Министерство образования и науки Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования Российский государственный гуманитарный университет (РГГУ)

Цифровое общество в культурно-исторической парадигме

МАТЕРИАЛЫ
МЕЖДУНАРОДНОЙ НАУЧНОЙ КОНФЕРЕНЦИИ

ПОД РЕДАКЦИЕЙ: Т. Д. МАРЦИНКОВСКОЙ, В. Р. ОРЕСТОВОЙ, О. В. ГАВРИЧЕНКО

МОСКВА 2018

DIRECTIONS OF STUDIES IN CYBERPSYCHOLOGY

Voiskounsky A.E.

Lomonosov Moscow State University, Moscow

Keywords: cyberpsychology, virtuality, anonymity, hybrid behavior, leveling up reputation, mobility, immersion, distributed behavior

Abstract. The leading directions of cyberpsychological studies describing human behavior (interactive, cognitive, gameplaying, consumer, etc.) on the Internet are introduced and discussed. These directions include: anonymity, hybrid behavior (transfer from virtual to real life and vice versa), leveling up reputation, mobility, immersion, distribution.

«ОТРАВЛЯЮЩИЕ АТАКИ» НА МАШИННОЕ ОБУЧЕНИЕ И СЕМАНТИЧЕСКИЕ СЕТИ: ВЕРСИЯ НАРРАТИВА УГРОЗЫ В ЦИФРОВОМ ОБЩЕСТВЕ[2]

Поддьяков А. Н.

Национальный исследовательский университет «Высшая школа экономики», Институт психологии РАН, г. Москва

Ключевые слова: троянское обучение, отравляющая атака, вербальная креативность, семантическая сеть, нарратив угрозы

Аннотация. Дается краткий обзор некоторых исследований дачи ложных подсказок и троянского обучения (обучения со скрытыми, не декларируемыми целями тому, о чем не подозревает обучаемый). Предлагается рабочая классификация, позволяющая различать ситуации ложных подсказок и троянского обучения, в разной степени связанные с речевой деятельностью как одной из важнейших для человека. С целью объединения исследовательского поля, отталкиваясь от аналогии с исследуемыми в настоящее время «отравляющими атаками» на системы машинного обучения», можно интерпретировать ложные подсказки и троянское обучение как «отравляющую атаку на семантическую сеть». Рассматриваются возможные аспекты проблемы, связанные с подавлением вербальной креативности и исполнения. В заключение ставится вопрос о динамике представленности нарративов о ложных подсказках (обучении со злым умыслом) в различные периоды и в различных обществах, в том числе цифровом.

В театральной среде рассказываются истории о розыгрышах, когда актера более опытные партнеры предостерегают перед выходом на сцену: «Ты там не скажи … вместо …» (за «не скажи» идет неправильное слово или смешно переиначенная реплика). И именно то, от чего якобы отговаривали, на сцене закономерно и произносится – на радость советчику и залу. Другой вариант – неправильная подсказка от суфлера, решившего, например, проучить зазнавшегося актера или актрису. Эти истории – часть

[2] Исследование поддержано РФФИ, проект 18-29-03167.

более широкого пласта нарративов, восходящих к таким общекультурным формам накопления и передачи социального опыта как народные сказки. В них нередко представлены ситуации, когда одни персонажи учат других тому, что для последних невыгодно или опасно: Баба-Яга учит Иванушку садиться на лопату, чтобы засунуть его в печь; лиса учит волка ловить рыбу на собственный хвост в проруби, в результате хвост примерзает, и волк его лишается; Братец Кролик учит Братца Лиса, как вести себя тому, кто хочет правдоподобно изобразить покойника при появлении друзей, и т. д.

Ложные наводки и подсказки, акты «троянского» (обманного) обучения стали объектом научного изучения (Лефевр, 1973; Поддьяков, 2006, 2011; Kline, 2015; Rhodes et al., 2015).

В.А. Лефевр в своей теории конфликтующих структур и рефлексивного управления ввел понятие «формирование доктрины противника посредством его обучения». Одно из проявлений такого обучения на материале спорта состоит в том, что «футболист-нападающий систематически сознательно "попадается" на определенное действие одного из защитников. В результате защитник закрепляет данное действие как стандарт противодействия данному нападающему, что и используется нападающим в решающий момент» (Лефевр, 1973, с. 51).

