Модель индивидуального и коллективного сознания на основе принципов “социального доказательства” и “свободной энергии”

Антон Колонин

akolonin@aigents.com

Telegram: akolonin

Новосибирский государственный университет

*НАСТОЯЩАЯ НАУКА

https://www.nsu.ru

AIGENTS

https://aigents.com

AGI IN RUSSIAN

https://agirussia.org

Какие вопросы нас волнуют?

Почему мы так поступаем и в это верим?

Почему мы сбиваемся во враждующие стаи?

Как нами ~~манипулируют~~ “мягко управляют” психотерапевты, политики, маркетологи и мошенники?

Чем на самом деле страшен сетевой супер-ИИ?

Как создавать эмпатичный ИИ?

“Принцип веры”

Наивность

...
Куда **спокойней раз поверить**,
Чем жить и **мыслить** каждый день.

Так **бойтесь тех**, в ком дух железный,
Кто преградил сомненьям путь.
В чьём сердце страх увидеть бездну
Сильней, чем страх в неё шагнуть.
…

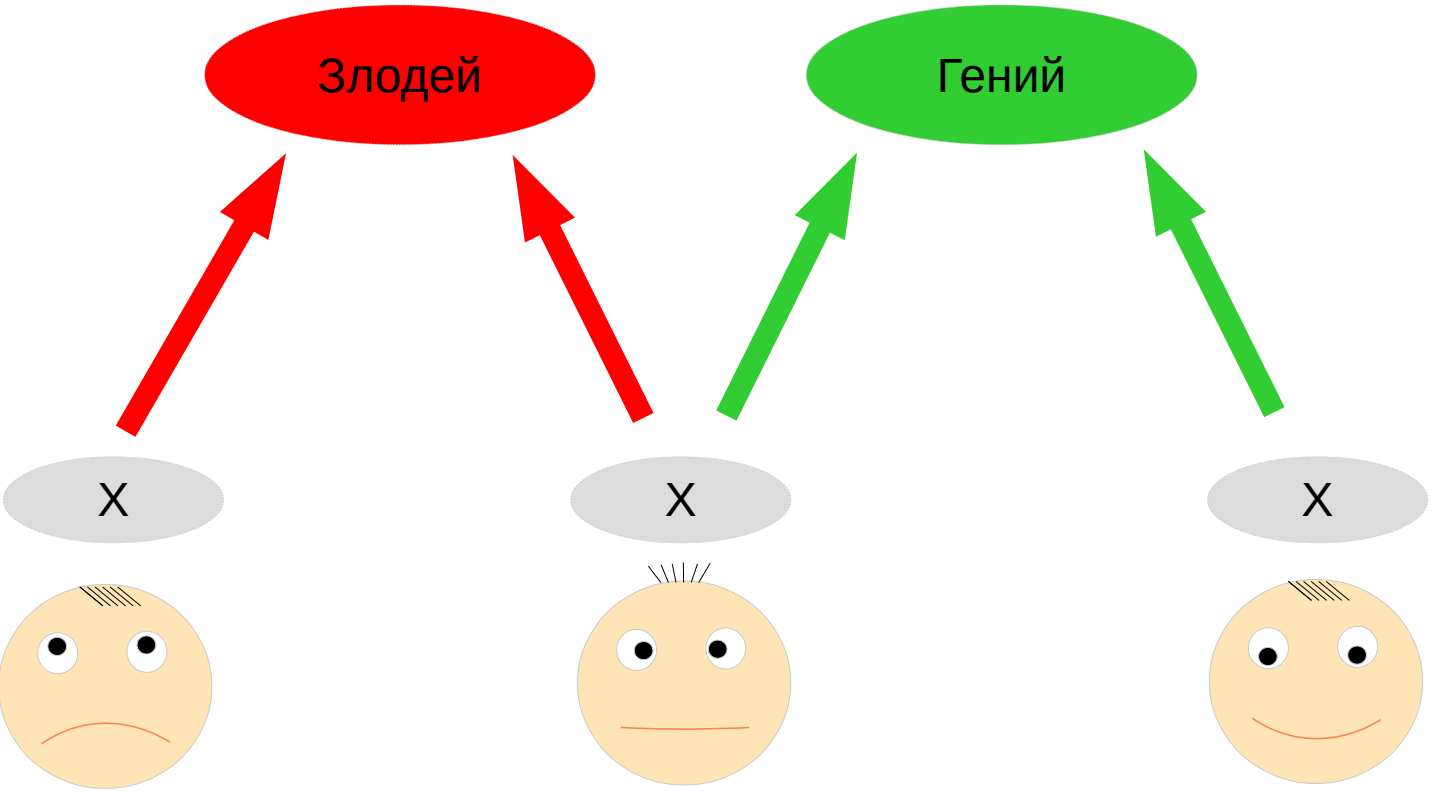
Наум Коржавин, 1963

http://www.belousenko.com/books/Korzhavin/korzhavin_naivnost.htm

“Стадное чувство”

Когда “мудрость толпы” затмевает “голос разума”

“Простота – хуже воровства”



Надежда на “однорукого консультанта”

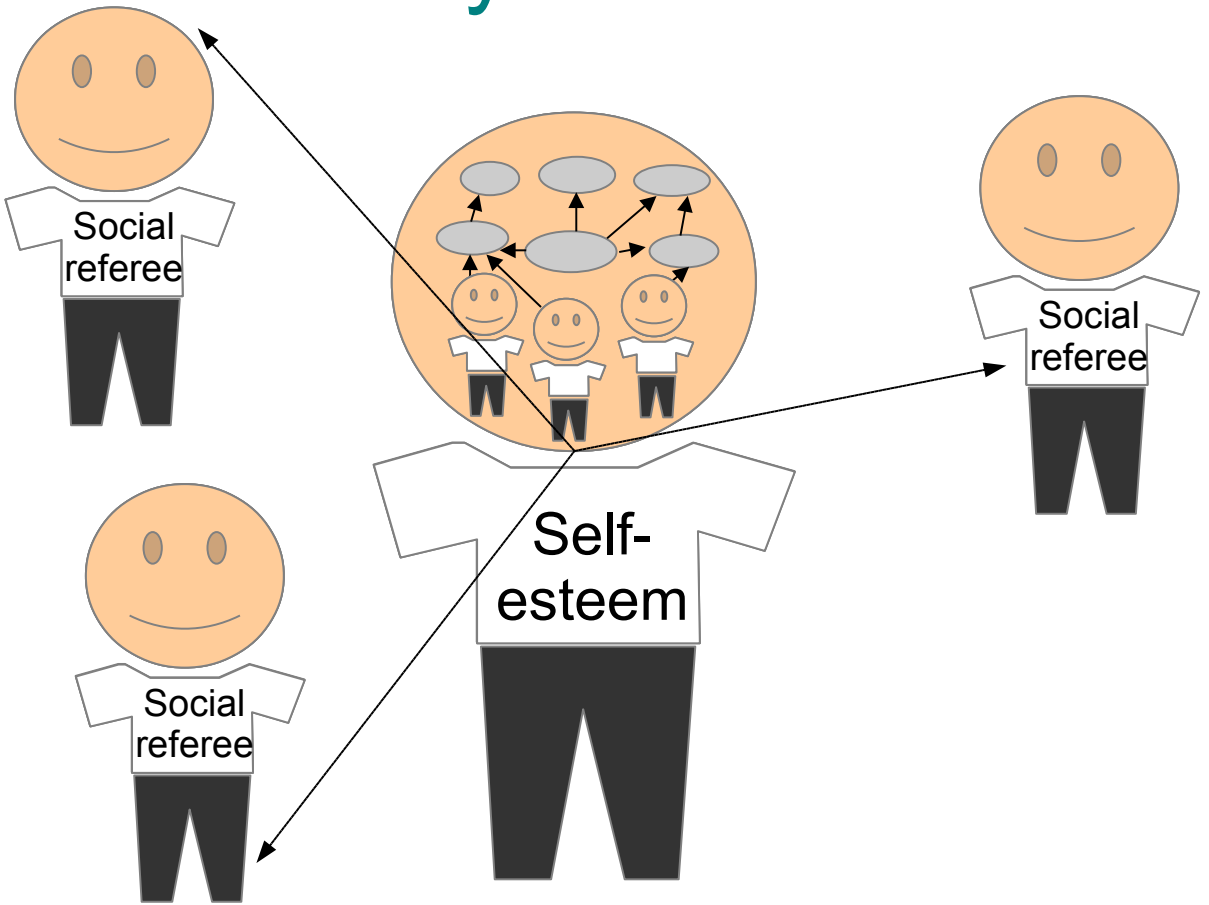
Computable cognitive model based on social evidence and restricted by resources

B.Goertzel, A.Kolonin, J.Pressing, C.Pennachin: Compassion-based artificial psyche design (2000)
https://www.goertzel.org/benzine/WakingUpFromTheEconomyOfDreams.htm

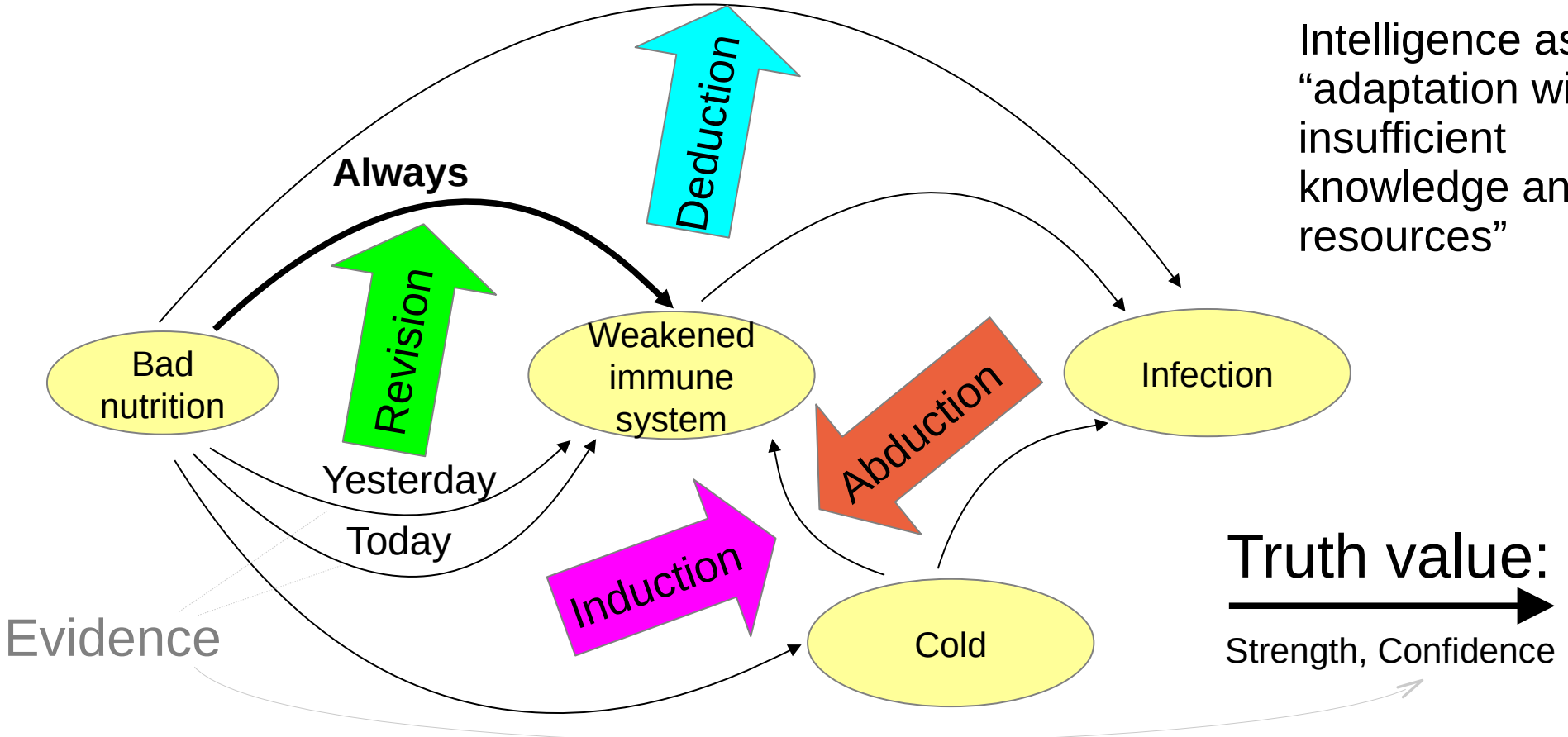
A.Kolonin: Computable cognitive model based on social evidence and restricted by resources (2015)
https://ieeexplore.ieee.org/document/7361869

Anton Kolonin, Evgenii Vityaev, Yuriy Orlov: Cognitive Architecture of Collective Intelligence Based on Social Evidence (2016)
https://www.sciencedirect.com/science/article/pii/S1877050916317239

A.Kolonin: Resource-Constrained Social Evidence Based Cognitive Model for Empathy-Driven Artificial Intelligence (2018)
https://link.springer.com/chapter/10.1007/978-3-319-97676-1_10



Pei Wang: Non-Axiomatic Reasoning System (NARS)



