the ventriloquism signals; 2) auditory space representation which is adapted by ventriloquism only in HC reference frame (HC & dHC models; "HC" arrow) or in a combination of HC and EC RFs (HEC & dHEC models; "HC" and "EC" arrow). In the dHC and dHEC models, auditory space adaptation by ventriloquism is reduced depending on the distance of the current FP from the training FP.

2 Methods

The model has two components: a saccade-related component characterizing the adaptation in auditory-saccade responses and auditory space representation adapted by ventriloquism signals in a combination of head-centered and eye-centered frames, in which the strength of adaptation can be eye-gaze-direction dependent (Fig. 1). There were 4 different model versions implemented, differing in 2 aspects. The first aspect is whether the ventriloquism aftereffect was

mixed of head- and eye-centered (HEC), or purely head-centered (HC). The second aspect is whether the gaze-direction-dependent modulation was considered (dHEC or dHC) or not (HEC or HC). The model versions were compared using AICc criterion in 4 different simulations using different data sets: no-shift, all data, central and peripheral.

3 Results

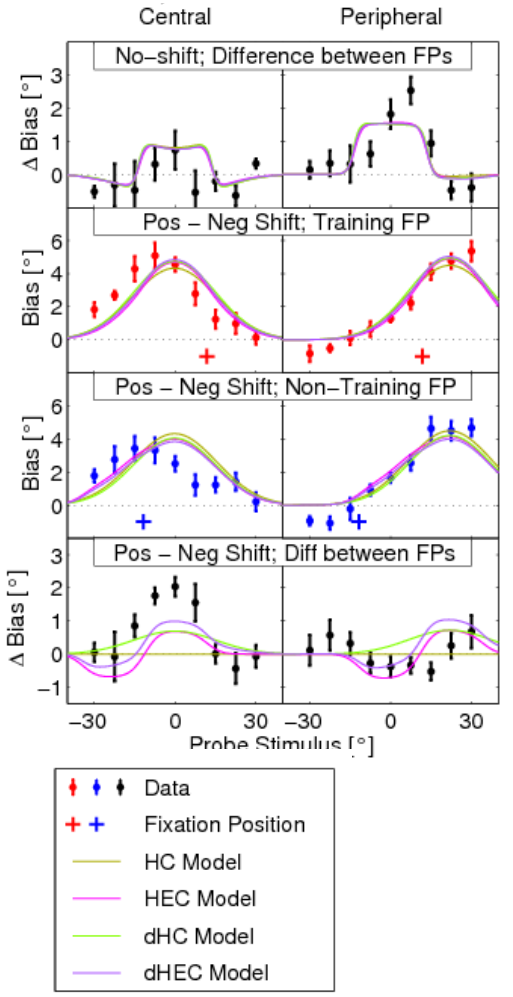


Fig. 2: Model predictions and data for the All Data simulation. Across-subject mean difference in biases from the training FP vs. non-training FP (±standard error of the mean) and model predictions for the no-shift data (top row), and for the positive-shift – negative-shift data (bottom row). The middle two rows show the positive-shift – negative-shift data separately for the training and non-training FPs.

Experimental data analysis confirmed that the VAE measured using saccades can be predicted based on observed ventriloquism effect. Overall, the model performed best when eye-centered signals were combined with head-centered signals with a gaze-direction-dependent modulation (dHEC) for all data simulation (Fig. 2). The model dHEC was 2.2 AICc points better than the second best model (dHC) here. However, for no-shift simulation where just data affected by aligned audiovisual pairs were selected, the HEC model provided the best fit to the data (previous version of simulation shown in Loksa & Kopco, 2021). HEC model here performed 2.4 AICc points better than the second best model, which is HC.

4 Discussion

There are likely to be two mechanisms by which visual signals are realigned with auditory signals. These mechanisms are combined to visually calibrate the auditory spatial representation in a mixed reference frame.

5 Acknowledgement

This work was supported by Science Grant Agency of the Slovak Republic VEGA 1/0355/20 and by Danube Region Strategy & The Slovak Research and Development Agency DS-FR-19-0025.

References

Recanzone, G.H., Rapidly induced auditory plasticity: The ventriloquism aftereffect. Proceedings of the National Academy of Sciences of the United States of America, 1998. 95(3): p. 869-875.

Woods, T.M. and G.H. Recanzone, Visually Induced Plasticity of Auditory Spatial Perception in Macaques. Current Biology, 2004. 14: p. 1559-1564.

Bertelson, P., et al., The aftereffects of ventriloquism: Patterns of spatial generalization. Perception and Psychophysics, 2006. 68(3): p. 428-436.

Kopco, N., et al., Reference Frame of the Ventriloquism Aftereffect. Journal of Neuroscience, 2009. 29(44): p. 13809-13814Kopco, N., et al., Hemisphere-specific properties of the ventriloquism aftereffect. J Acoust Soc Am, 2019. 146(2): p. EL177-183.

Loksa, P., Kopco, N., A model of the reference frame of the ventriloquism aftereffect. bioRxiv 2021.03.31.437664;

Špecifiká teórie mysle u pacientov so schizofréniou a bipolárnou afektívnou poruchou

Ivana Mirdalíková

Univerzita sv. Cyrila a Metoda v Trnave, Filozofická fakulta, Katedra psychológie
Nám J.Herdu 2, 917 01 Trnava
Nemocnica sv.Lukáša v Galante, Psychiatrické oddelenie
Hodská 373/38, 924 01 Galanta
mirdalikovai@gmail.com

Úvod

Jedinec si v priebehu života vytvára predstavy o fungovaní ľudskej psychiky. Na základe svojich predstáv je schopný porozumieť tomu, čo sa asi deje v mysli druhých ľudí. Táto schopnosť sa používa pri odhade skrytých úmyslov, ktoré nie sú v reči priamo vyjadrené. Ľudskí jedinci sa líšia v miere, v akej sú schopní ju uplatňovať. Narušená schopnosť sa spája najmä s poruchami autistického spektra (Baron-Cohen, 1999). Výsledky doterajších výskumov poukazujú na to, že práve schopnosť mentalizácie je narušená tak ako u pacientov so schizofréniou (Mazza et al., 2001; Kohler et al.,2010; Csukly et al., 2013), tak aj u pacientov s bipolárnou afektívnou poruchou (Kerr et al., 2003; Bora et al.2005). Niektorí autori opisujú vzájomný vzťah medzi teóriou mysle a emóciami (Bora et al., 2005; Csukly et al., 2013).
Hlavným cieľom výskumu bola deskripcia teórie mysle, emócií a spôsobu uvažovania u pacientov s ochorením zo schizofrénneho spektra a u pacientov s bipolárnou afektívnou poruchou a porovnanie so zdravou populáciou. Zároveň sme si kládli za cieľ zistiť, či existujú prítomné rozdiely v schopnosti mentalizácie, rozpoznávania emócií a spôsobe uvažovania medzi pacientmi ochorením rôznej diagnózy.

Výskumný súbor

Pre účely nášho experimentu boli zostavené tri výskumné súbory. Zberu dát predchádzalo stanovenie si explicitných kritérií, na základe ktorých boli jednotliví participanti zaradení do výskumu.Vzorka bola získavaná zámerným výberom. Výskumný súbor pozostával celkovo z 94 participantov. Klinickú populáciu reprezentovalo 30 pacientov s poruchami schizofrenického spektra a 30 pacientov s bipolárnou afektívnou poruchou (BAP). Pacienti boli v období experimentu v stabilizovanom stave, čo bolo inkluzívnou podmienkou pre zaradenie do výskumu. Zdravú populáciu tvorilo 34 participantov. Priemerný vek u kontrol bol M= 38,17, u schizofrenikov M= 34,96 a u bipolárnych pacientov M= 39,30.

Metóda

Na meranie schopnosti rozpoznávať mentálne stavy sme zvolili Faux Pas test, ktorý nebol doposiaľ v našich podmienkach používaný a zabezpečili sme jeho preklad. Test pozostáva z 20 príbehov, z toho 10 príbehov tvorí Faux Pas Stories a 10 príbehov je kontrolných. Faux pas príbeh hodnotí schopnosť rozpoznať a pochopiť, keď niekto neúmyselne povie alebo urobí niečo, čo zraňuje alebo uráža inú osobu. Správne identifikovať FP situáciu si vyžaduje pochopenie perspektívy vedomostí a pocitov protagonistov príbehu.
Schopnosť rozpoznávať emócie, sme merali prostredníctvom Testu čítanie mysle oči MIE. Test je tvorený súborom fotografií z oblasti očí, ktorý obsahuje celkovo 36 položiek. Úlohou participantov je vybrať jedno zo štyroch slov pri fotografii, ktoré najlepšie vystihuje to, čo si osoba myslí alebo cíti.
Preferenciu kognitívneho štýlu sme zisťovali dotazníkom CRT7. Test kognitívnej reflexie CRT7 sme sa rozhodli použiť na zachytenie preferencie participantov k rýchlym intuitívnym odpovediam, či analytickému módu spracovania problému. Správna (analytická/deliberatívna) odpoveď bola ohodnotená vždy 1 bodom. 7 bodov bol maximálny možný počet bodov v teste a vyššie skóre v teste poukazovalo na preferenciu analytického myslenia.
V klinickej populácií sme hodnotili závažnosť ochorenia škálou CGI a aktuálnu symptomatiku prostredníctvom Škály pozitívnych a negatívnych príznakov PANNS.

Výsledky

Výsledky výskumu priniesli zaujímavé zistenia, ktoré sú v značnej miere v súlade so zisteniami iných výskumov, ktoré sú prezentované v metaanalýzach. Klinická populácia mala v porovnaní so zdravou populáciou výraznejšie problémy s používaním teórie mysle (U=469; p<,01; d=1,00) a rozpoznávaním emócií (U=564,5; p<,01; d=0,79).
Hodnoty nadobudnuté vo Faux pas teste sa výrazne líšili medzi skupinami. V úrovni rozpoznávať mentálne stavy sa nám preukázali významne rozdiely medzi skupinami

(H(2)=27,139;p<,05). Vyjadrené v Cohenovom d, dosiahla veľkosť rozdielu veľmi veľký efekt (d=1,23) a v Eta squared (η2=0,276). Nasledovne bolo v Post Hoc analýze zisťované smerovanie rozdielu, kde sme zaznamenali štatisticky významné rozdiely vo všetkých skupinách. Medzi kontrolnou skupinou a pacientmi s BAP na hladine významnosti Z=-2,23 (p<,05), medzi kontrolami a schizofrenikmi na hladine významnosti Z=5,20 (p<,01) a takisto medzi klinickou populáciou navzájom Z=2,88 (p<,01).