Понятие «троянское обучение» (обучение со скрытыми, не декларируемыми целями тому, о чем не подозревает обучаемый) пересекается с понятием «формирование доктрины противника посредством его обучения», но не совпадает с ним (Поддьяков, 2011). Не всякое формирование доктрины противника посредством его обучения является скрытым троянским обучением. Например, в случае убежденности субъекта в своем явном превосходстве и в будущем проигрыше соперника он может предъявить ему свою и его доктрину для сопоставления в явном виде, не таясь, без скрытых манипуляций, в расчете на здравый смысл противостоящего субъекта. Также есть троянское обучение «с добрым умыслом», рассматривающее другого субъекта не как противника, а как не вполне разумного подопечного, которому из лучших побуждений стремятся помочь – например, вводя содержание, которому ученик не хочет обучаться, в особо привлекательной оболочке, и именно она выглядит для ученика главной составляющей (Boyle, 2001; White, 2004). При этом психологическая манипуляция и обучение тому, о чем не догадывается обучаемый, есть в ситуациях троянского обучения обоих типов – и со злым, и с добрым умыслом.

Мы провели теоретический анализ явления троянского обучения и серию эмпирических исследований с участием людей разного возраста (Поддьяков, 2006, 2011). Так, дошкольникам описывалась сказочная ситуация, в которой злые гиены решили поохотиться на беззащитных птенчиков, а храбрый львенок Симба решил их спасти. При этом и гиенам, и львенку для реализации их жестоких или же гуманных замыслов не хватает владения некоторыми знаниями и умениями. Ребенку задавались вопросы, надо ли учить гиен (или львенка) правильному или неправильному птичьему языку, учить ли их хорошо лазать по деревьям и т.д. Абсолютное большинство детей 5-6 лет давали ответы о необходимости правильного, эффективного обучения львенка и обманного обучения гиен. Это вполне согласуется с житейскими наблюдениями – на вопрос актера, играющего в спектакле волка, о том, куда убежали зайцы, дети отвечают так, чтобы обмануть его. Но в нашем эксперименте речь шла не просто об указании неправильного направления, а о некотором базовом понимании ребенком, что такое обучение, и понимании различных последствий правильного и обманного обучения в разных областях (Поддьяков, 2006).

В аппаратурном эксперименте M. Rhodes с коллегами показано, что дошкольники могут учить другого правильно, а могут, обманывая – создавая условия, чтобы другой сделал неправильные выводы из представленной ему информации. А именно, дети демонстрировали кукле работу технической игрушки на релевантных примерах, если взрослый просил ребенка показать кукле такие примеры, из которых можно узнать правило работы этой игрушки. Если же взрослый просил ребенка подшутить над куклой и запутать ее так, чтобы она пришла к неправильному заключению о работе устройства, дети подыскивали и показывали кукле нерелевантные примеры, провоцирующие ошибочный вывод (Rhodes et al., 2015). Задача требовала понимания логики работы устройства, умения строить умозаключения, а также социального интеллекта (встать на позицию другого, понять, из какой информации какие выводы он может сделать, и суметь обмануть его).

М. Клайн, ссылаясь на С.М.Камакау, пишет, что на Гавайях, где между жителями-рыбаками была очень высока конкуренция за рыбные ресурсы, дети должны были быть «скептичными учениками», поскольку имелся значимый риск стать жертвой обмана. Она также ставит более общую проблему «скептицизма» учащихся по отношению к информации, получаемой, возможно, от не вполне добросовестного «донора» (Kline, 2015).

Ситуации троянского обучения достаточны распространены в обычной жизни – таково мнение участников опроса – 393 россиян и 279 американцев от 16 до 59 лет. Более 80% респондентов – и россиян, и американцев – ответили, что обучение «со злым умыслом» бывает в реальной жизни и имеет место в школах и университетах. Около половины участников отмечали случаи, когда их учебе мешали из недружественных побуждений, а также пытались проводить по отношению к ним обучение «со злым умыслом». От 9 до 23% респондентов в разных подгруппах (в том числе некоторые профессиональные преподаватели) ответили, что сами проводили такое обучение по отношению к кому-то (Поддьяков, 2011).

В области машинного обучения, где системам искусственного интеллекта необходимы большие массивы обучающих примеров, изучаются возможности хакеров в отношении организации «отравляющих атак» на базы этих примеров (Jagielski et al., 2018). Речь идет о том, чтобы, скрыто подгрузив в базу минимальное количество особым образом подобранных примеров, нарушить процесс эффективного обучения и последующего принятия решений. (Подходящая метафора – минимально необходимая ложечка дегтя для порчи наибольшей бочки меда.) Понятно, что порча, «отравление» совокупности примеров, собранных для обучения распознавания болезни, может иметь серьезные практические следствия. Пока, к счастью, таких прецедентов не было, но превентивное исследование возможностей в этой области ведется – как и разработка контрмер (там же).