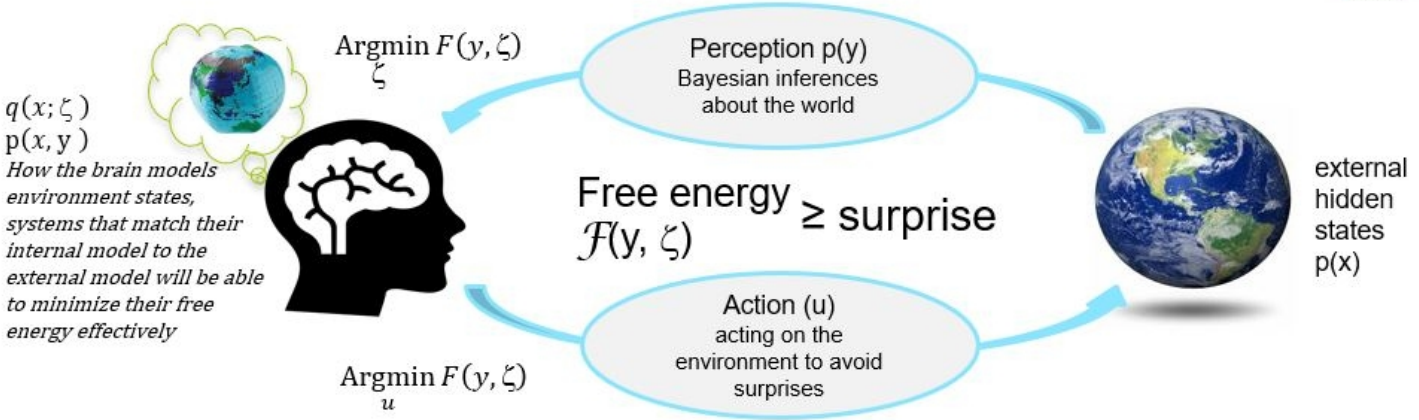
Pei Wang: Non-axiomatic reasoning system: exploring the essence of intelligence, 1993-1996
https://www.researchgate.net/publication/2690739_Non-Axiomatic_Reasoning_System_-_Exploring_the_Essence_of_Intelligence

Karl Friston: Free Energy Principle

Minimize free energy

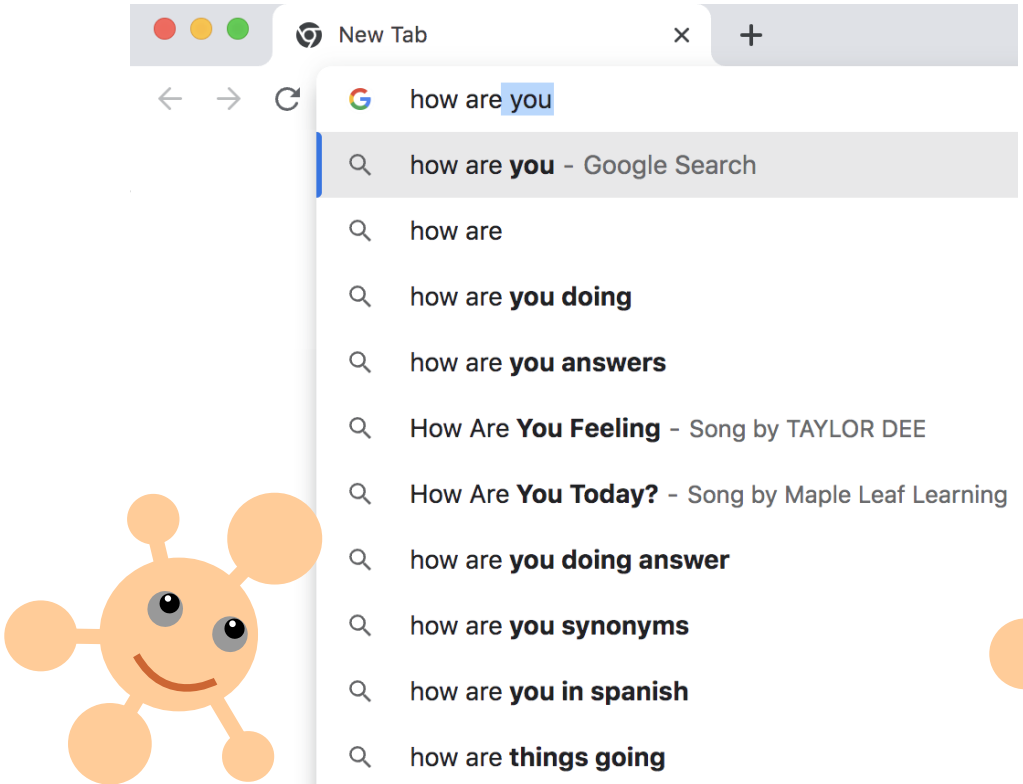
¨in summary, (i) agents resist a natural tendency to disorder by minimising a free-energy bound on surprise; (ii) this entails acting on the environment to avoid surprises, which (iii) rests on making Bayesian inferences about the world.¨

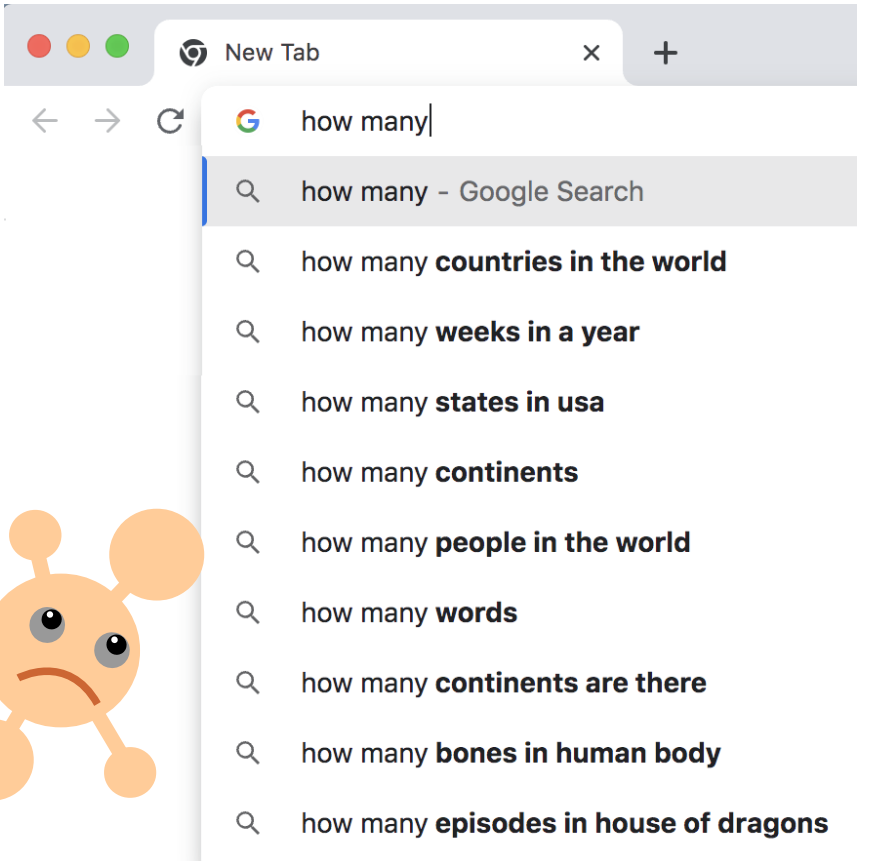
Prof. Karl Friston



Thomas Parr, Giovanni Pezzulo, Karl J. Friston:
Active Inference: The Free Energy Principle in Mind, Brain, and Behavior, 2022
https://direct.mit.edu/books/oa-monograph/5299/Active-InferenceThe-Free-Energy-Principle-in-Mind
Slide content source:
https://www.kaggle.com/code/charel/learn-by-example-active-inference-in-the-brain-2

К максимизации предсказуемости – через минимизацию неопределенности?

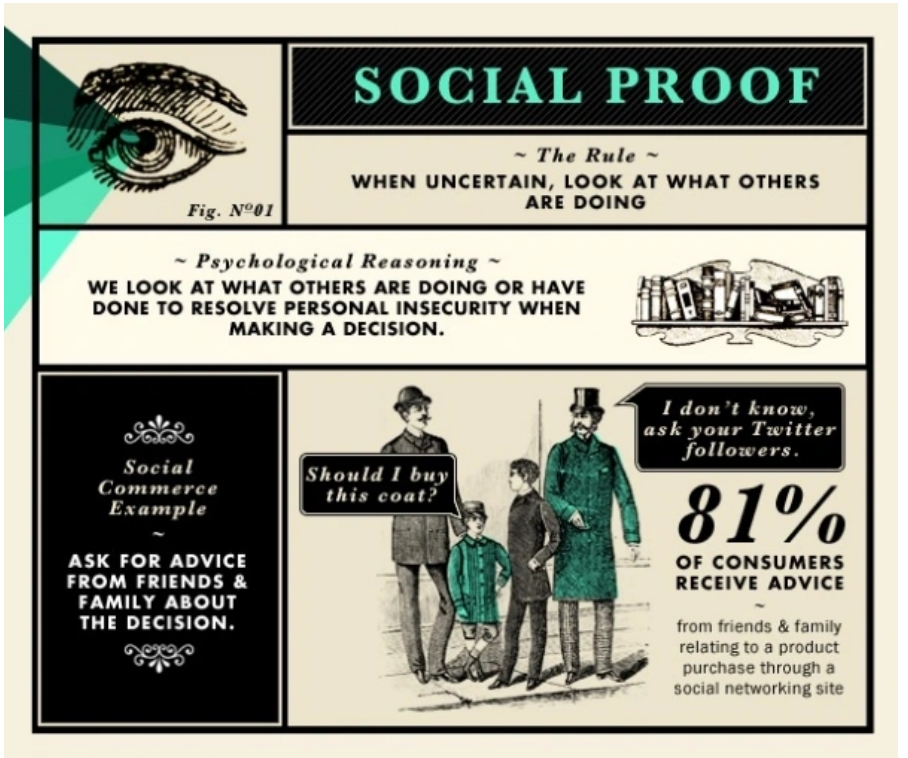




https://aclanthology.org/2022.emnlp-main.239/
https://arxiv.org/pdf/2303.02427.pdf

Robert Cialdini: Social Proof

Social proof is a psychological and social phenomenon wherein people copy the actions of others in choosing how to behave in a given situation. The term was coined by Robert Cialdini in his 1984 book Influence: Science and Practice, and the concept is also known as informational social influence.



https://www.amazon.com/Influence-Practice-Robert-B-Cialdini/dp/0205609996

Владимир Лефевр: Самоосознание как социальная рефлексия

линия. В лидере оказываются совмещенными оба начала – выполнение “рядовых” трудовых операций и специфической организующей деятельности.

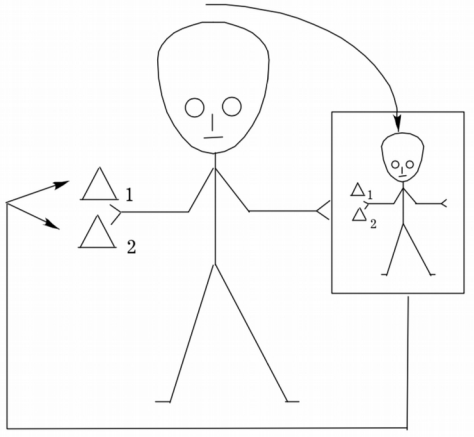
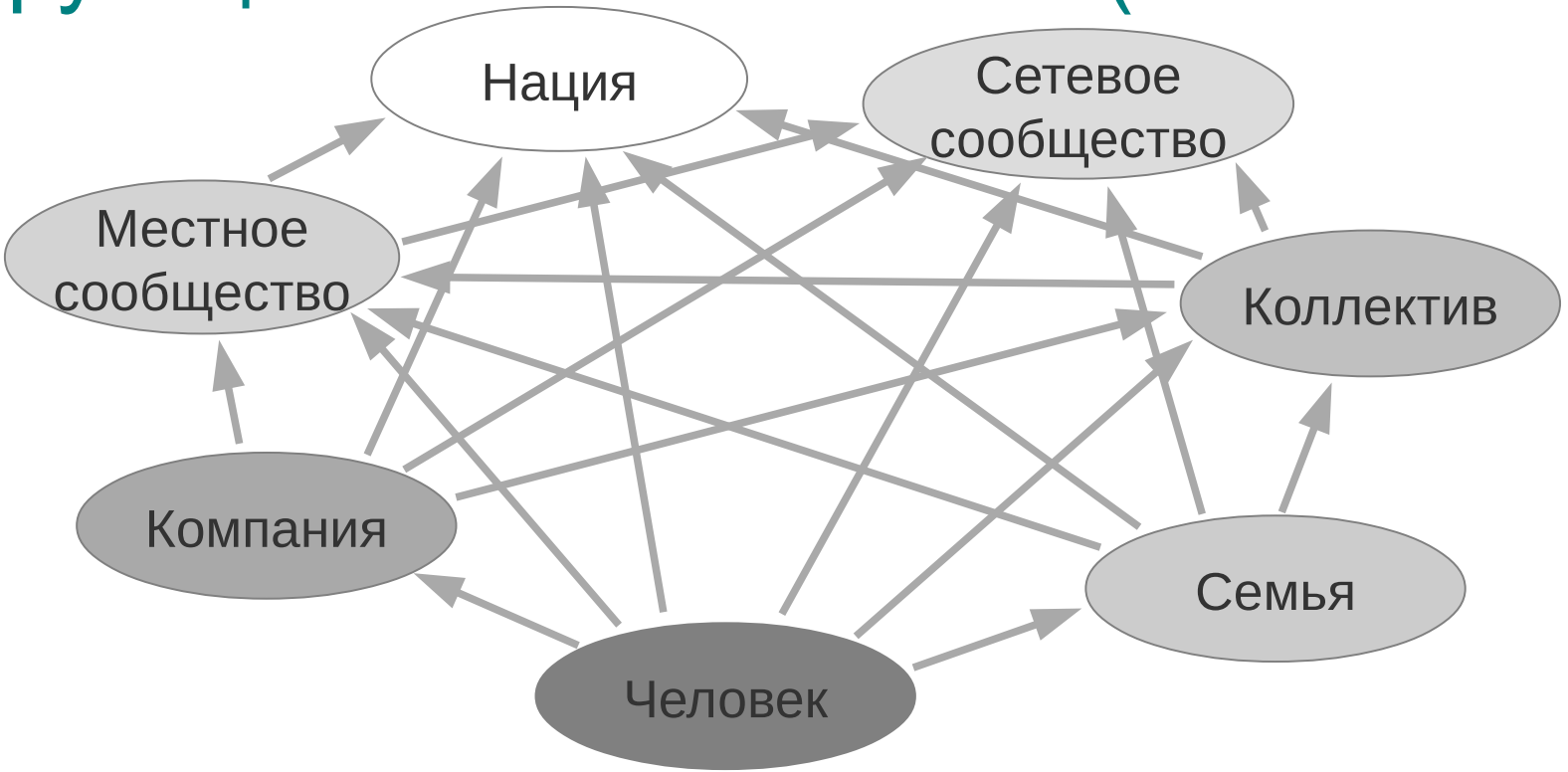