Hodnoty nadobudnuté v teste MIE sa výrazne líšili medzi skupinami.V úrovni rozpoznávať emócie sa nám taktiež preukázali významne rozdiely medzi skupinami (H(2)= 24,966;p<,05) vyjadrené v Cohenovom d dosiahla veľkosť rozdielu veľmi veľký efekt (d=1,162) a v Eta squared (η2=0,252). Výsledky Post Hoc testu pre test MIE, poukazujú na signifikantné rozdiely v dvoch porovnávaných dvojiciach: medzi pacientmi so schizofréniou a pacientmi s BAP (Z=3,49; p<,01) a taktiež medzi pacientmi so schizofréniou a zdravými kontrolami (Z=4,86; p<,01). Rozdiel medzi zdravou populáciou a pacientmi s BAP však nedosiahol v tomto prípade významné hodnoty (p=0,19).

Významné rozdiely sa nám nepreukázali medzi závažnosťou ochorenia a schopnosťou rozpoznávať mentálne stavy (p=0,359). Preukázal sa nám silný pozitívny vzťah medzi schopnosťou rozpoznávať mentálne stavy a schopnosťou rozpoznávať emócie (r=0,36; p<0,01).

V kognitívnom štýle sme zaznamenali rozdiely medzi klinickou a zdravou populáciou (U=558; p<,01). Na základe priemerných poradí evidujeme vyššie hodnoty v skupine kontrol. Medzi skupinami boli zistené signifikantné rozdiely (H(2)= 18,125;p<0,01) s veľkým efektom efektom (d=0,928; η2=0,177). Post Hoc analýza identifikovala rozdiely v dvoch porovnávaných skupinách, na hladine významnosti (Z=-2,07; p<,05) medzi kontrolami a pacientmi s BAP a na hladine významnosti (Z=4,25; p<,01) opäť medzi kontrolami a pacientmi so schizofréniou. Rozdiel medzi pacientmi so schizofréniou a pacientmi s BAP však nedosiahol signifikantné hodnoty (p=0,069).

Vzťahy medzi symptomatikou ochorenia a skúmanými doménami sa nám nepreukázali ako významné, s výnimkou preukázaného významného vzťahu medzi závažnosťou pozitívnych symptómov a kognitívnym štýlom.

Záver

Skúmaním viacerých spomínaných faktorov medzi rôznymi diagnózami, by sme chceli prispieť k hľadaniu spoločného mechanizmu medzi bipolárnou poruchou a schizofréniou a zisteniam, ktoré by vedeli byť užitočné pri práci s pacientmi v klinickom prostredí. Takisto je dôležité aby výsledky výskumu tvorili základ pre možné intervenčné programy.

Poďakovanie

Prezentovaný výskum je súčasťou dizertačnej práce. Týmto by som sa veľmi rada poďakovala doc.Mgr. Radovanovi Šiklovi, Ph.D. za cenné rady pomoc v odbornom nasmerovaní práce a nesmierne empatický prístup.

Literatúra

[1] S. Baron-Cohen: *Evolution of Theory of Mind?* 1.edition. Oxford: Oxford University press, 1999.

[2] E. Bora, S. Vahip, A.S. Gonul, F. Akdeniz, M. Alkan, M. Ogut, A. Eryavuz: Evidence for Theory of Mind Deficits in Euthymic Patients with Bipolar Disorder. In: *Acta Psychiatrica Scandinavica,* 112(2), 2005: 110-116.

[3] G.Csukly et al. : Emotion-Related Visual Mismatch Responses in Schizophrenia: Impairments and Correlation with Emotion Recognition. In: *Plos one*, 8(10), 2013.

[4] K.S. Kerr et al. : Theory of Mind Deficits in Bipolar Affective Disorder In: *Journal of Affective Disorders,* 73(3), 2003.

[5] C.G. Kohler, J.B. Walker, E.A. Martin, K.M. Healey, P.J. Moberg: Facial Emotion Perception in Schizophrenia: A Meta-analytic Review. In: *Schizophrenia Bulletin*, 36(5), 2010: 1009–1019.

[6] M. Mazza et al.: Selective Impairments of Theory of Mind in People with Schizophrenia. In: *Schizophrenia research,* 47(2), 2001: 299-308.

Systém rozpoznávání akcí integrující detekci objektů a jejich pohybů

Anastasia Ostapenko, Michal Vavrečka

Český institut informatiky, robotiky a kybernetiky, ČVUT
Jugoslávských partyzánů 1580/3, 160 00 Dejvice, Praha
Email: michal.vavrecka@cvut.cz

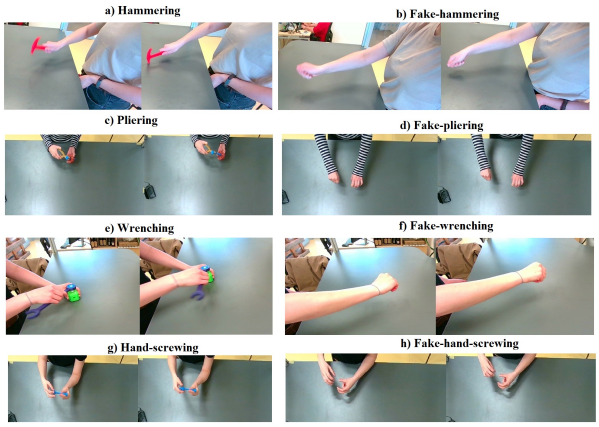
1 Úvod

Vyvinuli jsme nový systém rozpoznávání akcí založený na integraci informací ze dvou samostatných modulů. První modul je zodpovědný za detekci a kategorizaci pohybu. Druhý modul rozpoznává objekty a jejich polohu ve scéně. Informace z obou modulů jsou integrovány do třetího modulu, který rozpoznává akce na základě detekce objektů a jejich pohybu. Oproti tradičním systémům založeným na detekci pohybu jsme schopni rozpoznávat falešné akce (gesta) v případech, kdy ve scéně nejsou přítomny žádné kontextové objekty. Navíc detekujeme průměrnou rychlost kontextových objektů, abychom zvýšili přesnost rozpoznaných akcí. Pro otestování systému jsme vytvořili dataset osmi typů akcí, které zahrnují montážní akce s nástroji a také odpovídající falešné akce (gesta), které mají podobný pohyb, ale kde nejsou použity žádné nástroje. Náš systém rozpoznávání dosahuje 95,21 % přesnosti ve srovnání s 85,52 % u systému založeného pouze na detekci pohybu. Prokázali jsme, že kombinací výstupů ze dvou různých detektorů lze zlepšit celkové výsledky úlohy rozpoznávání akcí. Náš systém rozpoznávání lze použít v úlohách v reálném světě k odlišení skutečných akcí od gest. Systém je také vhodný pro detekci akcí ve spojitém záznamu, kde dokáže lépe identifikovat falešné pozitivní detekce akcí.

2 Metody

V naší práci používáme osm montážních úkonů: zatloukání, falešné zatloukání, ohýbání, falešné ohýbání, šroubování (šroubování matice klíčem), falešné šroubování, ruční šroubování (šroubování matice rukou), falešné ruční šroubování. Všechny typy akcí jsou zobrazeny na obrázku 1. Rozdíl mezi ”normálnímiä ”falešnýmiäkcemi spočívá v tom, že falešné akce nevyžadují nástroj. Pohyb těchto dvou typů akcí je však velmi podobný, což je pro běžné systémy rozpoznávání akcí matoucí, protože jsou navrženy tak, aby rozpoznávaly pohyb akce, a ne okolních objektů. Zde pomáhá zejména přidání druhého modulu, který je zodpovědný za rozpoznávání objektů.

Hlavní myšlenka našeho přístupu vychází z výše uvedené skutečnosti. Předpokládáme, že přiřazení úkonu k nástroji může zvýšit celkovou přesnost v případech, kdy typ úkonu není dobře identifikovatelný, např. zatloukání a falešné zatloukání. K ověření tohoto předpokladu jsme vytvořili následující architekturu.



Obr. 1: Ukázka typů akci

3 Architektura

Nejprve je natrénován modul pro rozpoznávání objektů (Bolya et al.,2019) na datasetu obsahujícím obrázky všech nástrojů potřebných pro vybrané činnosti, např. kladivo, destička atd. Poté jsou do natrénovaného modulu vloženy videosnímky akcí. Na každém snímku jsou detekovány objekty, jejich poloha v obraze a průměrná rychlost objektu v sekvenci. Modul pro rozpoznávání pohybu (Zhang et al., 2019) je trénován přímo na datasetu akcí. Po natrénování je každé video rozděleno na snímky a poté posláno do natrénovaného modulu. Výsledkem je předpovězená akce a pravděpodobnost této předpovědi. Výstupy z obou modulů se spojí a uloží jako dataset, který se pak použije k trénování třetího modulu.

Poslední a nejdůležitější částí našeho systému je modul, který přijímá všechna předzpracovaná data jako vstup a jako výstup vytváří konečnou klasifikaci. Pro tento krok jsme se rozhodli použít neuronovou síť, která se skládá z několika vrstev: vstupní vrstvy, dvou skrytých vrstev a výstupní vrstvy. Jako vstup bere data získaná ze dvou předchozích modulů: předpověď akce provedená modulem pohybu,

pravděpodobnost této předpovědi, objekt detekovaný modulem pro rozpoznání objektů, pravděpodobnost této detekce a vzdálenost, o kterou se průměrně pohyboval objekt mezi dvěma sousedními snímky. Na výstupu je výsledná klasifikace videa.

4 Výsledky

Náš systém jsme testovali na datasetu osmi typů akcí zastoupené přibližně 200 videi na akci. Výsledná přesnost rozpoznání pohybu je uvedena v tabulce 1, přesnost rozpoznání objektů je a tabulce 2. Během trénování detekce pohybu jsme experimentovali s počtem epoch a rychlostí učení. Nejlepšího výsledku jsme dosáhli, když jsme modul trénovali 120 epoch s rychlostí učení 0,01. Z tabulky je zřejmé, že falešné akce lépe odhalovány než akce vyžadující nástroj. To může být způsobeno tím, že v souborech dat je více falešných akcí. Rozdíl mezi počtem falešných akcí a akcí s nástrojem je však menší než 10 %. Akce s nástroji jsou často zaměňovány s odpovídajícími falešnými akcemi, protože jejich pohyb je podobný. Modul pro detekci objektů dosahuje výborných výsledků pro většinu tříd. Činnost tohoto modulu je velmi důležitá, protože přímo ovlivňují výsledky multimodálního integrátoru. Výsledky pro multimodální integrátor jsou v tabulce 3. Je zřejmé, že použití detekce objektu společně s detekcí pohybu zvyšuje přesnost rozpoznání akcí.