При этом, подчеркнем, здесь пока не идет речь об изменении самой обучаемости системы хакером при атаке. Атакуется, «отравляется» лишь массив предъявляемых примеров. Но теоретически возможны атаки именно на обучаемость – если рассматривать ее как потенциально подверженную влиянию техническую характеристику системы (Поддьяков, 2007).

Далее мы обратимся к пересечению некоторых из обозначенных выше тем. Это возможная модель снижения, подавления вербальной креативности и исполнения путем «отравляющей атаки» на семантическую сеть.

Предложим рабочую классификацию, позволяющую различать ситуации, в разной степени связанные с речевой деятельностью – одной из важнейших для человека.

А именно, ложные подсказки («отравляющие атаки», троянское обучение), в разной степени связанные с речевой деятельностью, можно классифицировать по параметрам «мишень – средство влияния» следующим образом.

1. Мишени:

а) речевая деятельность;

б) не речевая деятельность

Критерий различения: используется ли ложная подсказка («отравляющий пример», троянское обучение) для того, чтобы вызвать сбой, неуспех речевой или же другой – неречевой – деятельности.

2. Средства ложной подсказки (троянского обучения):

а) речевые (вербальные);

б) невербальные средства (картинка, указательный жест и т.д.).

Тогда в рамках этой рабочей классификации мы получаем 4 класса ситуаций.

1. Мишень – речевая деятельность, средство влияния – речевой, вербальной природы. Примеры в человеческом общении – актерские ложные подсказки, обучение не понимающих «иностранным языкам», нецензурным выражениям под видом благопристойных и пр.

2. Мишень – речевая деятельность, средство влияния – невербальные наводки, указания (например, указательный жест не на то место в учебнике, которое надо читать, а на другое, для соученика при опросе учительницей).

3. Мишень – неречевая деятельность, средство влияния – речевая деятельность (пример – вербальная инструкция новичку, как сделать что-то технологически неправильное, чтобы подшутить или же скомпрометировать, устранить потенциального конкурента).

4. Мишень – неречевая деятельность, средство влияния – невербальные наводки, указания (например, указательный жест, провоцирующий неправильное действие).

В предлагаемых терминах, ложные подсказки, «вредные советы» актеру («Ты там не скажи … вместо …») – это «отравляющая атака» на его семантическую сеть. То, от чего якобы предостерегают, становится «отравляющим», «заякоривающим», закрывающим всё остальное примером с неадекватно большим весом в семантической сети. Настолько большим, что произнесение на сцене озвученного до этого «шутником» неправильного варианта становится очень вероятным. При этом в живом общении людей огромное значение имеют и личные коммуникативные способности «подсказчика», включая невербальную составляющую, а также подверженность выбранной жертвы психологическому влиянию (вообще или только влиянию данного человека, в данных обстоятельствах).

При моделировании отравляющей атаки на виртуальную семантическую сеть речь может идти о влиянии посредством изменения весов узлов и связей. Напрашивающаяся грубая метафора – «дебилизация» сети в отношении возможностей поиска на ней интеллектуальных и креативных решений. Но, теоретически рассуждая, «отравляющая атака» может быть направлена и на неадекватное повышение креативности. Если семантическая сеть, поддерживающая разработку креативных рекламных слоганов, в результате хакерской атаки начнет выдавать чересчур креативные и потому «дико выглядящие», непонятные для большинства целевой аудитории слоганы, эту атаку тоже

можно считать «отравляющей» - интеллектуальная система сильно отклонилась от требуемого оптимума решений.

Анализируя динамику представленности нарративов о ложных подсказках (обучении со злым умыслом) в различные периоды и в различных обществах можно оценить воспринимаемую важность (опасность) этих явлений в общественном сознании. В том числе, представляет интерес вопрос, претерпит ли эта динамика изменения в «цифровом обществе», где идея троянского обучения (в виде отравляющих атак на системы машинного обучения) органично встраивается в представления о нарастающих возможностях взлома всего и вся – и, вероятно, укрепляет эти представления. Или же динамика представлений не изменится, а просто в массив сюжетов о ложных подсказках и обучении со злым умыслом будет добавлено еще несколько подтипов, что на настоящий момент тоже можно считать вероятным.