Рис. 9

Впервые индивидуальная деятельность оказывается *организованной*. Механизм, который раньше действовал в масштабе коллектива, переходит в индивидуаллную деятельность. Лидер превращается в *саморефлексивную систему*.

https://www.litres.ru/book/vladimir-lefevr/chto-takoe-odushevlennost-66216968/

Модель социума как гетерархия функциональных систем (П.К.Анохин)



Евгений Витяев: Сознание как логически непротиворечивая прогностическая модель реальности (1997-2017)
https://www.researchgate.net/publication/322977829_Soznanie_kak_logiceski_neprotivorecivaa_prognosticeskaa_model_realnosti
Anton Kolonin, Evgenii Vityaev, Yuriy Orlov: Cognitive Architecture of Collective Intelligence Based on Social Evidence (2016)
https://www.sciencedirect.com/science/article/pii/S1877050916317239

Все вместе

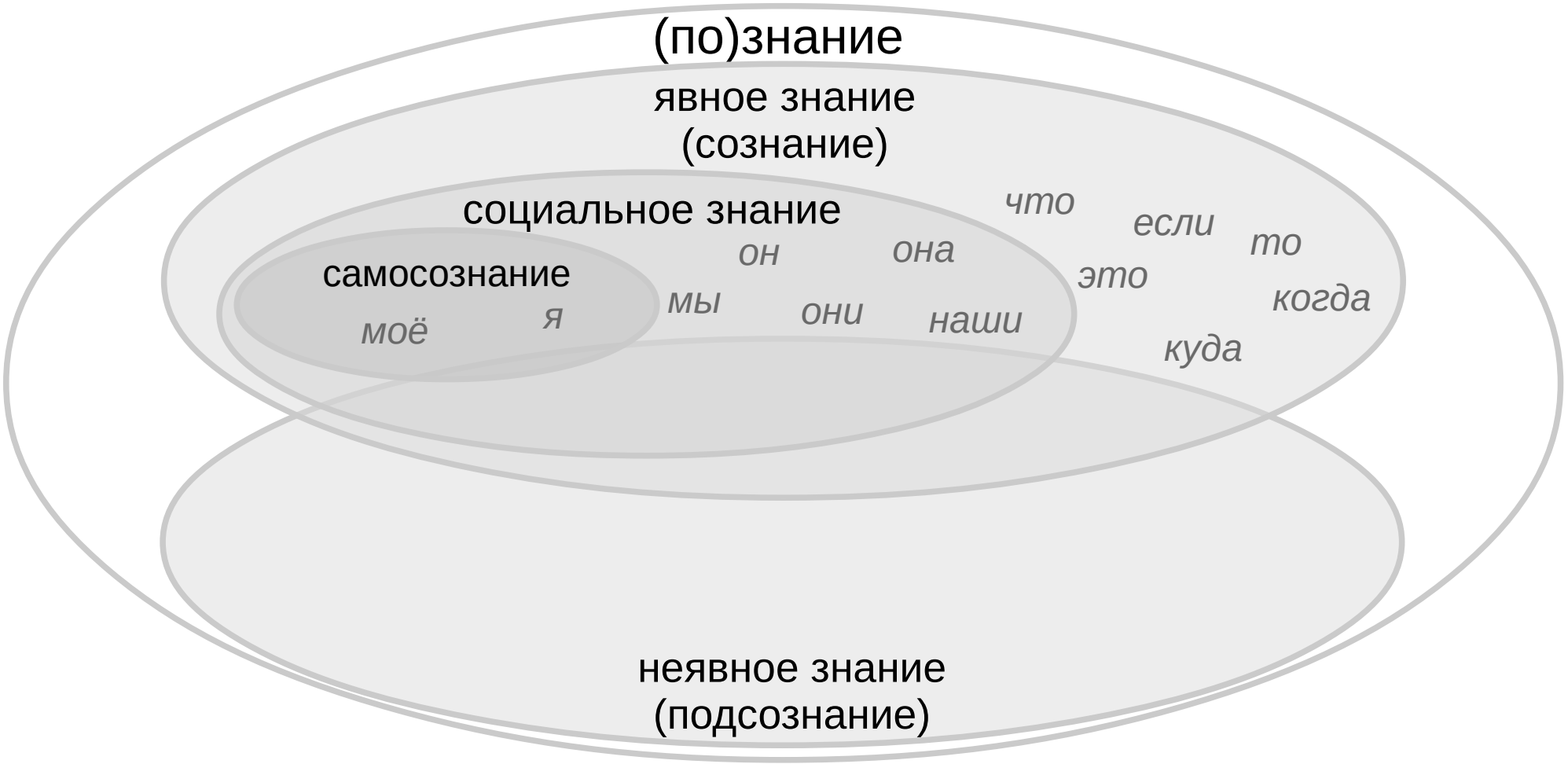
1. Минимизация потребления ресурсов (меньше думаем, быстрее принимаем решения) на рассмотрение различных гипотез и аспектов (принцип “ограниченных ресурсов”)

2. Максимизация вероятности ожидаемой позитивной оценки за свои действия и суждения (принцип “свободной энергии”)

3. Использование “мудрости толпы” в качестве как мерила для самооценки, так и для принятия решения в случае неопределённости (принцип “социального доказательства”)

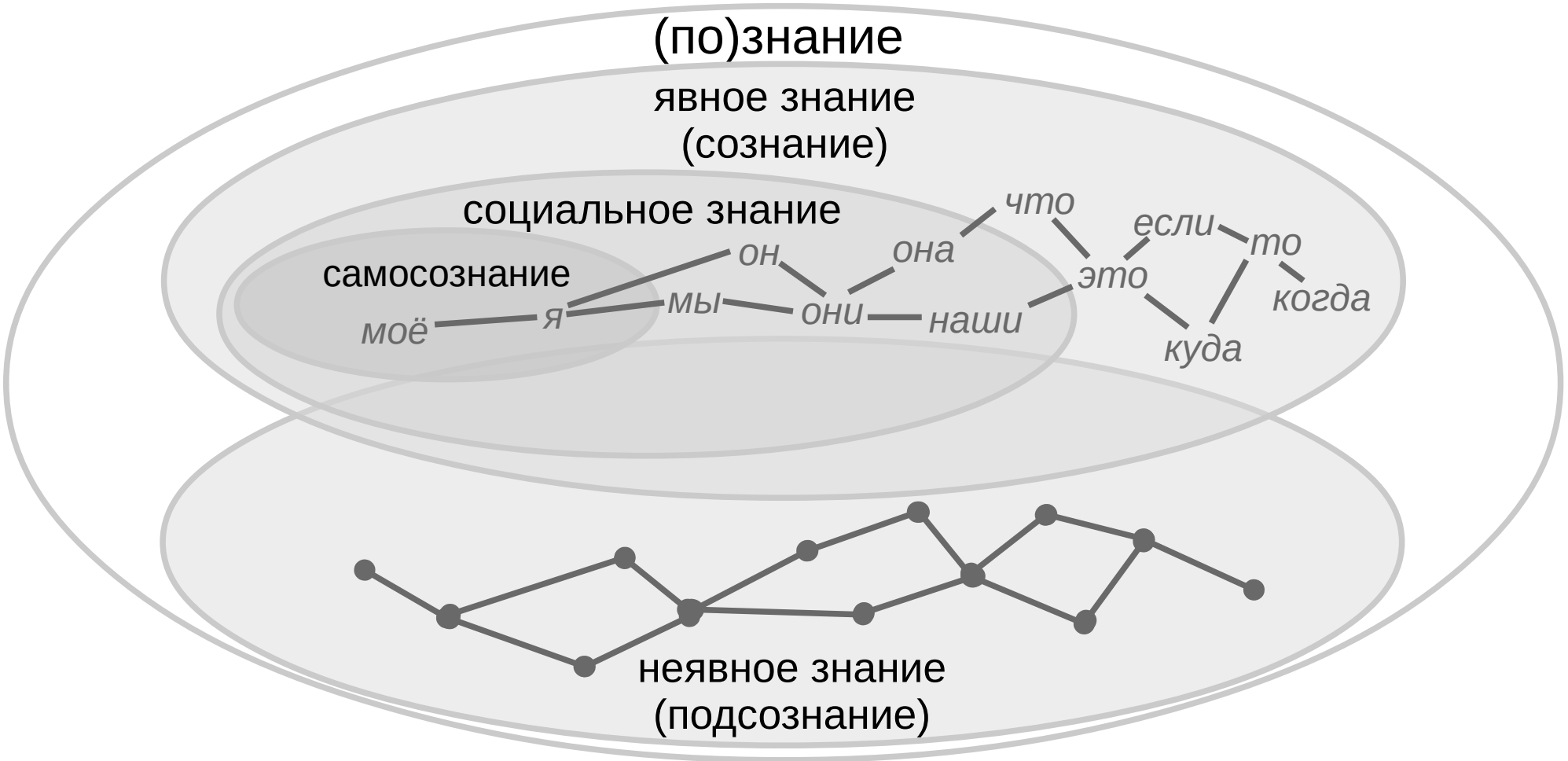
4. Социум как динамическая гетерархия “рефлексивных” “функциональных систем”.

СПознание – что в слове этом!?



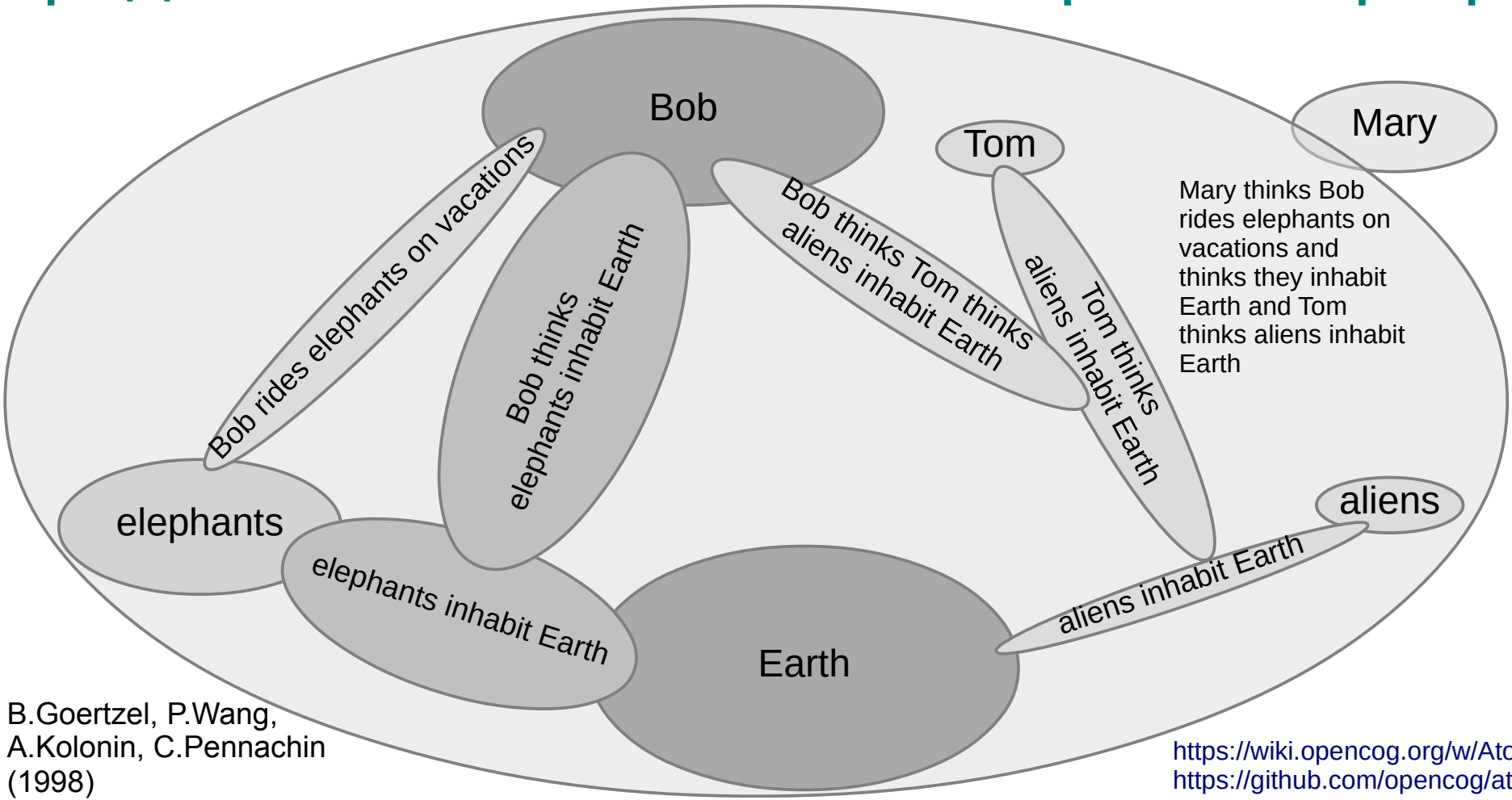
https://golos.id/ru--tekhnologii/@aigents/opredeleniya-ii