	Správnost [%]	Sensitivita [%]	Přesnost [%]
Hammering	87.50	0.0	N/A
Fake hammering	84.58	96.67	44.63
Pliering	82.08	23.33	25.93
Fake pliering	79.58	90.0	36.99
Hand-screwing	86.25	36.67	44.0
Fake hand-screwing	84.17	0.0	0.0
Wrenching	88.75	33.33	58.82
Fake wrenching	91.25	56.67	68.0
Total	85.52	42.08	39.76

Tab. 1: Tabulka přesnosti rozpoznání pro modul detekující pohyb

5 Diskuze

Jak bylo uvedeno dříve, modul nefunguje dobře, pokud je v souboru dat mnoho falešných akcí. K tomu dochází v důsledku vysoké míry podobnosti mezi falešnými akcemi a odpovídajícími akcemi s nástroji. Přidáním druhého modulu, který je zodpovědný za rozpoznávání objektů, se počet správně klasifikovaných videí znatelně zvýší. Sensitivita stoupla ze 42 % na 81 %. . Na druhou stranu sensitivita u některých falešných akcí (falešné zatloukání, falešné ohýbání) mírně klesá ve srovnání s výsledky základního klasifikátoru. Další nevýhodou našeho řešení je skutečnost, že je v současné konfiguraci použitelné pouze pro akce s nástroji.

	Správnost [%]	Sensitivita [%]	Přesnost [%]
Hammering	97.5	100.0	83.33
Pliering	88.33	46.67	53.85
Hand-screwing	91.25	90.0	60.0
Wrenching	96.67	73.33	100.0
Fake actions	90.42	86.67	93.69
Total	92.83	79.33	78.17

Tab. 2: Tabulka přesnosti rozpoznání pro modul detekující objekty

	Správnost [%]	Sensitivita [%]	Přesnost [%]
Hammering	99.58	100.0	96.77
Fake hammering	97.92	93.33	90.32
Pliering	94.58	80.0	77.42
Fake pliering	91.67	83.33	62.5
Hand-screwing	93.75	70.0	77.78
Fake hand-screwing	90.83	56.67	65.38
Wrenching	97.08	83.33	92.59
Fake wrenching	96.25	80.0	88.89
Total	95.21	80.83	81.46

Tab. 3: Tabulka přesnosti rozpoznání pro modul integrující informace o objektech a jejich pohybu

Poděkování

Tato práce byla podpořena projektem INAFYM (CZ.02.1.01/0.0/0.0/16_019/0000766).

6 Reference

Bolya, D., Zhou, C., Xiao, F., Lee, Y. J. (2019). Yolact: Real-time instance segmentation. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9157-9166).

Zhang, C., Zou, Y., Chen, G., Gan, L. (2019, October). PAN: Persistent appearance network with an efficient motion cue for fast action recognition. In Proceedings of the 27th ACM International Conference

Is none treatment for mental health problems better than a controversial one?

Klára Petrovická

FMFI UK, Comenius University in Bratislava
Mlynská dolina, 842 48 Bratislava
Email: petrovicka1@uniba.sk

Abstract

Nowadays, people worldwide face a lack of mental health workers, causing therapies to be out of reach for those who need them. Consequently, the digital applications of telemedicine, which are always instantly available, reveal a unique solution. However, significant barriers are preventing such solutions from integrating into clinical practice. This review paper discusses the advantages and obstacles to artificial intelligence-powered psychotherapy delivered by conversational chatbots. It aims to identify a gap-closing proposal by addressing the current state of (digital) therapeutic rapport and the app developmental process.

1 Introduction

Nowadays, artificial intelligence (AI) is being used in healthcare primarily in a complementary role to the superior health experts for diagnosis and treatment (Fiske, Henningsen, & Buyx, 2019; Sebri, Pizolli, Savioni, & Triberti, 2021). However, there is a recent outburst of research aiming to include AI also in mental health practise. Particularly in the form of mobile therapeutic apps (e.g. chatbots) as a mental health therapeutic tool, which explores patients' symptoms and addresses depression and anxiety over short-message services (Sebri et al., 2021). Importantly, such a device is proposed to be an outstanding tool, changing the mental health field due to its independence from expert human guidance (Fiske et al., 2019). Therefore, the AI chatbots may be superficial to other emerging applications of telemedicine, such as Web-based therapy (Mehrotra et al., 2017).

Given the median of 9 mental health workers (including one psychiatrist) per 100,000 people globally (WHO, 2018), and the up to 2-years long waiting period before the therapy starts in the UK (BMA, 2018), AI text-based conversational agents represent an optimal solution since they can offer anytime-available support (Inkster, Sarda, & Subramanian, 2018). To illustrate the advantage, the participants have checked in with the conversational chatbot on average 12.14 times, and even more, 17.71 times during the 2-week research period (Fitzpatrick, Darcy, & Vierhile, 2017; Ly, Ly, & Andersson, 2017; respectively). Whereas, if patients get even one appointment with a human therapist in two weeks, they will be considered lucky. Therefore, the internet-based AI-supported interventions seem like an obvious answer to the lack of personnel problem. Nevertheless, whether the AI-led therapies may be of equal quality, especially regarding the level of adherence and therapeutic rapport, require further investigation.

2 Adherence - rapport

There is prevailing evidence suggesting that the patient must adhere to the treatment well for a successful intervention. Specifically, to maintain a sufficient adherence level, the patients have to develop a robust working alliance with their therapist (Darcy et al., 2021). Inkster et al. (2018) reported significantly greater average improvement in high users than in low chatbot users demonstrating the importance of quality rapport. Moreover, self-reported psychological wellbeing and perceived stress differed between high chatbot users and participants without any support (those were put on the waiting list for a session with a human psychotherapist).

2.1 Empathy

Additionally, empathy skills presented in the conversation pre-determines a good-quality rapport (Darcy et al., 2021). Problematically, virtual therapeutic agents have been repetitively criticized for lacking empathy (Torous & Hsin, 2018). On the contrary, there is a recent trend proposing that the digital therapeutic rapport do not differ from the one with human therapists, and participants tend to report "feelings of empathy" while a conversation with their chatbot (D'Alfonso et al., 2017; Darcy et al., 2021; Inkster et al., 2018). Darcy et al. (2021) suggested that the conversational agent Woebot reported better engagement rates than in the previous study done by Fitzpatrick et al. (2017) due to the tool's

added empathetic and relational nature. Thus, the lack of empathy among virtual therapists seems outdated, and the application of chatbots over the waiting-list treatment could be growing in interest.

However, the assessment of empathy and therapeutic rapport in chatbots is mainly done by thematic analyses of qualitative data submitted by users/participants in the form of open-ended feedback (see Darcy et al., 2017; Fitzpatrick et al., 2017). Despite the priority-setting work by the James Lind Alliance (Hollis et al., 2018) identifying the therapeutic rapport as a top research priority, not a single study included rapport as a primary outcome of their analyses in a recent review of mobile health apps (Henson et al., 2019). Until the conversational rapport in mental health is explicitly analysed, the research findings are purely estimations with the tendency to keep the AI role in mental health supplementary to human therapists.

3 App development

Furthermore, the digital rapport is not ignored only by academics but also in the apps development process. Torous and Hsin (2018) argue it is due to the app developers' lack of awareness of such needs. Furthermore, researchers criticize that the apps development process only occurs in academic settings, and the scientists involved have limited to no actual patient contact (Arnberg et al., 2014; Torous & Hsin, 2018). Therefore, while the results of the interventions seem impressive on paper, they lack effectiveness in real-world settings. In contrast to Torous and Hsin's (2018) criticism, two recent studies included anonymous active chatbot users as their participants (see Darcy et al., 2021; Inkster et al., 2018). Nevertheless, these studies also face several limitations, including a lack of variation in their sample size.

Moreover, another alarming limitation of these studies cannot go unmentioned; there are financial relations between the researchers conducting the study (see Darcy et al., 2021; Fitzpatrick et al., 2017; Inkster et al., 2018) and the respective chatbot companies (Woebot, Wysa). Notably, Ly's et al. (2017) is the only reviewed study analyzing a therapeutic chatbot, yet not carrying said conflict of interest. Henceforth, further studies with independent investigators are certainly needed.

3.1 Barriers to clinical practice

Additionally, next to the lack of proper interventions development, mental health care providers also lack training and guidance to integrate AI interventions into clinical practice (Fiske et al., 2019). Significantly, AI will not have any powerful effect without a structural change in the health system and the shift in the health care providers' attitude (Dowie & Kaltoft, 2018). However, mental health professionals do not have a very positive tendency towards AI as psychotherapists (Sebri et al., 2021). It has been pointed out that this perceiving gap between the rapid development of AI-supported chatbots and the real-world clinical practice is due to the ethical issues related to the application of said technology (Fiske et al., 2019; McDonnell, Rooney, & Flood, 2013).

To conclude, there is a persistent matter of including trained mental health carers in the app development and the need for interdisciplinary and independent investigation teams (Darcy et al., 2021; Fitzpatrick et al., 2017). While said remark has been highlighted in the recent guideline for clinical app evaluation published by the American Psychiatric Association (APA, 2017), it is not being actively addressed. In addition, appropriate guidelines for care providers are needed to target the clinical practice successful application.

4 Summary

Nowadays, the burden of mental health disorders is even greater due to the Covid-19 pandemic, which presented unintended barriers to seeking treatment and increased the number of people suffering mental-health-related discomfort (Bueno-Notivol et al., 2021). Internet-based interventions provide the apparent solution. While there is still little known regarding the direct long-term effects of AI-supported intervention, it has shown to be a better intervention for treating mild cases of mental health discomfort than not engaging in any treatment. However, there is a unique challenge of assessing the quality of the digital rapport and the ability to perform empathy by the virtual agents. Consequently, the apps' development needs to shift towards these targets to overcome the gap to clinical practice so the AI therapists may fulfill their potential.

References

American Psychiatric Association [APA] (2017). *App Evaluation Model.* Retrieved from: https://www.psychiatry.org/psychiatrists/practice/mental-health-apps/app-evaluation-model

Arnberg, F.K., Linton, S.J., Hultcrantz, M., Heintz, E., & Jonsson, U. (2014). Internet-delivered psychological treatments for mood and anxiety disorders: A systematic review of their efficacy, safety, and cost-effectiveness. *PLOS One, 9*(5), e98118. doi: 10.1371/journal.pone.0098118.

British Medical Association [BMA] (2018). *The impact of COVID-19 on mental health in England; Supporting services to go beyond parity of esteem*. Retrieved from: https://www.bma.org.uk/media/2750/bma-the-impact-of-covid-19-on-mental-health-in england.pdf

Bueno-Notivol, J., Gracia-García, P., Olaya, B., Lasheras, I., López-Antón, R., & Santabárbara, J. (2021). Prevalence of depression during the COVID-19 outbreak: A meta-analysis of community-based studies. *International Journal of Clinical and Health Psychology, 21*(1), 100196. doi: 10.1016/j.ijchp.2020.07.007.