Литература.

Лефевр В.А. Конфликтующие структуры. М.: Советское радио, 1973.

Поддьяков А.Н. Противодействие обучению конкурентов и троянское обучение в информационных технологиях // 1-ая Международная конференция по бизнес-информатике. Труды международной научно-практической конференции, 9-11 октября 2007 г. Звенигород, 2007. С. 261-269. http://www.hse.ru/data/142/913/1235/zvenigorod.doc.

Поддьяков А.Н. Психология конкуренции в обучении. М.: Изд. дом ГУ-ВШЭ, 2006.

Поддьяков А.Н. Троянское обучение в экономическом сознании и поведении // Культура и экономическое поведение / Под ред. Н.М.Лебедевой, А.Н. Татарко. М.: МАКС Пресс, 2011. С. 421-444.

Boyle M. The Computer as a Trojan horse *// Journal of Computer Assisted Learning.* 2001. Vol. 17. P. 251-262.

Jagielski M., Oprea A., Biggio B., Liu C., Nita-Rotaru C., Li B. Manipulating machine learning: poisoning attacks and countermeasures for regression learning. 2018. https://arxiv.org/abs/1804.00308.

Kline M. How to learn about teaching: an evolutionary framework for the study of teaching behavior in humans and other animals // Behavioral and Brain Sciences. 2015. Vol. 38. e31. doi:10.1017/S0140525X14000090

Rhodes M., Bonawitz E., Shafto P., Chen A., Caglar L. Controlling the message: preschoolers' use of information to teach and deceive others // Frontiers in Psychology. 2015. 6. 867. doi:10.3389/fpsyg.2015.00867

White A.L. The pedagogical Trojan horse: handheld technologies in the secondary mathematics classroom // *Proceedings of the 2nd National Conference on Graphing Calculators.* October 4-6, 2004. P. 105-112.

Сведения об авторе.

Поддьяков Александр Николаевич. Доктор психологических наук, профессор. Национальный исследовательский университет «Высшая школа экономики».

POISONING ATTACKS ON MACHINE LEARNING AND SEMANTIC NETWORKS: A VERSION OF A THREAT NARRATIVE IN DIGITAL SOCIETY

Poddiakov, A.

National Research University Higher School of Economics, Institute of Psychology of RAS, Moscow

Key words: “Trojan horse” teaching, poisoning attack, verbal creativity, semantic network, threat narrative

Abstract. A brief review of studies of false tips and “Trojan horse” teaching is presented. A classification giving opportunity to distinguish between situations of false tips (“Trojan horse” teaching) differing in their relations to verbal activity is described. Based on an analogy with “poisoning attacks on machine learning”, one can interpret some verbal false tips and “Trojan horse” teaching as “poisoning attacks on a semantic network”. Various aspects of inhibition of verbal creativity and performance are considered. An issue of dynamics of narratives of false tips and teaching “with evil intent” in various periods and various societies including digital one is introduced.

HOMO DIGITAL: ТРАНСФОРМАЦИИ ИДЕНТИЧНОСТИ В ИНФОРМАЦИОННОЙ КУЛЬТУРЕ

Гусельцева М.С.

ФГБНУ «Психологический институт РАО», Москва

Ключевые слова: методология, информационная культура, цифровая среда, идентичность, транзитивность, виртуальность, современность.

Аннотация: Современность и происходящие в ней изменения являются сферой трансдисциплинарного познания. В психологических исследованиях идентичности важны контексты глобализации, транзитивности, виртуальности, а также микширование тенденций и появление новых системных качеств. В изучении информационной культуры произошел переход от анализа общих характеристик к дифференциации слоев; появились разновидности информационной грамотности: компьютерная, цифровая, коммуникативная, сетевая (нетикет), медиаграмотность; меняются коммуникативные стратегии и нормы. В качестве устойчивых трендов представлены самоорганизация сетевых сообществ, персонализация, вариативность цифрового поведения. Посредством интерактивности, персонификации цифровой среды, творчества новых культурных практик и норм развивается субъектность. Неочевидным последствием информатизации культуры становится тот факт, что цифровая среда латентно влияет на всех, даже на тех, кто не пользуется электронной почтой, смартфоном и компьютером.

НОВЫЕ ВЫЗОВЫ СОВРЕМЕННОСТИ

Современность как предмет изучения является областью трансдисциплинарного познания: представители разных наук описывают ее особенности, выделяют ведущие тенденции. В психологических исследованиях современность предстает в аспектах