СПознание – что в слове этом!?



https://golos.id/ru--tekhnologii/@aigents/opredeleniya-ii

Представление знаний в гипер-мета-графах

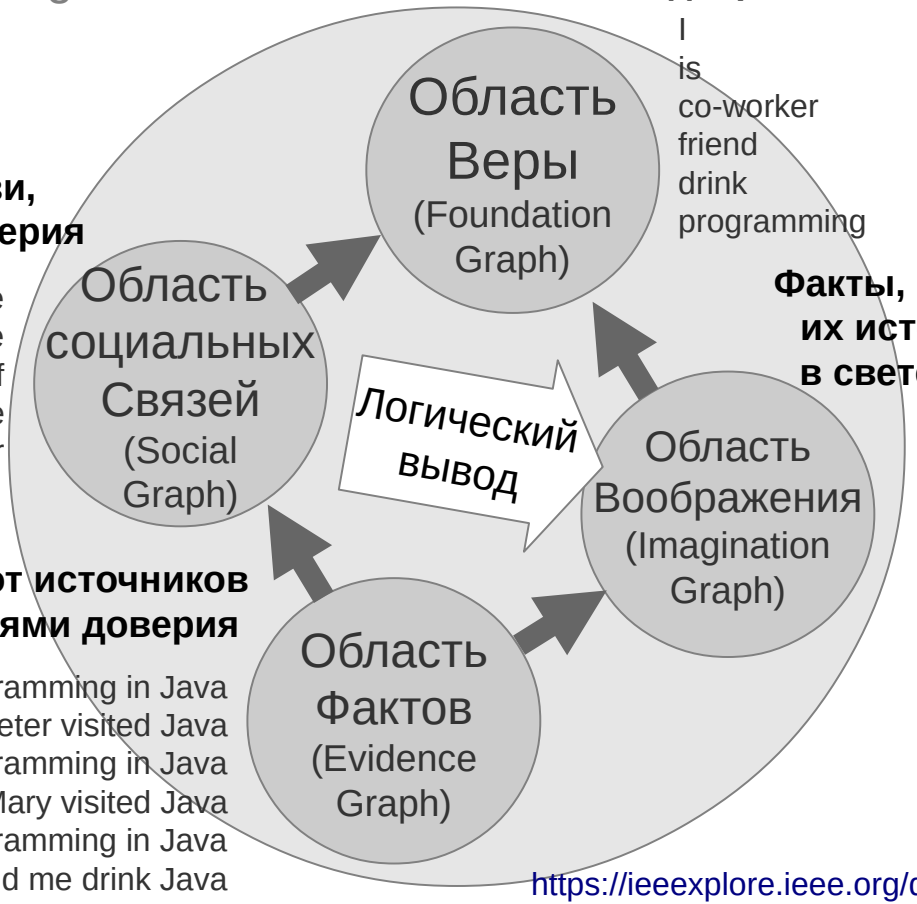


B.Goertzel, P.Wang, A.Kolonin, C.Pennachin (1998)

https://wiki.opencog.org/w/AtomSpace
https://github.com/opencog/atomspace

Социально-доказательная модель сознания

Social evidence based cognitive model



Не подвергаемые сомнению истины

I
is
co-worker
friend
drink
programming

Социальные связи, с учетом уровня доверия

Mary is friend of mine
Peter is friend of mine
Bob is co-worker of mine
Bob is friend of Peter

Факты, взвешенные на основе их источников и оценщиков, в свете непреложных истин

Java
(9)
Java coffee
(4)
Drink Java coffee
(4)
Program in Java language
(3)
Java island
(2)
Visit Java island
(2)

Факты, полученные от источников с различными уровнями доверия

1997: Bob is programming in Java
1998: Peter visited Java
1999: I am programming in Java
2011: Mary visited Java
2012: Peter is programming in Java
Yesterday: Peter and me drink Java
Today: Mary and me drink Java

https://ieeexplore.ieee.org/document/7361869
https://www.sciencedirect.com/science/article/pii/S1877050916317239
https://link.springer.com/chapter/10.1007/978-3-319-97676-1_10

Imagination and Action driven by Believed Evidence

i – person of consideration
j – concept or action of consideration
l – person's belief item (of foundation graph of size *L)*
B_{il} – person's *i* mental attachment to *l*

$$P^B_{ij} = \Sigma_{l=1,L}(E^B_{ijl} * B_{il})$$ *(why we think so)*

E^B_{ik} – concept *j* agreement or compatibility with *l* in mind of *i*
P_{ij} – concept *j* power for *i*

$$A^B_{ij} = \Sigma_{l=1,L}(C^B_{ijl} * B_{il})$$ *(why we act so)*

C^B_{ik} – action *j* agreement or compatibility with *l* in view of *i*
A_{ij} – action *j* preference for *i*

Imagination and Action driven by Social Evidence

i – person of consideration
j – concept or action of consideration
k – person's correspondent (of social graph of size *K)*
S_{ik} – person's *i* social bind to *k*

recursion

$$P^S_{ij} = \Sigma_{k=1,K}(E^S_{ijk} * S_{ik})$$ *(why we think so)*

E^S_{ik} – concept *j* expression or confirmation by *k* in view of *i*
P_{ij} – concept *j* power for *i*

$$A^S_{ij} = \Sigma_{k=1,K}(C^S_{ijk} * S_{ik})$$ *(why we act so)*

C^S_{ik} – action *j* acceptance or approval by *k* in view of *i*
A_{ij} – action *j* preference for *i*

Imagination driven by Believed Social Evidence

i – person of consideration
j – concept or action of consideration (exposed evidence)
l – person's belief item (of foundation graph of size *L* - personal preference base)
k – person's correspondent (of social graph of size *K* - social reference base)
B_{il} – person's *i* mental attachment to *l* (personal preference)
S_{ik} – person's *i* social bind to *k* (social reference)

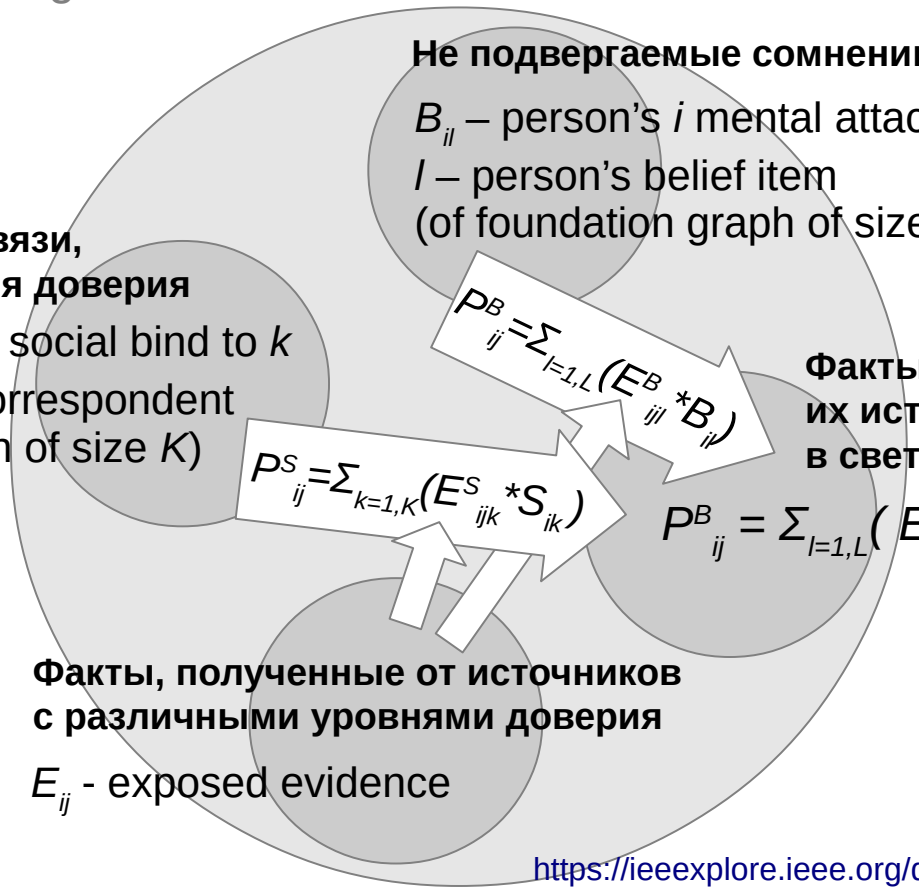
recursion

$$P^B_{ij} = \Sigma_{l=1,L}(E^B_{ijl} * B_{il}) * \Sigma_{k=1,K}(E^S_{ijk} * S_{ik})$$ *(why we think so)*

E^B_{ik} – concept *j* agreement or compatibility with *l* in mind of *I* (believed evidence)
E^S_{ik} – concept *j* expression or confirmation by *k* in view of *I* (social evidence)
P_{ij} – concept *j* power for *I* (personal evidence)

Социально-доказательная модель сознания

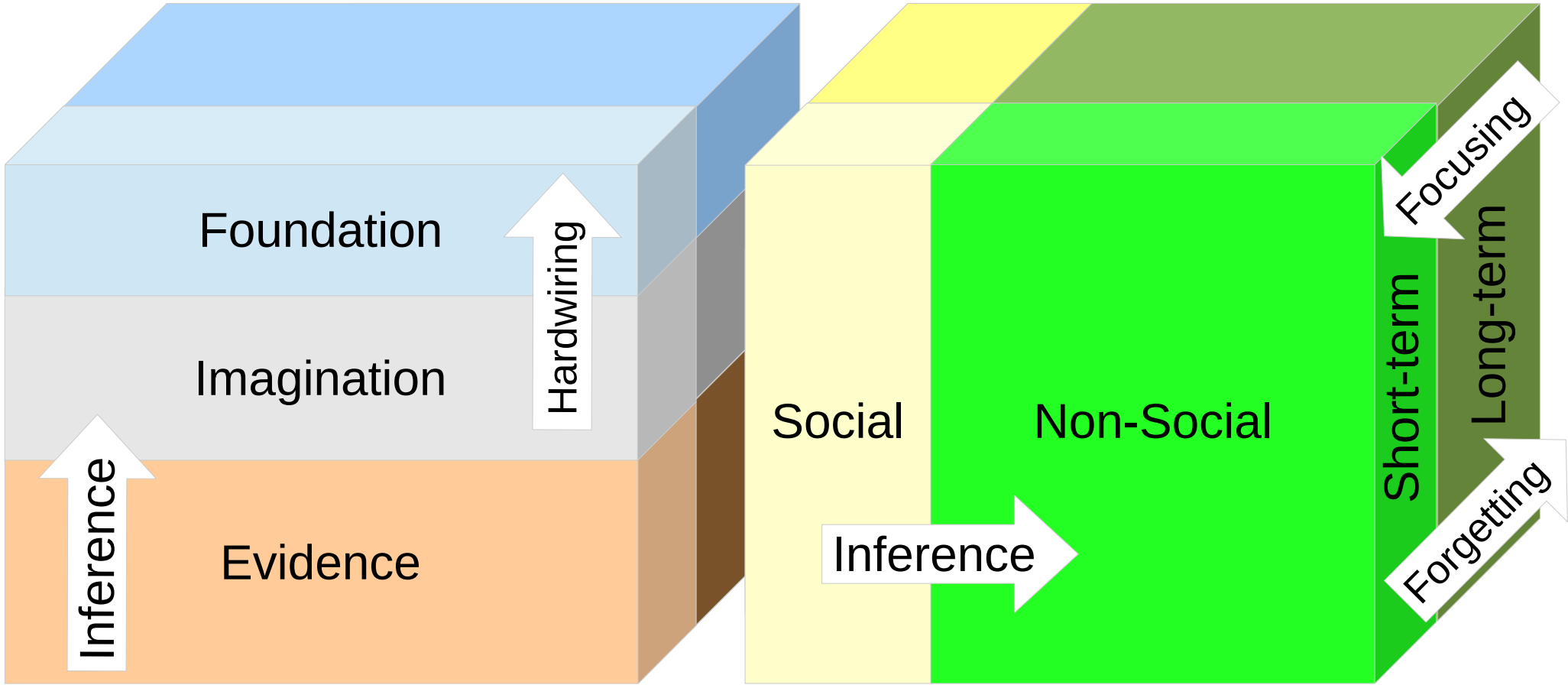
Social evidence based cognitive model



https://ieeexplore.ieee.org/document/7361869
https://www.sciencedirect.com/science/article/pii/S1877050916317239
https://link.springer.com/chapter/10.1007/978-3-319-97676-1_10