D'Alfonso, S., Santesteban-Echarri, O., Rice, S., Wadley, G., Lederman, R., Miles, C., …, & Alvarez-Jimenez, M. (2017). Artificial Intelligence-assisted online social therapy for youth mental health. *Frontiers Psychology, 8*, 796. doi: 10.3389/fpsyg.2017.00796.

Darcy, A., Daniels, J., Salinger, D., Wicks, P., & Robinson, A. (2021). Evidence of human-level bonds established with a digital conversational agent: Cross-sectional, retrospective observational study. *Journal of Medical Internet Research Formative Research, 5*(5), e27868. doi: 10.2196/27868.

Dowie, J., & Kaltoft, M.K. (2018). The future of health is self-production and co-creation based on apomediative decision support. *Medical Sciences, 6*(3), 66. doi: 10.3390/medsci6030066.

Fiske, A., Henningsen, P., & Buyx, A. (2019). Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of Medical Internet Research, 21*(5), e13216. doi: 10.2196/13216.

Fitzpatrick, K.K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *Journal of Medical Internet Research Mental Health, 4*(2), e19. doi: 10.2196/mental.7785.

Henson, P., Wisniewski, H., Hollis, C., Keshavan, M., & Torous, J. (2019). Digital mental health apps and the therapeutic alliance: Initial review. *British Journal of Psychiatry Open, 5*(1), e15. doi: 10.1192/bjo.2018.86.

Hollis, C., Sampson, S., Simons, L., Davies, E.B., Churchill, R., Betton, V., ..., & Tomlin, A. (2018). Identifying research priorities for digital technology in mental health care: Results of the James Lind Alliance Priority Setting Partnership. *The Lancet Psychiatry, 5*(10), 845-854. doi: 10.1016/S2215-0366(18)30296-7.

Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental wellbeing: Real-world data evaluation mixed methods study. *Journal of Medical Internet research mHealth and uHealth, 6*(11), e12106. doi: 10.2196/12106.

Ly, K.H., Ly, A.-M., & Andersson, G. (2017). A fully automated conversational agent for promoting mental wellbeing: A pilot RCT using mixed methods. *Internet Intervention, 10*, 39-46. doi: 10.1016/j.invent.2017.10.002.

McDonnell, D., Rooney, B., & Flood, C. (2013). Attitudes to computerized psychotherapy: A survey of psychotherapists In: *Cyberpsychology and New Media.* Psychology Press Taylor & Francis Group, 190–202.

Mehrotra, S., Kumar, S., Sudhir, P., Rao, G.N., Thirthalli, J., & Gandotra, A. (2017). Unguided mental health self-help apps: Reflections on challenges through a clinician's lens. *Indian Journal of Psychological Medicine, 39*(5), 707-711 doi: 10.4103/IJPSYM.IJPSYM_151_17.q

Sebri, V., Pizzoli, S.F.M., Savioni, L., & Triberti, S. (2021). Artificial Intelligence in mental health: Professionals; attitudes towards AI as a psychotherapist. *Annual Review of CyberTherapy and Telemedicine, 18*.

Torous, J., & Hsin, H (2018). Empowering the digital therapeutic relationship: Virtual clinics for digital health interventions. *Nature Partner Journals Digital Medicine, 1*, 16. doi: 10.1038/s41746-018-0028-2.

World Health Organization [WHO] (2018). *Mental Health Atlas 2017*. Retrieved from: https://www.who.int/publications/i/item/9789241514019

Učení se reprezentace peripersonálního prostoru pomocí neuronových sítí

Zdeněk Straka, Matěj Hoffmann

Katedra kybernetiky, Fakulta elektrotechnická, ČVUT v Praze
Karlovo náměstí 13, Praha
Email: straka.zdenek@fel.cvut.cz

Abstrakt

Schopnost predikovat kontakt mezi tělem a okolním světem je klíčová pro přežití. Předpokládá se, že klíčovou roli v této schopnosti hrají multisenzorní vizuo-taktilní či audio-taktilní neurony, které reagují na stimuly v blízkém okolí těla – v tzv. peripersonálním prostoru. V příspěvku představíme, jak lze mechanismy těchto reprezentací peripersonálního prostoru zkoumat pomocí modelů umělých neuronových sítí. Představíme hlubokou neuronovou síť pro predikci budoucích vizuo-taktilních vstupů či model peripersonálního prostoru kombinující Omezený Boltzmannův stroj a dopřednou neuronovou síť.

1 Úvod

Pro bezpečnou interakci člověka s jeho okolím je klíčová schopnost předvídat nárazy hrozící především od pohybujících se předmětů. Předpokládá se, že pro tyto predikce hrají zásadní roli fronto-parietální oblasti mozku, které jsou z velké části tvořeny multisenzorními neurony (viz např. (Cléry a spol., 2015)). Tyto neurony reagují na stimuly v blízkém okolí těla – v tzv. Peri-Personálním Prostoru (PPP). Přestože existuje celá řada empirických pozorování souvisejících s reprezentacemi PPP, pochopení mechanismů učení a fungování těchto reprezentací stále chybí. Pro lepší pochopení těchto mechanismů jsme navrhli modely založené na umělých neuronových sítích a jejich vlastnosti porovnali s empirickými poznatky o těchto reprezentacích.

Empirické poznatky o reprezentacích peripersonálního prostoru jsou získávány v behaviorálních a neurofyziologických experimentech. Na základě těchto experimentů se ukázalo, že tyto reprezentace mají například tyto vlastnosti:

- rozděluje prostor na blízký a vzdálený (Serino, 2019),
- velikost PPP roste se zvyšující se rychlostí objektu blížícího se k tělu (Fogassi a spol., 1996; Noel a spol., 2018),
- reprezentace PPP je plastická (např. (Galli a spol., 2015)).

Vlastnosti naučených modelů reprezentací PPP jsou porovnány s uvedenými empiricky pozorovanými vlastnostmi. Tímto způsobem je možné vyloučit některé hypotetické mechanismy. Zajímavé vlastnosti modelů mohou navíc být experimentálně otestovány v nových empirických studích.

V současné době existuje velmi omezený počet neurálních modelů PPP. Magosso a spol. (2010) představili biologicky motivovanou neuronovou síť modelující reprezentaci PPP. Tato síť se skládá z multisenzorní populace neuronů, která je propojena s unimodálními populacemi vizuálních a taktilních neuronů. Varianty této sítě umí např. modelovat rozšíření PPP při zvýšení rychlosti stimulu (Noel a spol., 2018). Bertoni a spol. (2021) navrhli neuronovou síť, která se učí statistické závislosti mezi vizuálními, taktilními a proprioceptivními vstupy. Tato reprezentace PPP se dokázala naučit přikotvení receptivních polí k částem těla. Přestože tyto modely jistě přispěly k porozumění mechanismů PPP, stále dokáží vysvětlit pouze značně omezený počet známých empirických vlastností PPP.

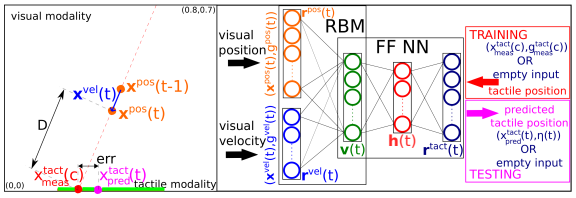
Ani jedna z uvedených sítí nemodeluje mechanismus reprezentace PPP jako formu predikce možného zasažení těla objekty z okolí. To je však populární kvalitativní vysvětlení vzniku reprezentací PPP ve výzkumné komunitě (viz např. (Dijkerman a Medendorp, 2021)). Proto jsme navrhli modely PPP, které se učí predikovat zásah těla.

2 Vizuo-taktilní prediktivní modely reprezentací PPP

Představíme dva modely pro vizuo-taktilní predikci – na základě vizuálního stimulu je predikován taktilní stimul. První model má na vstupu předzpracovaná data – 2D pozici stimulu a vektor rychlosti. Druhý model pracuje přímo s obrazovými daty 3D prostředí.

2.1 Model pro předzpracovaná vstupní data

Neuronová síť se skládá z Omezeného Boltzmannova stroje a dopředné neuronové sítě (Straka a Hoffmann, 2017). Architektura sítě je na Obr. 1. První část sítě se (bez učitele) učí reprezentovat pozici stimulu a vektor rychlosti. Druhá část sítě se učí predikovat budoucí po-



Obr. 1: Scénář a architektura. VLEVO: 2D experimentální scénář. Trajektorie stimulu je znázorněna oranžovou barvou. Taktilní modalita má na obrázku zelenou barvu. **VPRAVO:** Architektura neuronové sítě a znázornění procesu trénování a testování (predikování).

zici taktilního stimulu. Vstupní i výstupní neurální populace umožňují zakódovat také neurčitost stimulu. Pro rychlejší stimuly došlo k rozšíření PPP. Použitý mechanismus učení byl tedy, i přes svou jednoduchost, dostačující pro dosažení této důležité vlastnosti z empirických studií reprezentací PPP.

2.2 Model pro nepředzpracovaná vstupní data

Tato neuronová síť vznikla rozšířením sítě PreCNet (Straka a spol., 2020) pro predikci příštího snímku videa – sekvence snímků je dána na vstup a úkolem sítě je predikovat následující snímek. PreCNet je hluboká síť založena na myšlence prediktivního kódování. Aby bylo možné síť použít pro reprezentaci PPP, byla síť rozšířena o taktilní modalitu a rozdílnou váhu falešně pozitivní a negativní chyby (Pitoňák, 2020). V současné době stále probíhá analýza vlastností reprezentace PPP, která byla naučena touto sítí.

3 Závěr

Reprezentace blízkého okolí těla je klíčová pro přežití. Pochopení vzniku a fungování mechnismů těchto reprezentací stále chybí. Ve vědecké komunitě je populární hypotéza, že tato reprezentace je spojena s predikováním možného kontaktu těla s předměty z okolí. Proto jsme navrhli dva typy modelů neuronových sítí, které tuto hypotézu zkoumají. Podařilo se nám ukázat, že prediktivní mechanismus učení je postačující pro dosažení některých klíčových vlastností pozorovaných v empirických studiích PPP. Pro budoucí výzkum považujeme za důležité na základě vlastností našich modelů navrhnout empirické experimenty, které mohou posílit či vyvrátit správnost našich modelů.

Poděkování

Tento příspěvek vznikl za finanční podpory Grantové agentury České republiky (GAČR), projekt č. 20-24186X.

Reference

Bertoni, T., Magosso, E. a Serino, A. (2021). From statistical regularities in multisensory inputs to peripersonal space representation and body ownership: Insights from a neural network model. *European Journal of Neuroscience*, 53(2):611–636.

Cléry, J., Guipponi, O., Wardak, C. a Hamed, S. B. (2015). Neuronal bases of peripersonal and extrapersonal spaces, their plasticity and their dynamics: knowns and unknowns. *Neuropsychologia*, 70:313–326.