Социально-доказательная модель сознания

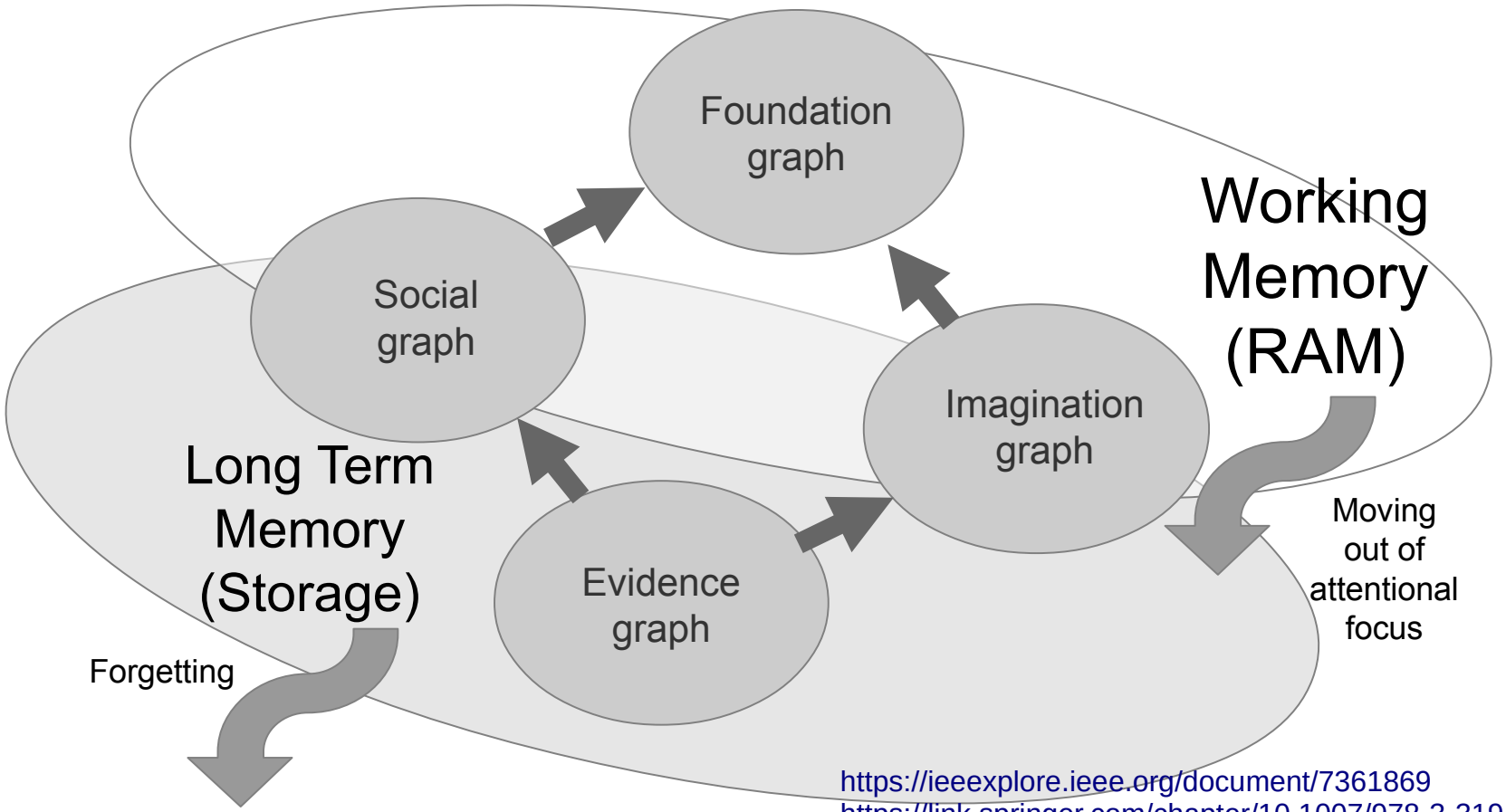
Temporally and socially scoped evidence, supported with short-term and long-term memory capabilities.



https://www.sciencedirect.com/science/article/pii/S1877050916317239

Social evidence-based resource-constrained cognitive model

Applying resource constraints: short-term and long-term memories.

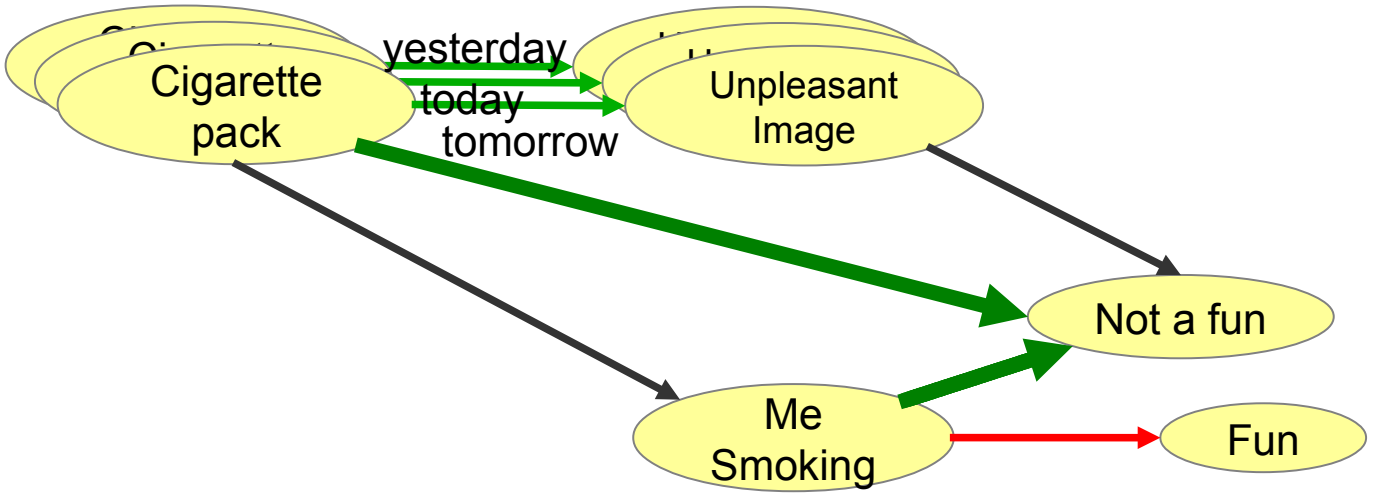


https://ieeexplore.ieee.org/document/7361869
https://link.springer.com/chapter/10.1007/978-3-319-97676-1_10

Social evidence-based cognitive model

Understanding applied belief change

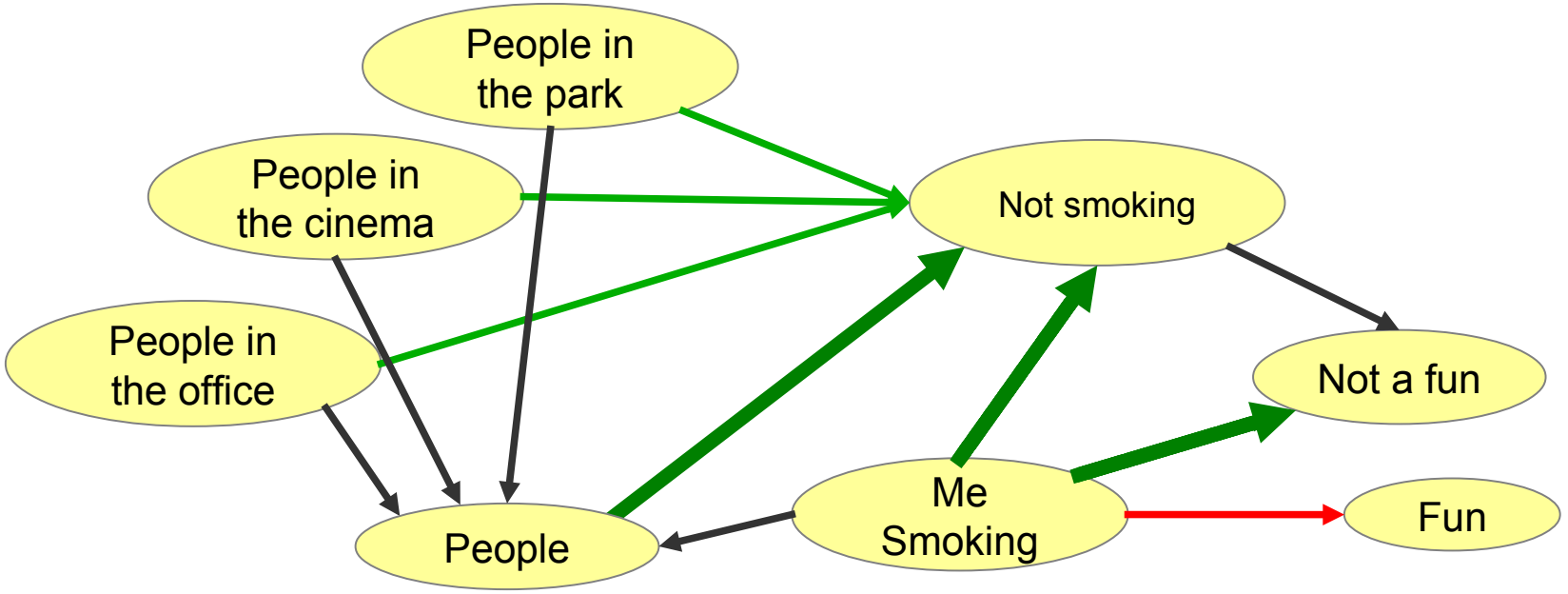
Method: “Redundant temporal evidence”



Social evidence-based cognitive model

Understanding applied belief change

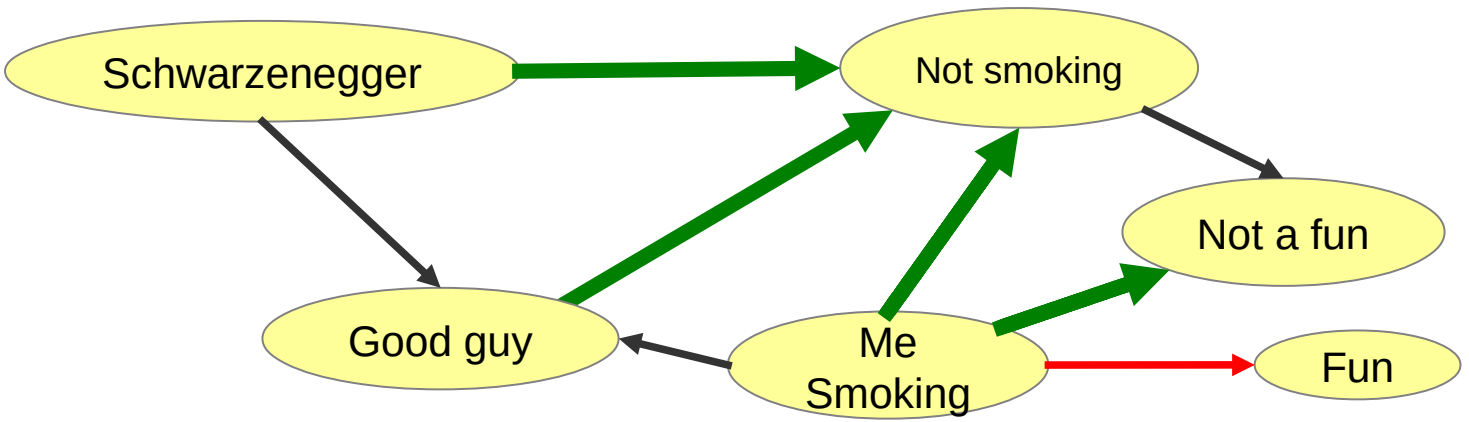
Method: “Redundant Social Evidence”



Social evidence-based cognitive model

Understanding applied belief change

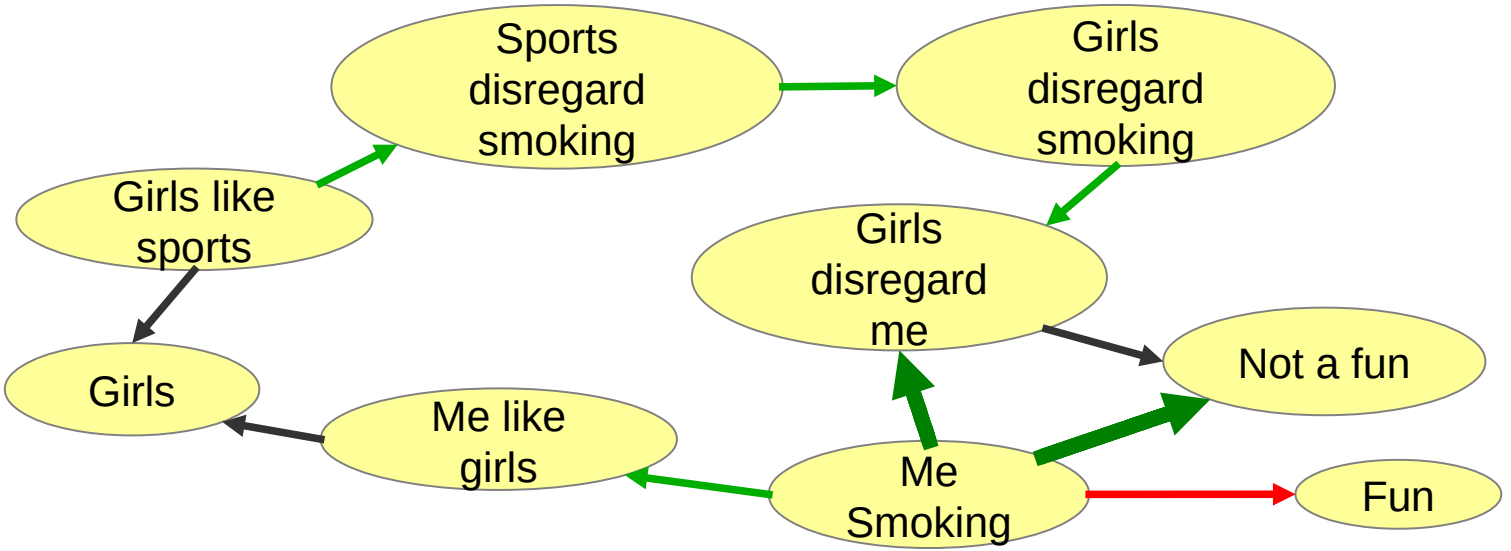
Method: “Highly valable social evidence”