Dijkerman, H. a Medendorp, W. (2021). Visuotactile predictive mechanisms of peripersonal space. V *The world at our fingertips: a multidisciplinary exploration of peripersonal space*, str. 81–100. Oxford University Press.

Fogassi, L., Gallese, V., Fadiga, L., Luppino, G., Matelli, M. a Rizzolatti, G. (1996). Coding of peripersonal space in inferior premotor cortex (area F4). *Journal of neurophysiology*, 76(1):141–157.

Galli, G., Noel, J. P., Canzoneri, E., Blanke, O. a Serino, A. (2015). The wheelchair as a full-body tool extending the peripersonal space. *Frontiers in psychology*, 6:639.

Magosso, E., Zavaglia, M., Serino, A., Di Pellegrino, G. a Ursino, M. (2010). Visuotactile representation of peripersonal space: a neural network study. *Neural computation*, 22(1):190–243.

Noel, J.-P., Blanke, O., Magosso, E. a Serino, A. (2018). Neural adaptation accounts for the dynamic resizing of peripersonal space: evidence from a psychophysical-computational approach. *Journal of neurophysiology*, 119(6):2307–2333.

Pitoňák, A. (2020). Aplikace prediktivního kódování pro vizuo-taktilní integraci. Bakalářská práce, České vysoké učení technické v Praze.

Serino, A. (2019). Peripersonal space (pps) as a multisensory interface between the individual and the environment, defining the space of the self. *Neuroscience & Biobehavioral Reviews*, 99:138–159.

Straka, Z. a Hoffmann, M. (2017). Learning a peripersonal space representation as a visuo-tactile prediction task. V *International conference on artificial neural networks*, str. 101–109. Springer.

Straka, Z., Svoboda, T. a Hoffmann, M. (2020). PreCNet: next frame video prediction based on predictive coding. *arXiv preprint arXiv:2004.14878*.

Potenciál augmentované reality během letu

Šašinka, Čeněk[1], Stachoň, Zdeněk[2], Chmelařová, Kateřina[3], Jochecová, Kateřina[3]

Katedra informačních studií a knihovnctví[1],
Geografický ústav[2]
Psychologický ústav[3]

Masarykova Univerzita

cenek.sasinka@mail.muni.cz, zstachon@geogr.muni.cz, 393900@mail.muni.cz, katerina.jochecova@mail.muni.cz

Abstrakt

Tento příspěvek se věnuje efektivitě získávání informací během letu z brýlí s augmentovanou realitou a z běžného telefonu. Podle Swellera (2010) je míra interaktivity podnětů podstatná pro míru kognitivní zátěže. Pokud lze každý podnět zpracovávat samostatně, zátěž je mírná, ale pokud na sebe podněty navazují, zvyšuje se. Pilotování jakéhokoliv prostředku je tedy kognitivně velmi zatěžující, protože pilot během letu musí získávat velké množství informací z prostředí a adekvátně reagovat, jinak ohrozí (nejen) svůj život.

Rozšířená (augmentovaná) realita (AR) se využívá v širokém spektru odvětví. Kromě výuky se také často objevuje při výzkumu nepozornosti řidičů jako alternativa k mobilním telefonům (He et al., 2018; Sawyeret al., 2014). Z těchto výzkumů vyplývá, že během užívání brýlí s AR je soustředěnost řidičů vyšší, jejich rekce rychlejší a jízda přesnější. Obecně AR v dané situaci snižuje kognitivní zátěž účastníků. Můžeme tedy předpokládat, že i během letu budou brýle vhodnější.

Vzhledem k nebezpečnosti prvního testování využití AR během reálného letu došlo k vytvoření simulace, při níž byla použita nahrávka z reálného letu (obr. 1). Vytvořili jsme realistickou aplikaci pro brýle s augmentovanou realitou (VUZIX m100), která ukazovala výšku a rychlost letu paraglidu. Účastníci poté absolvovali dva lety, pokaždé s jiným zařízením (brýle s AR nebo telefon připevněný na stehně, jak je zvykem u skutečných paraglidistů, obr. 2). Během letu účastníci kromě hlášení údajů ze zařízení odpovídali na otázky ohledně prostředí, které sloužily ke zvyšování kognitivní zátěže. Tento design odpovídá realitě paraglidingu a lze tedy očekávat, že v reálném světě budou piloti také hodnotit svoji kognitivní zátěž nižší při užití AR a objektivní chování tomu bude odpovídat.

Obr. 1. Ukázka z nahrávky letu.

Obr. 2. Ukázka letu s využitím brýlí s AR

Literatura

[1] He, J., McCarley, J.S., Crager, K., Jadliwala M., Hua, L., & Huang, S. (2018). Does wearable device bring distraction closer to drivers? Comparing smartphones and Google Glass. *Applied Ergonomics*, 70, 156–166. https://doi.org/10.1016/j.apergo.2018.02.022

[2] Sawyer, B. D., Finomore, V., Calvo, A., & Hancock, P. A. (2014). Google Glass: A Driver Distraction Cause or Cure?. *Human Factors and Ergonomics Society*, 56(7), 1307–1321. https://doi.org/10.1016/j.apergo.2018.02.022

[3] Sweller, J. (2010). Element Interactivity and Intrinsic, Extraneous, and Germane Cognitive Load. *Educational Psychology Review*, 22, 123–138. https://doi.org/10.1007/s10648-010-9128-5

Kolaborativní imerzivní virtuální realita v praxi a výzkumu

Čeněk Šašinka [1], Jiří Chmelík [2], Alžběta Šašinková [1] a Zdeněk Stachoň [1]

Katedra informačních studií aknihovnictví [1¨]
Katedra vizuální informatiky [2]
Masarykova univerzita

cenek.sasinka@mail.muni.cz, jchmelik@mail.muni.cz, asasinkova@mail.muni.cz, zstachon@geogr.muni.cz

Abstrakt

Příspěvek je zaměřen na dvě hlavní témata. Prvním z nich je potenciál a omezení využívání kolaborativní imerzivní virtuální reality v praxi, mj. ve vzdělávání a výzkumu. Druhé se zaměřuje na způsob výzkumu dotčených témat v kolaborativní imerzivní virtuální realitě. Výzkumný tým Masarykovy univerzity se tématem potenciálu kolaborativní imerzivní virtuální reality zabývá několik let. Šašinka et al. (2018) publikovali empirickou kvalitativní studii, která prozkoumala možnosti a omezení využití kolaborativního imerzivního virtuálního prostředí ve vzdělávání, specificky v kontextu výuky geografie (výškopisu). Výsledky studie naznačily oblasti, ve kterých imerzivní virtuální prostředí ukazují největší potenciál. Pro účely zmíněného výzkumu vzniklo i unikátní experimentální softwarové řešení. Jochecová et al. (2022) téma dále rozvíjeli a zkoumali názor doménových expertů (učitelů zeměpisu) na možnosti a limity využití kolaborativní imerzivní virtuální reality ve výuce a rovněž na způsob její implementace. Tato studie již probíhala na nové platformě eDIVE (eDucation in Immersive Virtual Environments), která vznikla jako součást projektu aplikovaného výzkumu, ve kterém jsou vytvářeny výukové scénáře jak geografii, tak angličtinu. Návrh aplikovaného projektu, širší koncepce využití virtuální reality ve formálním i neformálním vzdělávání (mj. se zapojením knihoven) a stejně jako koncepce platformy eDIVE byla připravena v interdisciplinární spolupráci týmem výzkumníků (Šašinka, Chmelík, Šašinková, Stachoň) na Katedře ionformačních studií a knihovnictví a Katedře vizuální informatiky Masarykovi univerzity. Platforma eDIVE byla následně využita přímo v celosemestrálním kurzu angličtiny pro neanglické obory na Filozofické fakultě Masarykovy univerzity, který sloužil zároveň pro ověření efektivity výuky ve virtuální realitě. Cílem příspěvku je rovněž diskutovat možnosti výzkumu a analýzy dat získávaných ve virtuální realitě.

Obr. 1. Ukázka výuky angličtiny v kolaborativním prostředí eDIVE

Poděkování

Tato publikace je podpořena projektem TL03000346 „Vzdělávání v kolaborativní imerzivní virtuální realitě“ (EduInCIVE) finacovaným Technologickou agenturou ČR v programu ÉTA

Literatúra

[1] Šašinka, Č., Stachoň, Z., Sedlák, M., Chmelík, J., Herman, L., Kubíček, P., Šašinková, A., ... & Juřík, V. (2018). Collaborative immersive virtual environments for education in geography. *ISPRS International Journal of Geo-Information*, 8(1), 3.

[2] Jochecová, K., Černý, M., Stachoň, Z., Švedová, H., Káčová, N., Chmelík, J., ... & Šašinka, Č. (2022). Geography Education in a Collaborative Virtual Environment: A Qualitative Study on Geography Teachers. *ISPRS International Journal of Geo-Information, 11(3),* 180.

Inkrementální imitační učení pomocí variačního autoenkodéru

Gabriela Šejnová, Karla Štěpánová

Český institut informatiky, robotiky a kybernetiky, ČVUT
Jugoslávských partyzánů 1580/3, 160 00 Dejvice, Praha
Email: gabriela.sejnova@cvut.cz

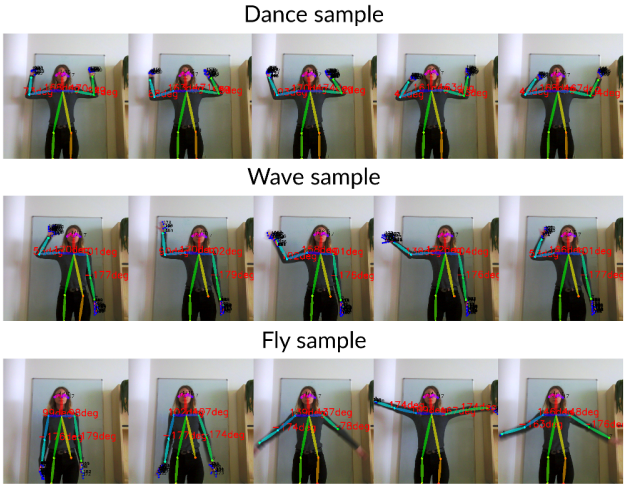
Abstrakt

Variační Autoenkodér (VAE) je generativní model založený na umělých neuronových sítích, který je schopný mapovat vstupní data do spojitého latentního prostoru na základě podobnosti. Ačkoliv VAE nachází uplatnění v nejrůznějších oblastech strojového učení, jeho funkčnost je zpravidla podmíněná velkým množstvím trénovacích dat. V tomto příspěvku se zabýváme využitím VAE v robotické úloze, a to pro inkrementální učení různých typů akcí demonstrovaných člověkem. Výhodou modelu v této úloze je možnost generovat příklady již dříve naučených akcí a předcházet tak jejich zapomínání během dalšího učení. Protože je však příprava trénovacích příkladů v tomto případě časově náročná, ověřujeme, jaký minimální počet demonstrací postačuje k úspěšnému naučení dané úlohy. Průběh i výsledky učení ukazujeme a porovnáváme na humanoidním robotovi Pepper.