Social evidence-based cognitive model

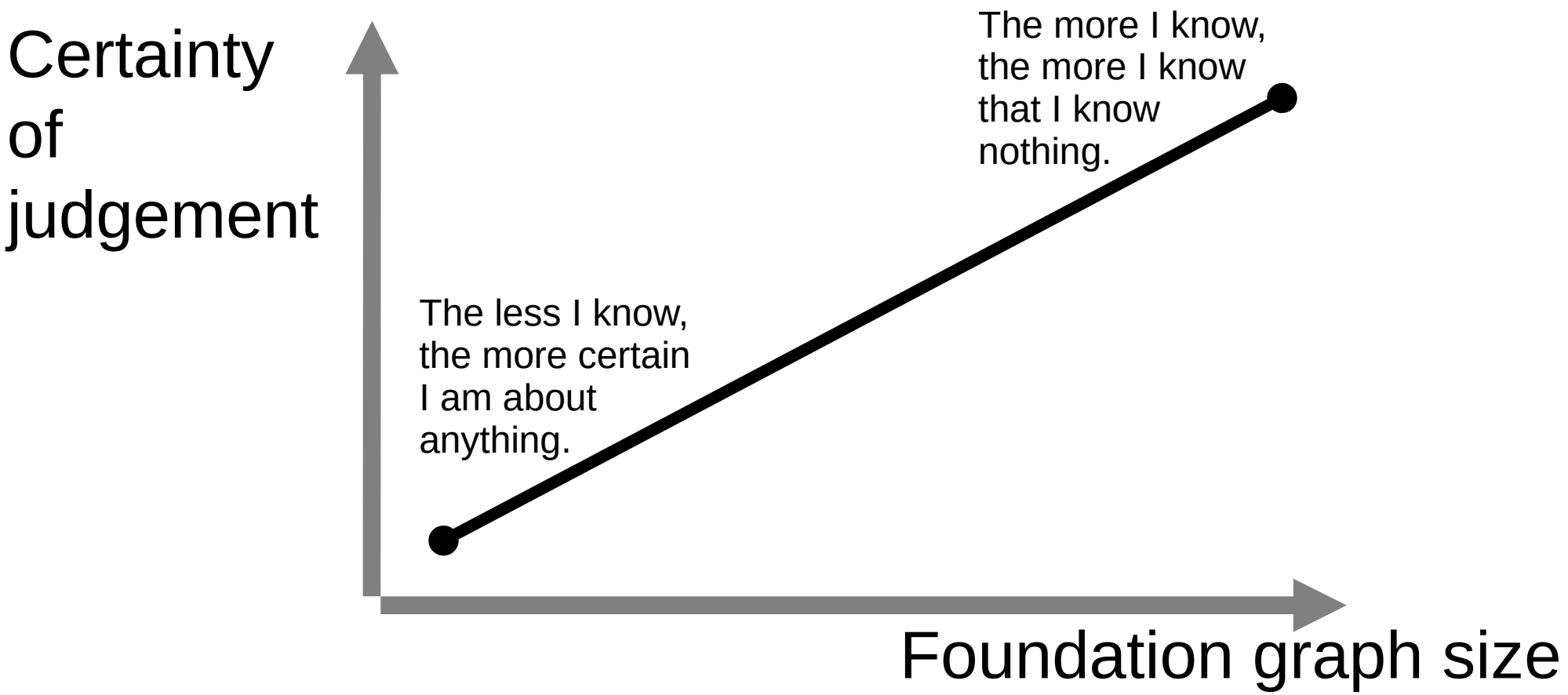
Understanding applied belief change

Method: “Implicit evidence injection”



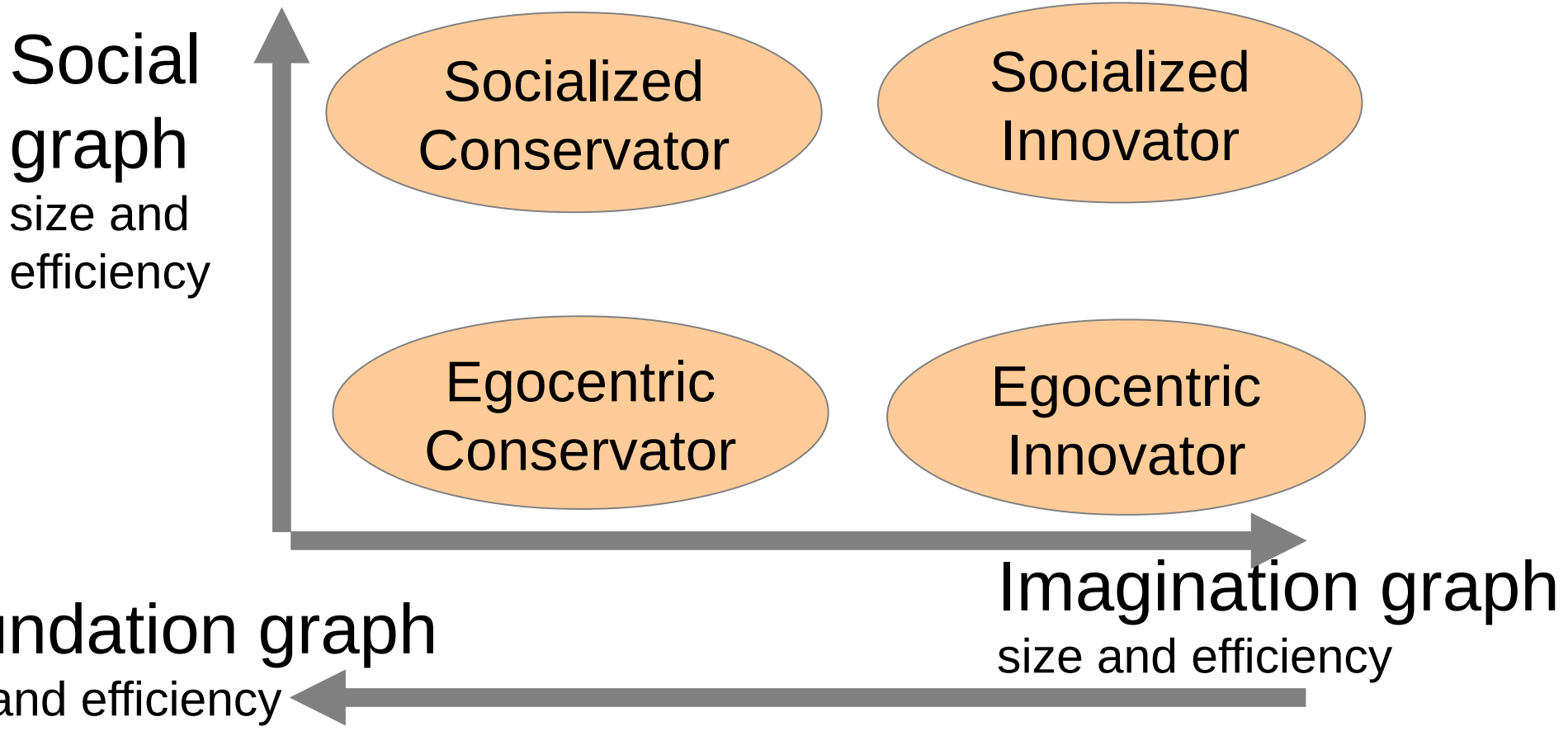
Social evidence-based resource-constrained cognitive model