1 Úvod

Během posledních let se variační autoenkódery (Kingma a Welling, 2013) zařadily mezi nejčastěji používané generativní modely z důvodu širokého spektra jejich využití. Lze je mj. využít pro rekonstrukci a generování obrazových dat (Liu a spol., 2017) nebo pro rozpoznávání a reprodukci akcí (Petrovich a spol., 2021). Jejich univerzálnost (schopnost generovat i klasifikovat data zároveň) by se mohla uplatnit i v úloze imitačního učení robotických akcí v reálném světě - zde však narážíme na problém omezeného množství trénovacích příkladů, které je schopen člověk jako demonstrátor poskytnout. Navíc zde obvykle od modelu očekáváme schopnost doučovat se postupně nové úlohy, aniž by zapomínal akce již dříve naučené. Dalším problémem, zvláště důležitým pro modely VAE, je, jak hodnotit kvalitu výsledků výhradně na základě vizuální domény.

V tomto příspěvku se zabýváme adaptací vybraného variačního autoenkodéru ACTOR Petrovich a spol. (2021) pro úlohu imitačního učení. Akce reprezentujeme jako sekvenci úhlů kloubů rukou získaných pomocí předtrénovaného modelu OpenPose (Cao a spol., 2017) z obrazových dat. Ověřujeme, kolik demonstrací je třeba pro úspěšně natrénování modelu ACTOR a jaká strategie učení zabraňuje zapomínání dříve naučených typů akcí. Také navrhujeme evaluaci vygenerovanách akcí pomocí statistických parametrů vytvořených nad trénovacím datasetem. Výsledky učení demonstrujeme na humanoidním robotovi Pepper.



Obr. 1: Ukázka tří námi použitých akcí pro imitační učení: *tancovat* (nahoře), *mávat* (uprostřed) a *létat* (dole). Pozice a úhly kloubů rukou jsme získali z RGB obrazu pomocí knihovny OpenPose (Cao a spol., 2017).

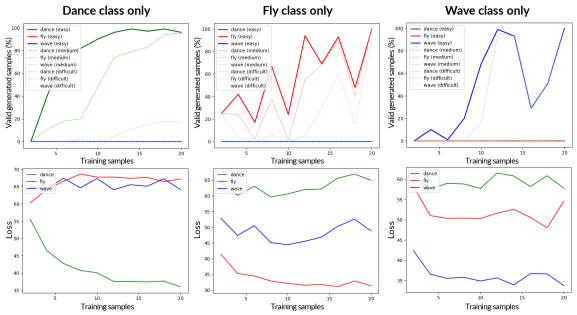
2 Experiment

2.1 Tvorba datasetu

Pro náš experiment jsme zvolily tři jednoduché typy akcí, které se má robot na základě demonstrace naučit: akce *tancovat*, *mávat* a *létat*. Pro porovnání různých přístupů jsme nahrály trénovací dataset sestávající z 20 ukázek pro každou akci, tedy 60 sekvencí (každá o délce 27-33 časových kroků). Pro evaluaci jsme navíc připravily 10 příkladů pro každou třídu (30 sekvencí).

2.2 Trénovací strategie

Pro trénování jsme použily variační autoenkodér ACTOR (Petrovich a spol., 2021) využívající neuronové sítě zvané Transformery určené pro zpracování sekvenčních dat.



Obr. 2: Přehled výsledků pro VAE modely trénované pouze na akci *tancovat*, *létat* nebo *mávat*. Horní grafy ukazují percentuální úspěšnost dle počtu správně vygenerovaných akcí. Jednotlivé typy křivek znázorňují různě přísná kriteria pro evaluaci. Spodní grafy ukazují celkovou chybu vypočtenou na 10 testovacích příkladech.

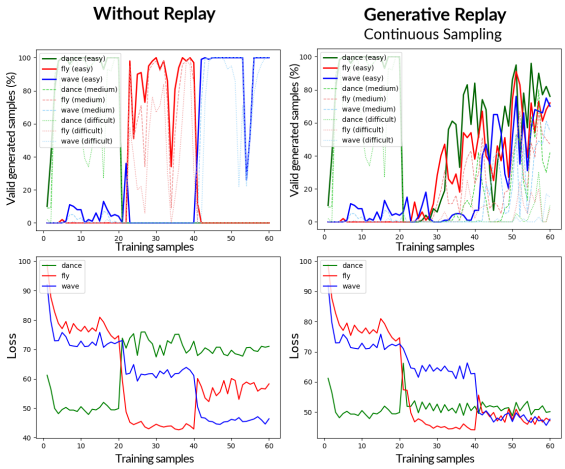
V prvním setu experimentů jsme trénovaly pro každý typ akce zvlášť jeden model a sledovali jsme, po kolika příkladech se model danou akci naučí generovat. V druhé části jsme trénovaly jeden model pro všechny typy akcí prezentovaných postupně za sebou. Zde je však třeba přidat kompenzační mechanismus zabraňující zapomínání. Proto vždy při změně typu akce během učení uložíme aktuální váhy modelu. Ty pak používáme během dalšího učení pro generování příkladů daného typu akce - tato data průběžně přimícháváme mezi nová data, čímž si model neustále "opakuje" i dříve naučené akce.

3 Výsledky

Na Obr. 2 ukazujeme výsledky pro VAE modely specializované vždy na jeden typ akce. Vždy po N (osa x) předložených příkladech model vygeneruje 100 náhodných vzorků, u nichž ověřujeme jejich "správnost", tzn. jestli jejich statistické parametry (např. střední hodnota pozice kloubů napříč celou akcí) odpovídají průměrným statistickým parametrům celého trénovacího datasetu. Na základě těchto parametrů jsme stanovily tři prahové hodnoty dle striktnosti. Horní grafy na Obr. 2 tedy ukazují procento vzorků vyhodnocených jako správné dle těchto tří prahů. Na Obr. 3 porovnáváme výsledky učení před (vlevo) a po (vpravo) přidání mechanismu proti zapomínání předchozích akcí. Horní grafy opět ukazují percentuální úspěšnost vygenerovaných vzorků po předložení N akcí dané třídy.

4 Závěr

Ověřily jsme, že pro natrénování VAE inkrementálním způsobem na akcích z reálného světa stačí i malé množství (cca 20) trénovacích příkladů. Námi použitá



Obr. 3: Trénovací strategie u společného modelu pro všechny akce. Horní grafy vyjadřují procento správně vygenerovaných akcí, spodní grafy ukazují celkovou chybu vypočtenou na 10 testovacích příkladech. Na ose x je počet příkladů (1-20 je akce *tancovat*, 21-40 akce *létat* a 41-60 akce *mávat*). Ukazujeme scénář bez doučování (vlevo) a s průběžným doučováním (vpravo) předchozích akcí.

data byla však relativně jednoduchá, u komplexnějších dat by bylo jistě třeba dataset zvětšit. Dále jsme ukázaly, že je možné použít i jeden VAE model podmiňovaný různými typy akcí, ale pouze s kompenzačním mechanismem zabraňujícím zapomínání.

Poděkování

Tato práce byla financována Grantovou agenturou ČR (grantem č. GA21-31000S) a Studentskou grantovou agenturou ČVUT, projekt č. SGS21/184/OHK3/3T/37).

Reference

Cao, Z., Simon, T., Wei, S.-E. a Sheikh, Y. (2017). Real-time multi-person 2d pose estimation using part affinity fields. V *Proceedings of the IEEE conference on computer vision and pattern recognition*, str. 7291–7299.

Kingma, D. P. a Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Liu, M.-Y., Breuel, T. a Kautz, J. (2017). Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30.

Petrovich, M., Black, M. J. a Varol, G. (2021). Action-conditioned 3d human motion synthesis with transformer vae. V *Proceedings of the IEEE/CVF International Conference on Computer Vision*, str. 10985–10995.

Personifikovaný robotický chatbot založený na kompozičních dialozích

Michal Vavrečka, Gabriela Šejnová, Petr Schimperk

Český institut informatiky, robotiky a kybernetiky, ČVUT
Jugoslávských partyzánů 3, 160 00 Praha
Email: michal.vavrecka@cvut.cz

1 Úvod

Představujeme novou architekturu chatbota implementovanou v humanoidním robotu Pepper, který je schopen komunikovat v přirozeném jazyce kompozičním způsobem. Oproti předchozím řešením založeným buď na metodě odpovědí na otázky, nebo na dialogových stromech, navrhujeme novou metodu, která konstruuje dialog z atomických bloků, jež lze použít zaměnitelně. Umožňuje tak vytvořit stovky smysluplných dialogů z pouhých několika stavebních bloků. Většina dialogů je zaměřena na preference a názory uživatelů, které jsou uloženy v databázi a převedeny do vektorů příznaků. Chatbot je schopen tyto preference porovnávat s ostatními uživateli nebo s daty z internetu, takže jej lze použít jako doporučovací systém. Pro zlepšení uživatelského zážitku má chatbot vlastní osobnost v podobě předdefinovaných názorů a preferencí pro každou otázku položenou během komunikace. Vtělení chatbota do humanoidního robota umožňuje další zlepšení interakce, neboť robot je schopen detekovat obličej uživatele a zapamatovat si jej pro budoucí dialogy, udržovat oční kontakt a gestikulovat během konverzace. Nakonec uvádíme výsledky předběžného hodnocení se skupinou dobrovolníků, kteří porovnávali našeho chatbota s jiným systémem.

2 Architektura

Architekturu chatbota můžeme rozdělit na moduly závislé na robotovi a konverzační jádro nezávislé na robotovi, které lze v případě potřeby použít jako samostatného textového chatbota. Moduly závislé na robotovi jsou zodpovědné za interakci s uživatelem, jako je detekce člověka nebo rozpoznávání obličeje a řeči. Podrobně je popisujeme v části 3. Zde se zaměříme na konverzační jádro, které se skládá ze strukturálních modulů tvořících hlavní komponenty. Strukturální moduly jsou následující:

- **Báze dat uživatelů** - drží informace o všech uživatelích, jejich názorech a preferencích a také o osobnosti robota
- **Manažer dialogu** - hlavní řídicí prvek toku chatbota, který vybírá další akci na základě zpětné vazby uživatele
- **Výběr témat** - konverzační část, v níž se robot ptá uživatele na jeho preference a odpovídá mu na základě své osobnosti uložené v bázi
- **Doporučovací systém** - algoritmus porovnává odpovědi uživatele s databází nebo internetovými zdroji a doporučuje uživateli podobný obsah
- **Hra/kvíz** - dodatečný zábavní modul nabízí hry související s tématem, které usnadňují komunikaci

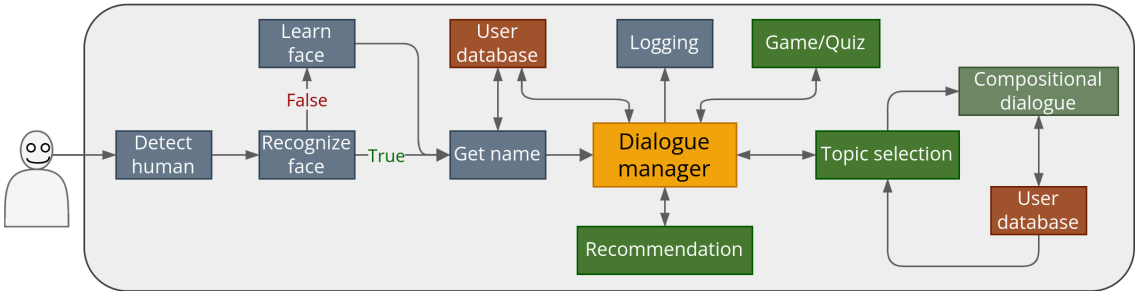
Vazby mezi jednotlivými strukturálními moduly jsou znázorněny na obrázku 1. Protože je architektura plně modulární, je možné chatbota spustit s libovolnou podmnožinou těchto komponent (s ostatními vypnutými) s výjimkou správce dialogu, který je vždy nezbytný pro proces výběru akce.

3 Robot

Vycházeli jsme z rozhraní API pro Python, které vyvinul náš výzkumný tým. Toto API je obal kolem frameworku qi od společnosti Aldebaran, který ovládá Pepper pomocí Pythonu 2.7. Hlavní třída se nazývá "Pepper", nejprve provede připojení k instanci robota se zadanou IP adresou a portem. Poté lze pomocí příkazů pohybovat robotem, snímat vstupy z kamery nebo spustit rozhraní Google Speech Recognition API pro získání přepisu řeči. Pepper má vestavěnou knihovnu pro rozpoznávání řeči, kterou poskytuje společnost Aldebaran. Modul ALSpeechRecognition dává robotovi možnost rozpoznávat předem definovaná slova nebo fráze v několika jazycích. Protože nám však tato knihovna neposkytuje možnost rozpoznávat libovolná slova bez jakýchkoli omezení, bylo v této práci použito rozhraní Google Speech API. Toto rozhraní API zpracovává zvuk přenášený z mikrofonu robota a převádí jej na řeč (text). Řečový modul je vybaven kontextovými gesty, která zvyšují věrohodnost konverzace. Pro detekci lidí a učení/rozpoznávání obličejů používáme vestavěné moduly, které jsou součástí knihovny qi.

4 Výsledky

Provedli jsme předběžné hodnocení robotické verze chatbota se skupinou dobrovolníků. Robot byl umístěn ve veřejném prostoru a uživatelé s ním mohli hovořit.



Obr. 1: Přehled architektury chatbota. Jakmile je detekován člověk, robot se pokusí rozpoznat jeho obličej. Pokud se mu to nepodaří, zapamatuje si obličej a naučí se jméno uživatele. Chatbot využívá informace o osobě uložené v databázi uživatelů k poskytování personalizovaného obsahu. Modul správce dialogu zvolí další nejvhodnější akci - buď Doporučení (nabídne nové tipy z oblasti preferencí uživatele), Hru/Kvíz (související s probíraným tématem), nebo Výběr tématu, po kterém následuje Kompoziční dialog (shromažďuje nové informace o uživateli).

.

Konverzace byly zaznamenávány a sloužily k vylepšení architektury na základě zpětné vazby od uživatelů i analýzy dialogů. Během testování byl robot uživateli pozitivně přijat, někteří si k němu dokonce vybudovali vztah. Naše podrobná analýza reálných dialogů odhalila, že uživatelé se základními předchozími znalostmi o tom, jak mluvit s robotem (krátké odpovědi, strategie střídání), byli s dialogem spokojenější než uživatelé bez předchozích zkušeností - ti mají obvykle vysoké požadavky a očekávání, které nelze pomocí nejmodernějších metod splnit. Kvalitu interakce jsme hodnotili také pomocí dotazníku, který zahrnuje základní aspekty sociální interakce, a to fyzický aspekt komunikace (porozumění a výslovnost), emocionální aspekty (věrohodnost, empatie) sociální aspekty (kvalita dialogu, adekvátnost) a celkové hodnocení (zda se interakce s robotem podobá komunikaci mezi lidmi). Uživatelé měli každý aspekt označit na stupnici od 1 do 5 (nižší je lepší). Pro porovnání rozdílů mezi chatboty někteří uživatelé hovořili s chatbotem Alquist (Pichl a spol. (2020)) implementovaným v robotu Pepper. Při porovnání hodnocení obou chatbotů dosáhl náš systém lepšího průměrného skóre (2,5 SD=(0,4)) ve srovnání s chatbotem Alquist (3,1 SD=(0,6)). Největší rozdíl byl identifikován v kvalitě dialogu. Lze to přičíst direktivnosti a kompoziční struktuře dialogů v našem chatbotu ve srovnání s dialogovými stromy přijatými v chatbotu Alquist.

5 Závěr

Vyvinuli jsme architekturu chatbota ztělesněného v humanoidním robotovi, který je schopen napodobit lidské komunikační dovednosti. Je založena na několika modulech, které jsou spojeny do jednoho rámce. Každý modul zastupuje určitou schopnost. Hlavní výhoda naší architektury spočívá v kompozičním způsobu generování otázek, který nám umožňuje vytvářet mnoho dialogových kombinací z několika málo atomárních částí. Také přítomnost zaměnitelných částí dialogu je v oblasti konverzační strategie chatbotů novou metodou. Přítomnost konzistentní osobnosti chatbota zlepšuje kvalitu dialogu. Doporučovací systém založený na datech získaných od uživatelů je spolehlivější než anonymní názory a preference získané z internetu (za předpokladu, že je shromážděno dostatečné množství dat). Robotické ztělesnění systému zlepšuje věrohodnost chatbota, protože nabízí i neverbální aspekty komunikace.
Do budoucna plánujeme architekturu rozšířit. Kvalita doporučovacího systému se bude zvyšovat s množstvím dat nasbíraných mezi uživateli. Výběr témat a podtémat je také závislý na datech získaných od uživatelů. Analýza dialogů odhalí stabilní komunikační strategie a tyto šablony lze převzít pro výběr témat s novým uživatelem. Plánujeme také implementovat generativní model pro vytváření otázek, abychom proces dotazování plně automatizovali. Je také možné využít generativní modely k navrhování různých typů otázek ke stejnému tématu, abychom eliminovali stereotypnost dialogů.

Poděkování

Tato práce byla podpořena Technologickou agenturou ČR grantem č. TL02000362, projektem INAFYM (CZ.02.1.01/0.0/0.0/16_019/0000766) a Studentskou grantovou agenturou ČVUT, projekt SGS21/184/OHK3/3T/37.

Reference

Pichl, J., Marek, P., Konrád, J., Matulík, M. a Šedivỳ, J. (2020). Alquist 2.0: Alexa prize socialbot based on sub-dialogue models. *arXiv preprint arXiv:2011.03259*.

O procitnutí hmoty[1]

Jiří Wiedermann
Centrum Karla Čapka pro výzkum hodnot ve vědě a technice
Ústav informatiky AV ČR
Pod Vodárenskou věží 2, 182 07 Praha
Česká republika
Email: jiri.wiedermann@cs.cas.cz

Abstrakt

Na rozdíl od tradiční výpočetní teorie mysli nebudeme chápat vlastnosti hmoty a vědomí jako reflexi jejího okamžitého stavu, nýbrž jako výraz vlastností jejího chování během potenciálně nekonečné interakce s prostředím. V tomto kontextu budou základními kameny mentálních aktivit události - fyzikální jevy, zprostředkované senzo-motorickými jednotkami interaktivního agenta a projevující se ve změnách jeho chování. Změna výpočetního paradigmatu nám dovolí řešení dvou obtížných navzájem souvisejících metafyzických problémů: Russellova (1927) problému hmoty a Chalmersova (1996) problému vědomí. Dokážeme platnost následujícího principu procitnutí hmoty: *minimální strojové vědomí a strojově vědomé zakoušení*[2] *jsou fundamentální vlastností živé i neživé hmoty.* Odpovíme na otázky: (i), které procesy dávají vzniknout strojově vědomému zakoušení, (ii), jak fungují příslušné mechanismy, a (iii), jak a proč pomocí těchto mechanismů vzniká strojově vědomé zakoušení, a jaký je jeho účel. Tím vlastně definujeme postačující podmínky pro existenci jisté formy vědomí a vědomého zakoušení. Domníváme se, že do té míry, do jaké naše modelování zachycuje základní aspekty vyšší mentální činnosti mozku, naše závěry platí i pro lidské vědomí a zakoušení.

Rozšířený abstrakt

Cílem příspěvku je popsat výsledky probíhajícího projektu (Wiedermann & van Leeuwen, 2022), jenž se zabývá dodnes nevyřešenou filozofickou otázkou vztahu vědomí a fyzického světa. Jak je možné, že hmotný objekt, jakým je mozek, disponuje nehmotným vědomím? Jak fungují příslušné mechanismy? Jak vzniká vědomé zakoušení (prožívání)? A proč vůbec máme tyto schopnosti? K čemu jsou dobré, jaký je jejich smysl? Je mozek jediným příkladem hmoty, jež má vědomí?

Tyto a podobné otázky formulovali zejména dva významní filozofové moderní éry. Prvním z nich byl významný britský filozof a matematik Bernard Russell, jenž konstatoval, že *„o fundamentální kvalitě mentálních jevů mimo nás nevíme dost na to, abychom řekli, jestli se liší, anebo neliší od těch mentálních jevů, jež právě zakoušíme"* (1927, s. 221; 1950/1995). Tuto myšlenku později popularizoval britský filozof Galen Strawson (2016, 2017) pod názvem *„obtížný problém hmoty"*: jaké jsou fundamentální kvality fyzikálních jevů, resp. obecně, hmoty? Samozřejmě, tyto kvality musí zahrnovat *„mentální jevy, jež právě zakoušíme"*, tj., kválie, resp. fenomenální zakoušení, či obecněji, vědomí. Tato formulace přímo vede na další problém z oblasti metafyziky, jenž formuloval australský filozof David Chalmers v devadesátých letech 20. století (Chalmers, 1995, 1996) jako tzv. *„obtížný problém vědomí":* vysvětlit, proč a jak máme kválie či fenomenální zakoušení.