Predicting individual behavior

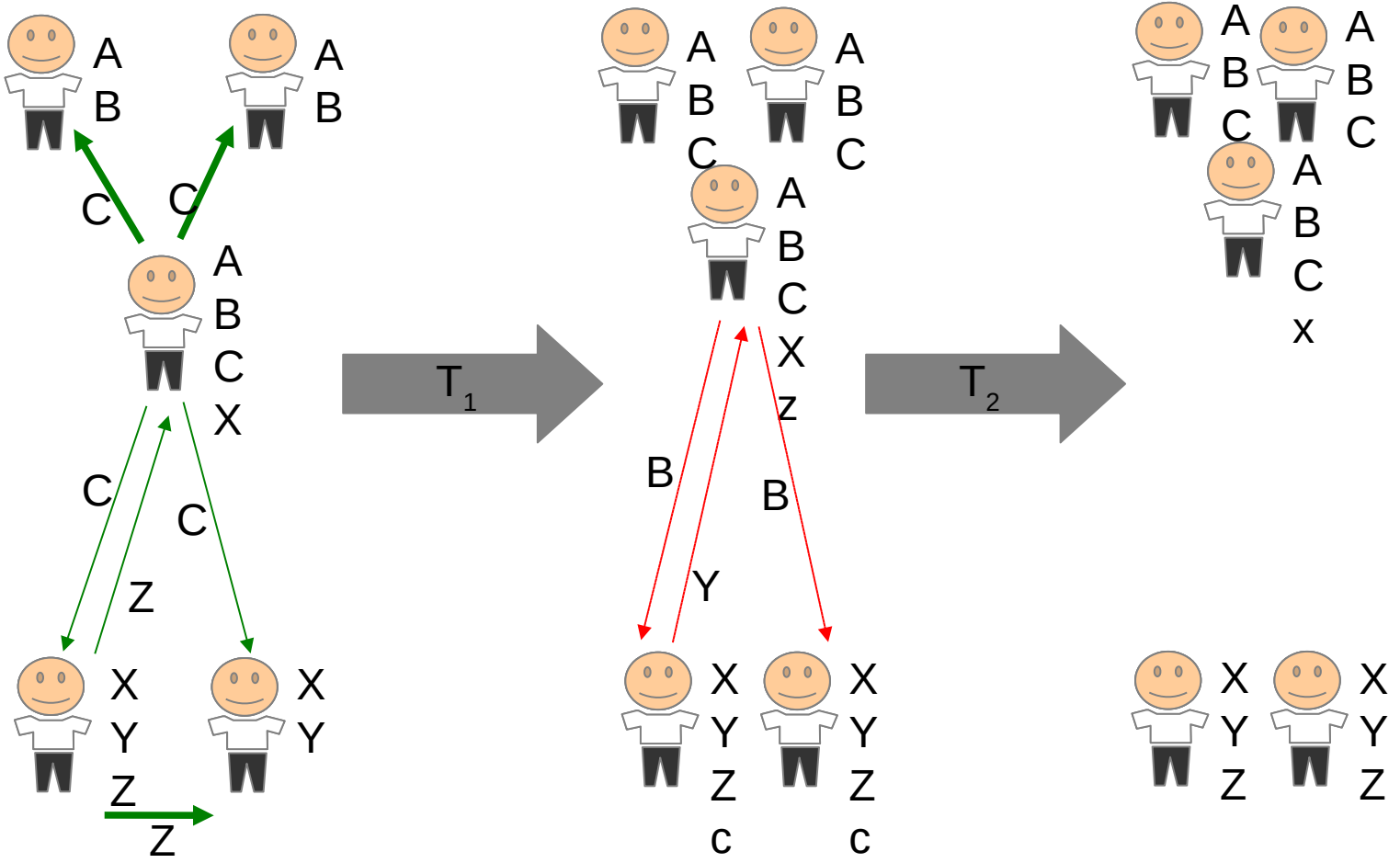


Social evidence-based resource-constrained cognitive model

Predicting individual behavior

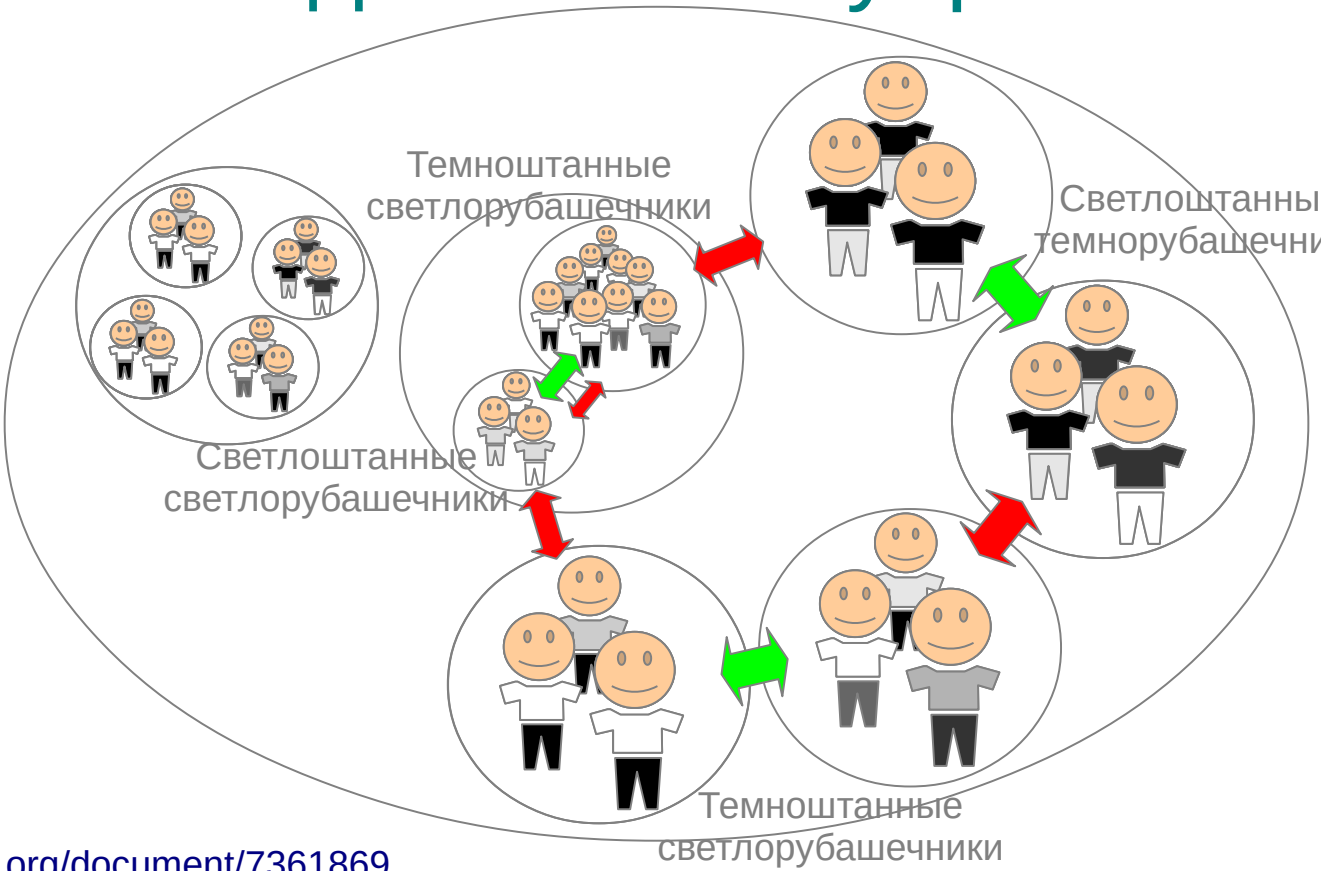


Эффект естественной социальной поляризации



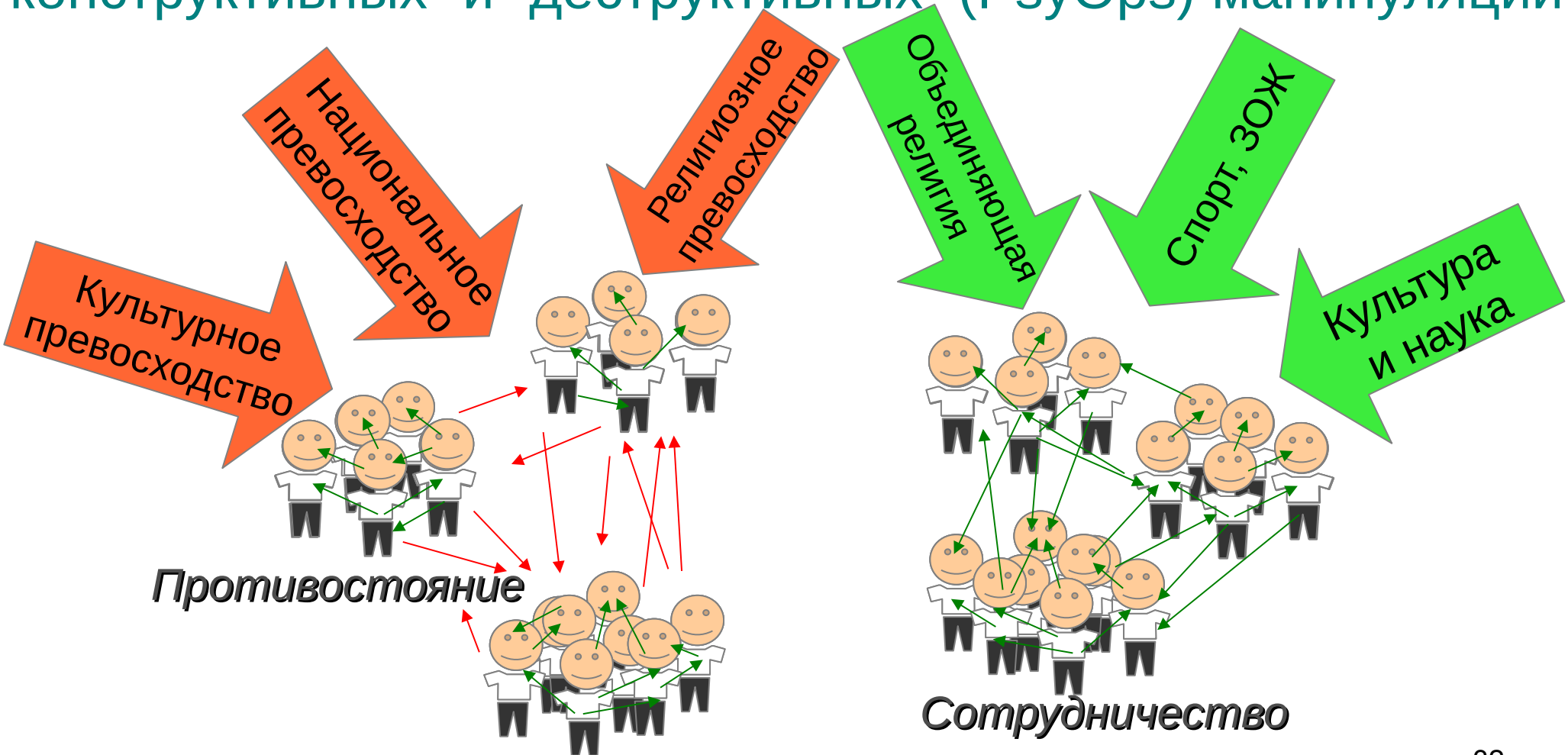
https://ieeexplore.ieee.org/document/7361869
https://www.sciencedirect.com/science/article/pii/S1877050916317239

Возможность моделирования и предсказания социальной динамики и управления ей

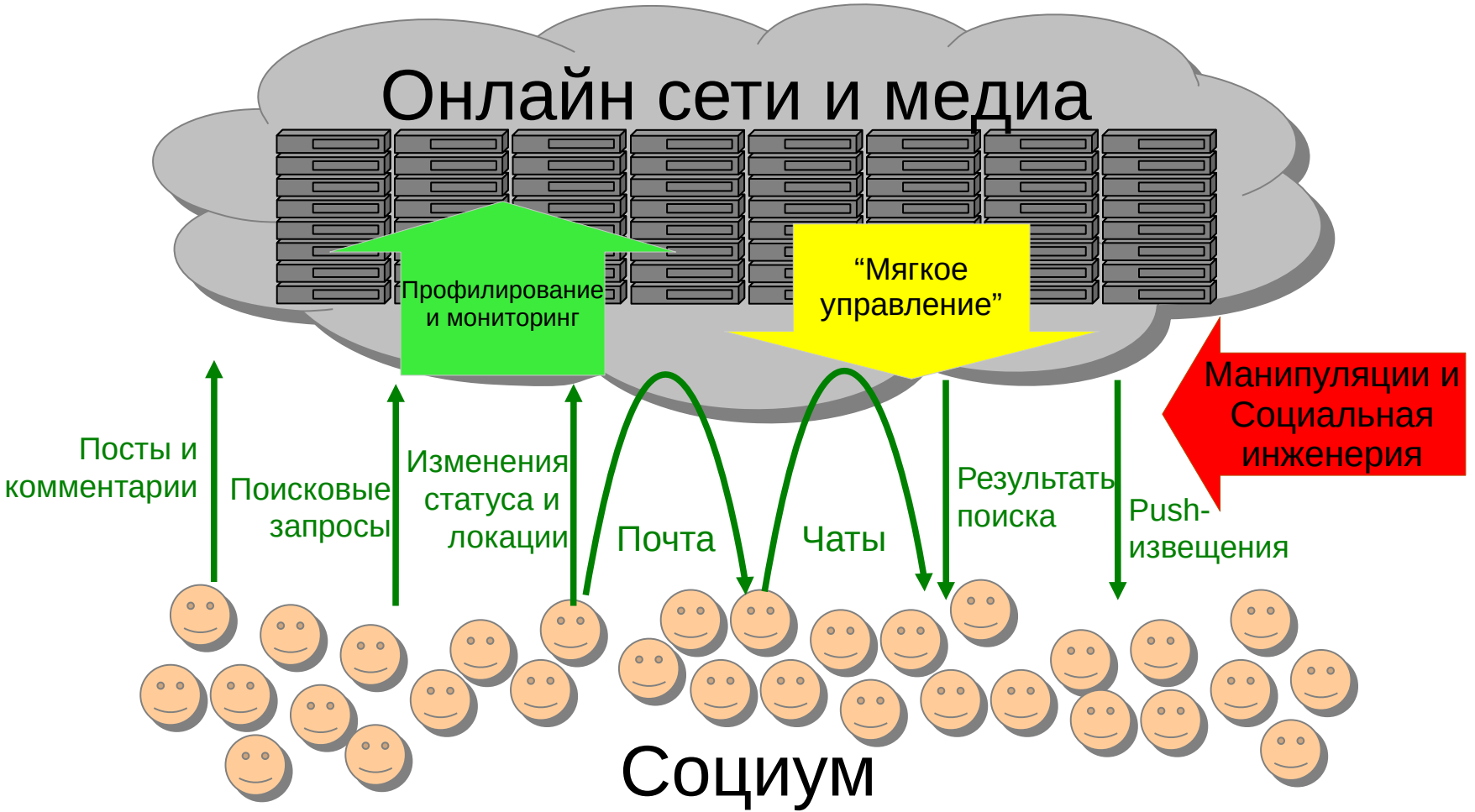


https://ieeexplore.ieee.org/document/7361869
https://www.sciencedirect.com/science/article/pii/S1877050916317239

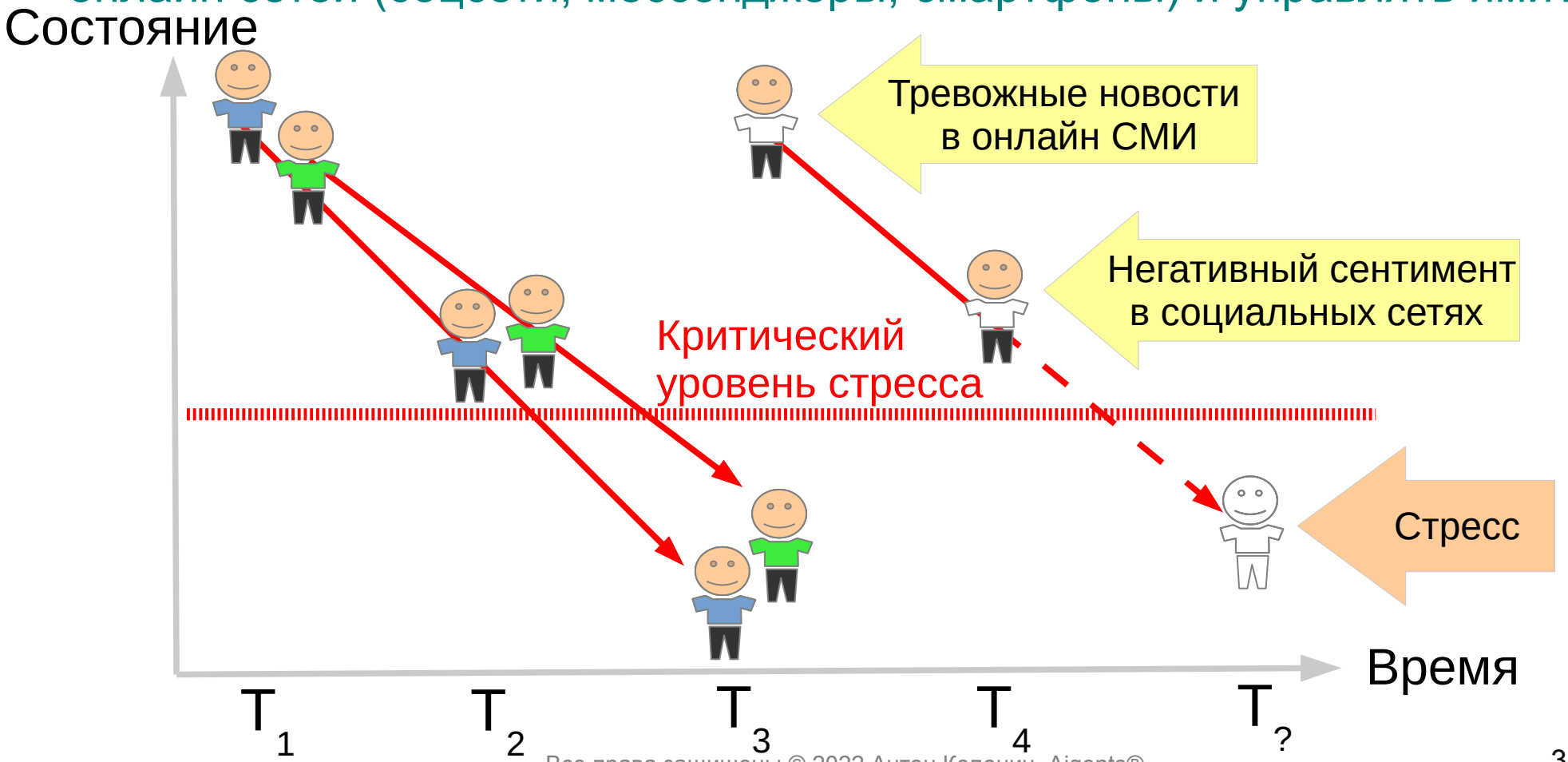
Особенности человеческой психики как основа для “конструктивных” и “деструктивных” (PsyOps) манипуляций