Na tuto a podobné otázky odpovíme z hlediska umělé inteligence, kybernetiky a robotiky. Ukážeme, že platí následující princip procitnutí hmoty: *schopnost mít minimální strojové vědomí a strojově-vědomé zakoušení je fundamentální vlastností živé i neživé hmoty*, a vysvětlíme, proč tomu tak je a jak to funguje. Tím vyřešíme oba problémy.

V porovnání s klasickým metafyzickým myšlenkovým rámcem úvah o vztahu mysli a hmoty nám řešení obou problému umožní dvě zásadní změny.

První změnou je radikální odklon od řešení problému vědomí v kontextu věd o mozku ve prospěch obecného řešení Russellova obtížného problému hmoty, tj. řešení nejen v rámci živé hmoty.

Druhou změnou je změna tradičního paradigmatu, doposud používaného ve výpočetní teorii mysli.

[1] Tato práce vznikla na základě společného výzkumu s Janem van Leeuwenem, Utrecht, NL, viz (Wiedermann & van Leeuwen, 2022), za částečné podpory institucionálního plánu ÚI AV ČR RVO 67985807 a programu Strategie AV21.

[2] Termín „zakoušení" často požíval Ivan M. Havel pro označení fenomenálních prožitků. My jej používáme z důvodu, že jinak bychom museli hovořit o prožitcích hmoty, což se nehodí z důvodu, že máme na mysli i neživou hmotu.

Změněné paradigma nahlíží na vlastnosti hmoty nikoliv jako na reflexi jejího okamžitého stavu, nýbrž jako výraz vlastností jejího chování během potenciálně nekonečné interakce s prostředím. V tomto kontextu jsou základními kameny mentálních aktivit události (fyzikální jevy), zprostředkované senzo-motorickými jednotkami agenta a projevující se ve změnách chování agenta.

Modelem, jenž stojí za změněným paradigmatem, je *interaktivní vtělený behaviorální agent*. Formálně se jedná o interaktivní *transducer* – počítačem řízený překladač nekonečného proudu tzv. *situací* na odpovídající *chování*. Situace jsou zprostředkovány agentovi prostřednictvím jeho vnějších a vnitřních senzorů, chování realizují agentovy motorické jednotky. V obecním případě je agent realizován pomocí poměrně složité, ale reálné technologie, asi na úrovni současných nejprogresivnějších robotů anebo samořiditelných vozidel.

Interaktivní transducer jako model agenta umožní definovat sémantické vlastnosti chování (procesů, událostí), které generuje agent, což je velmi blízké pohledu B. Russella (1927). Změněný model umožňuje vidět strukturu interakcí - dvojic *(současná situace, odpovídající chování)* - a výpočetně ji ovlivňovat tak, aby chování agenta splňovalo netriviální sémantické vlastnosti v každé situaci, se kterou agent setká. Netriviální sémantické vlastnosti jsou vlastnosti, jimiž se odlišuje chování agenta od jiného agenta, jenž takové chování nesplňuje. Příkladem takových vlastností u člověka jsou jeho mentální schopnosti, u našeho agenta to budou jejich strojové ekvivalenty, vyvolávající podobné chování, jako u lidí.

Zásadní výsledek, od kterého odvineme princip procitnutí hmoty, je *princip kauzality*, jenž tvrdí, že vlastnost agenta splnit danou netriviální sémantickou vlastnost se nevyhnutně projeví na jeho chování, které se nutně v některých situacích musí lišit od chování agenta, jenž danou vlastnost nesplňuje. To nám dále umožní definovat pojem *strojového zakoušení* jako fyzikální jev (úkaz, fenomén), jenž nezávisí na materiální podstatě (substrátu) agenta a je způsoben chováním agenta. Tento fyzikální jev je doprovázen autentickými senzo-motorickými a výpočetními akcemi, jež jsou nutné pro zabezpečení plnění dané netriviální sémantické vlastnosti a jež dodávají strojovému zakoušení potřebné charakteristické znaky interního sebe-pozorování události či aktivity, které se agent účastní (Wiedermann & van Leeuwen, 2022).

Tyto charakteristické znaky jsou i), privátní, nepřístupné externímu pozorovateli, ii), příslušné strojové pocity není možné přesně a srozumitelně sdělit externímu pozorovateli, iii), představují fundamentální (intrinsickou) znalost příslušnou danému agentovi a typu zakoušení, a iv), tato znalost nejde oddělit od daného zakoušení.

Na tomto základě odpovíme v kontextu našeho modelu na tři otázky, jejichž odpovědi jsou dle Chalmerse (1996) postačující podmínkou pro vznik vědomého zakoušení.

Za prvé, budeme definovat procesy, jež dávají vzniknout strojovému vědomému zakoušení. To jsou procesy, splňující specifické netriviální vlastnosti, definované ve čtyřech principech minimálního strojového vědomí. Těmito principy jsou princip *sebe-poznání, sebe-pozorování, sebe-uvědomění a sebe-informování* (Wiedermann & van Leeuwen, 2019, 2020, 2021a, 2021b, 2022). Popíšeme vlastnosti těchto procesů a ukážeme, že je lze výpočetně realizovat.

Za druhé, ukážeme, jak v našem modelu fungují mechanismy generující minimální strojové vědomí a strojově vědomé zakoušení.

A za třetí zdůvodníme, proč pomocí těchto mechanismů vzniká strojové vědomé zakoušení a jaký je jeho účel.

Podle naší teorie je účelem strojového zakoušení nejen poskytnout agentovi strojový pocit *„jaké to je, být tím agentem, kterým agent právě je a zakoušet to, co právě agent zakouší“*, ale i definovat pro agenta reprezentace nových pojmů pro koncepty - strojové prožitky a emoce - které v reálném světě neexistují, ale vznikají právě v souvislosti se strojovým zakoušením. To dále umožní stejně konstruovaným agentům operujícím ve stejném prostředí komunikovat o těchto konceptech (tj. o jejich vnitřním zakoušení) a budovat nad nimi další abstraktní koncepty.

Tyto tři podmínky dohromady definují postačující podmínky pro vznik strojového vědomého zakoušení a tím pádem řeší, v rámci našeho modelu, oba dva obtížné problémy metafyziky. Dále vedou k formulaci následujícího principu procitnutí hmoty:

Minimální strojové vědomí a strojově vědomé zakoušení jsou fundamentální vlastností živé i neživé hmoty.

Jsme přesvědčeni, že do té míry, do jaké naše metafyzické modelování zachycuje základní aspekty mentální činnosti mozku, naše závěry platí i pro lidské vědomí a zakoušení.

Reference

Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies* 2 (3):200-19.

Chalmers, D. (1996). *The Conscious Mind,* New York: Oxford University Press,

Russell, B. (1927a). *An Outline of Philosophy.* London: Routledge.

Russell, B. (1927b). *The Analysis of Matter.* London: Kegan Paul, Trench, Trubner; New York: Harcourt, Brace.

Russell, B. (1950/1995). Mind and Matter. *In Portraits from Memory.* Nottingham: Spokesman.

Wiedermann, J. & van Leeuwen, J. (2019). Finite State Machines with Feedback: An Architecture Supporting Minimal Machine Consciousness. In: Manea F., Martin B., Paulusma D., Primiero G. (Eds): *Computability in Europe: Computing with Foresight and Industry* (CiE 2019), Lecture Notes in Computer Science, Vol. 11558, Springer, 2019, pp. 286-297

Wiedermann, J. & van Leeuwen, J. (2020). Towards minimally conscious cyber-physical systems: A design philosophy. Technical Report UU-PCS-2020-02, Center for Philosophy of Computer Science, Department of Information and Computing Sciences, Utrecht University, https://www.cs.uu.nl/groups/AD/UU-PCS-2020-02.pdf

Wiedermann, J. & van Leeuwen, J. (2021a). Towards minimally conscious finite-state controlled cyber-physical systems: A manifesto. In: T. Bureš et al. (Eds.), SOFSEM 2021: Theory and Practice of Computer Science, Lecture Notes in Computer Science Vol 12607, Springer-Verlag, 2021, pp. 43-55

Wiedermann, J. & van Leeuwen, (2021b). Validating Non-trivial Semantic Properties of Autonomous Robots. Presented at the conference Philosophy & Theory of Artificial Intelligence, PT-AI 2021, Gothenburg 2021; in: V. Muller et al. (Eds.), *Proceedings PT-AI 2021*, Springer (in print)

Wiedermann, J. & van Leeuwen, J. (2022). On non-trivial semantic properties of physical events. Manuscript in preparation.

Rejstřík autorů

A
Adamus, Magdalena 8
Andrejková, Gabriela 162

B
Ballová Mikušková, Eva 8
Bečková, Iveta 34, 160
Brezina, Ivan 15
Bujok, Petr 98
Burič, Roman 19

C
Chmelík, Jiří 176
Chmelařová, Kateřina 175
Cienciala, Luděk 144
Ciencialová, Lucie 144
Cimrová, Barbora 23, 34

Č
Čavojová, Vladimíra 15, 122

D
Dostál, Daniel 66

F
Fandl, Matej 28
Farkaš, Igor 23, 34, 74, 84, 160
Friebe, Kassandra 112

G
Galasová, Miroslava 39

H
Hoffmann, Matěj 45, 112, 173
Hoza, Petr 52
Hvorecký, Juraj 58

J
Jochecová, Kateřina 175
Juřík, Vojtěch 91, 106
Jurkovičová, Lenka 106

K
Knott, Alistair 126
Kopčo, Norbert 162, 164
Korečko, Štefan 23

L
Lúčny, Andrej 62
Linková, Stanislava 162
Lokša, Peter 164

M
Macháčková, Lenka 66
Malinovská, Kristína 74, 79, 112
Malinovský, Ľudovít 79
Marko, Martin 23
Mirdalíková, Ivana 166

O
Ostapenko, Anastasia 168

P
Páleník, Jan 106
Pócoš, Štefan 34, 160
Pešán, Jan 91
Pecháč, Matej 84
Petrovická, Klára 170
Poláková, Radka 98

R
Rošťáková, Zuzana 23
Rosipal, Roman 23
Ružičková, Alexandra 106

S
Sagar, Mark 126
Samporová, Sabína 112

Schimperk, Petr 179
Sobota, Branislav 23
Stachoň, Zdeněk 175, 176
Straka, Zdeněk 173
Svoboda, Aleš 117

Š

Šašinka, Čeněk 175, 176
Šašinková, Alžběta 176
Šejnová, Gabriela 177, 179
Šrol, Jakub 122
Štěpánová, Karla 177

T

Takáč, Martin 28, 126
Tomašková, Silvia 133

V

Vadinský, Ondřej 52, 137
Valenta, Daniel 144
Vavrečka, Michal 168, 179
Vavrečková, Šárka 151

W

Wiedermann, Jiří 181