Онлайн сети и медиа – взаимодействие с социумом

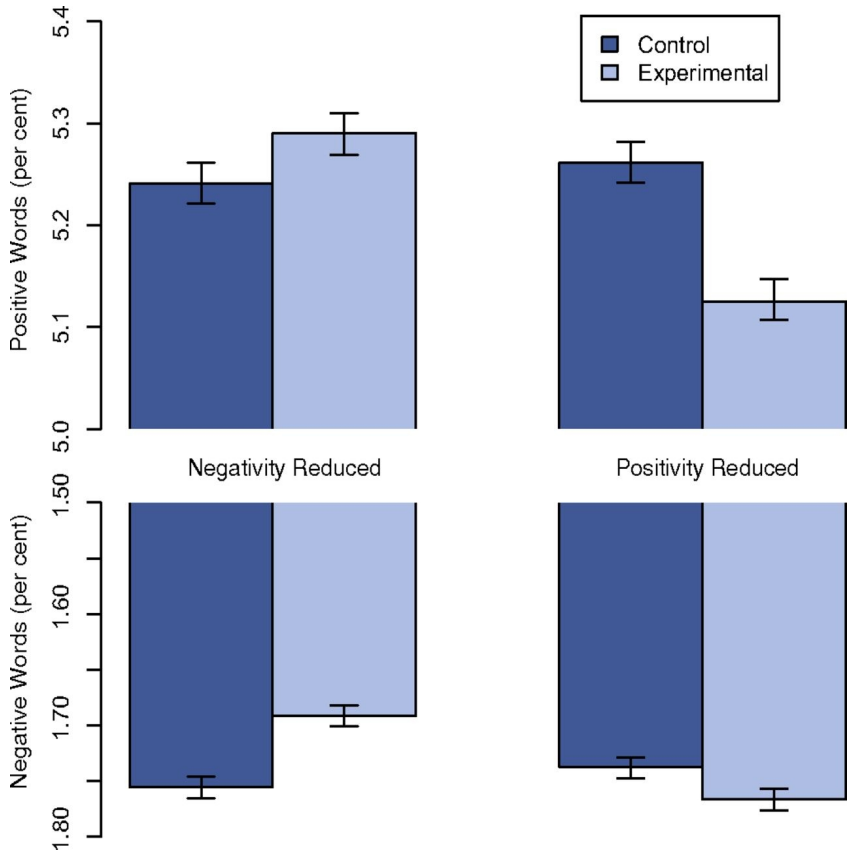


Можно ли с помощью машинного обучения предсказывать критические измененения в психоэмоциональном состоянии на основе данных из онлайн сетей (соцсети, мессенджеры, смартфоны) и управлять ими?



Как тональность текста влияет на поведение?

“Эксперимент Facebook”: Искуственно меняя “среднее настроение” в личной новостной ленте, можно синхронно “опускать” или “поднимать” настроение у миллиардов человек.



Изменение числа позитивных и негативных слов в сообщениях пользователей Facebook в экспериментальной и контрольной группе (передача негативного эмоционального контекста).

Kramer A.D., Guillory J.E., Hancock J.T. Experimental evidence of massive-scale emotional contagion through social networks. *PNAS* 2014;111(24):8788-8790. https://www.pnas.org/doi/10.1073/pnas.1320040111

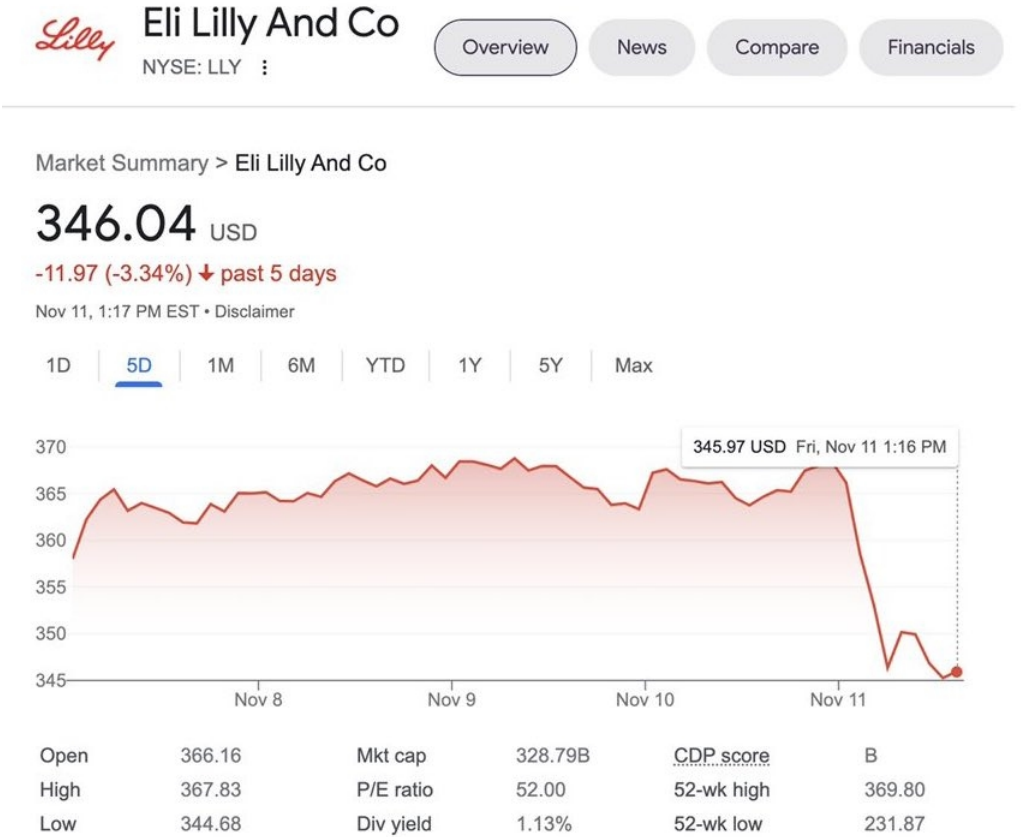
(слайд подготовлен Ю.Л.Орловым)

Как манипуляции в тексте влияют на поведение?

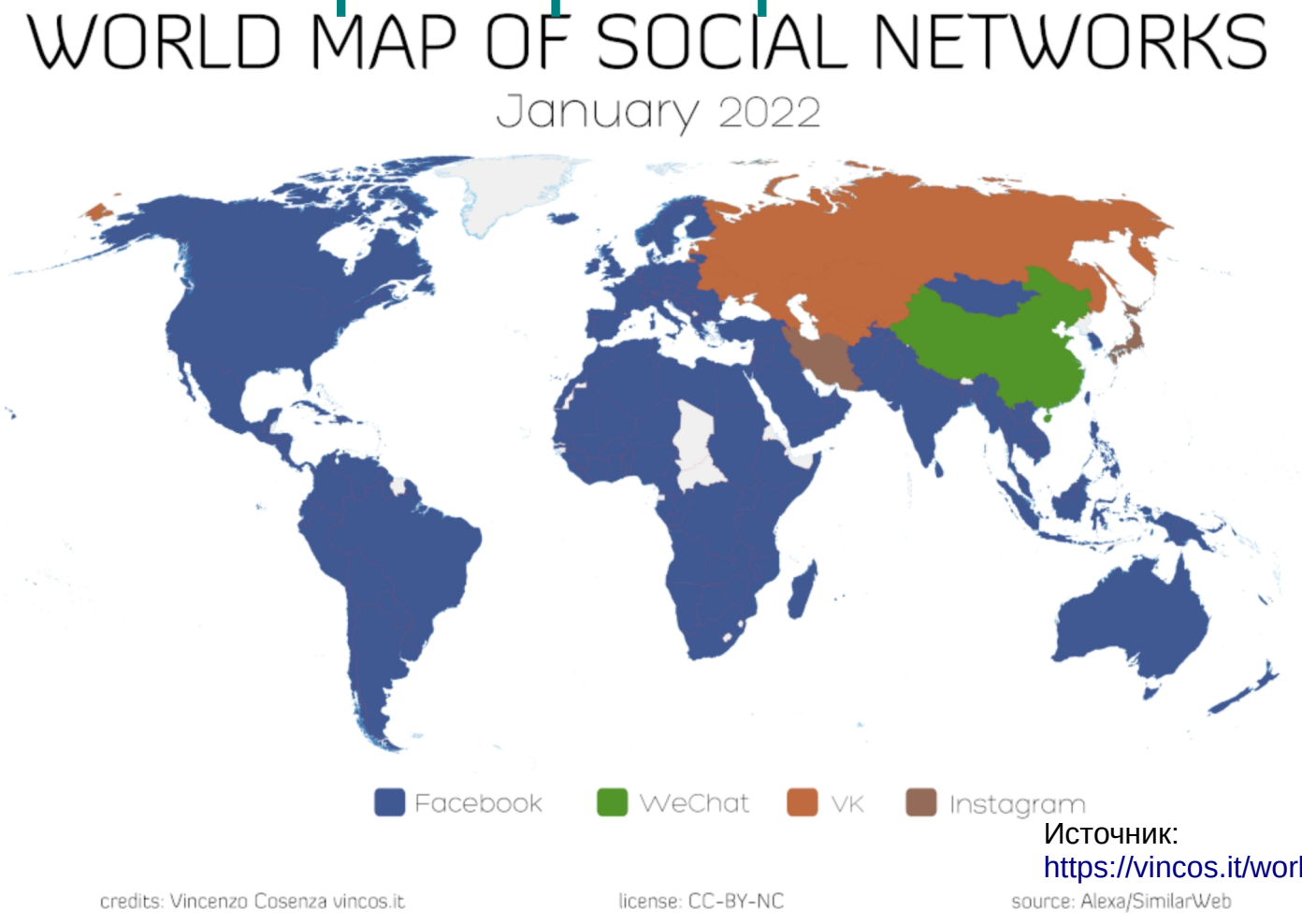
Одна фэйковая новость, распространяемая в онлайн СМИ в США в 2021 году в течении 48 минут привела к взлету и обрушению рынка криптовалют

Как манипуляции в тексте влияют на поведение?

Одна фэйковая новость, опубликованная в Twitter 11 ноября 2022 года обрушила акции компании по производству инсулина минимум на сутки



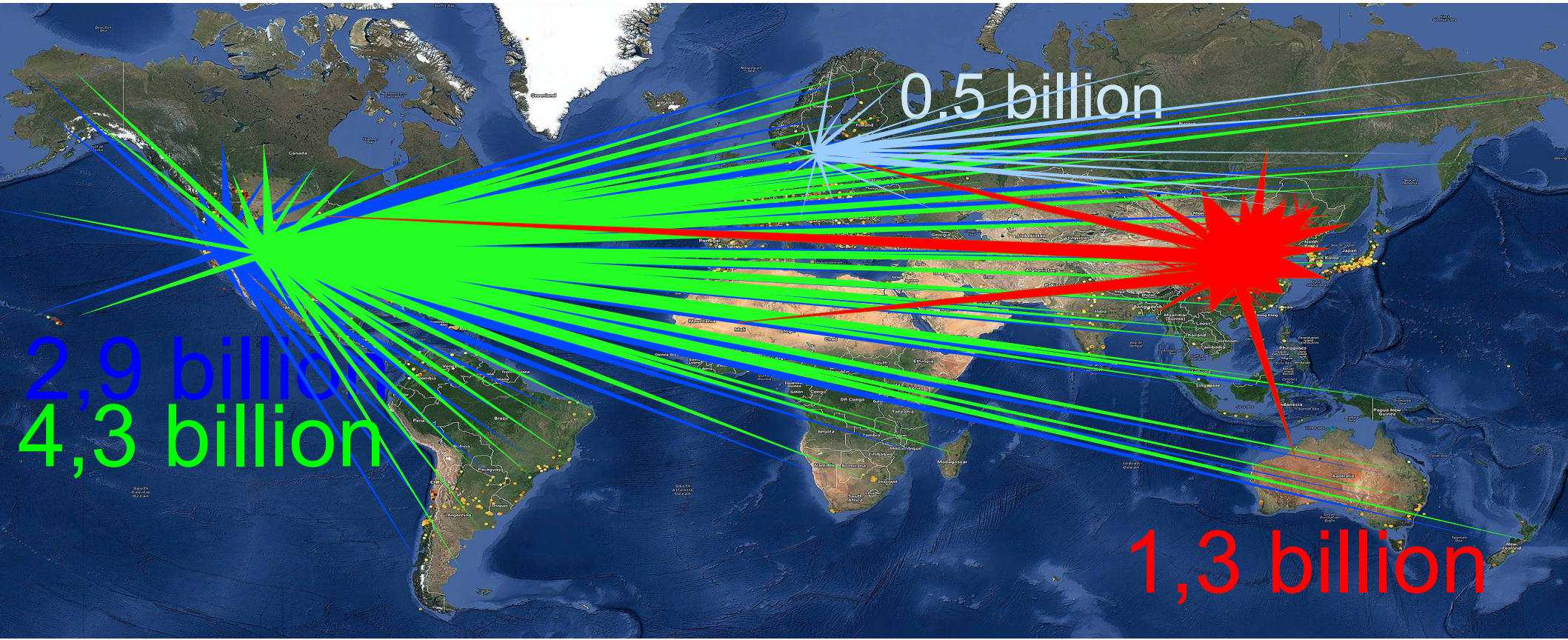
Масштабы распространения на карте



Источник:
https://vincos.it/world-map-of-social-networks/

Планетарное распространение онлайн сетей

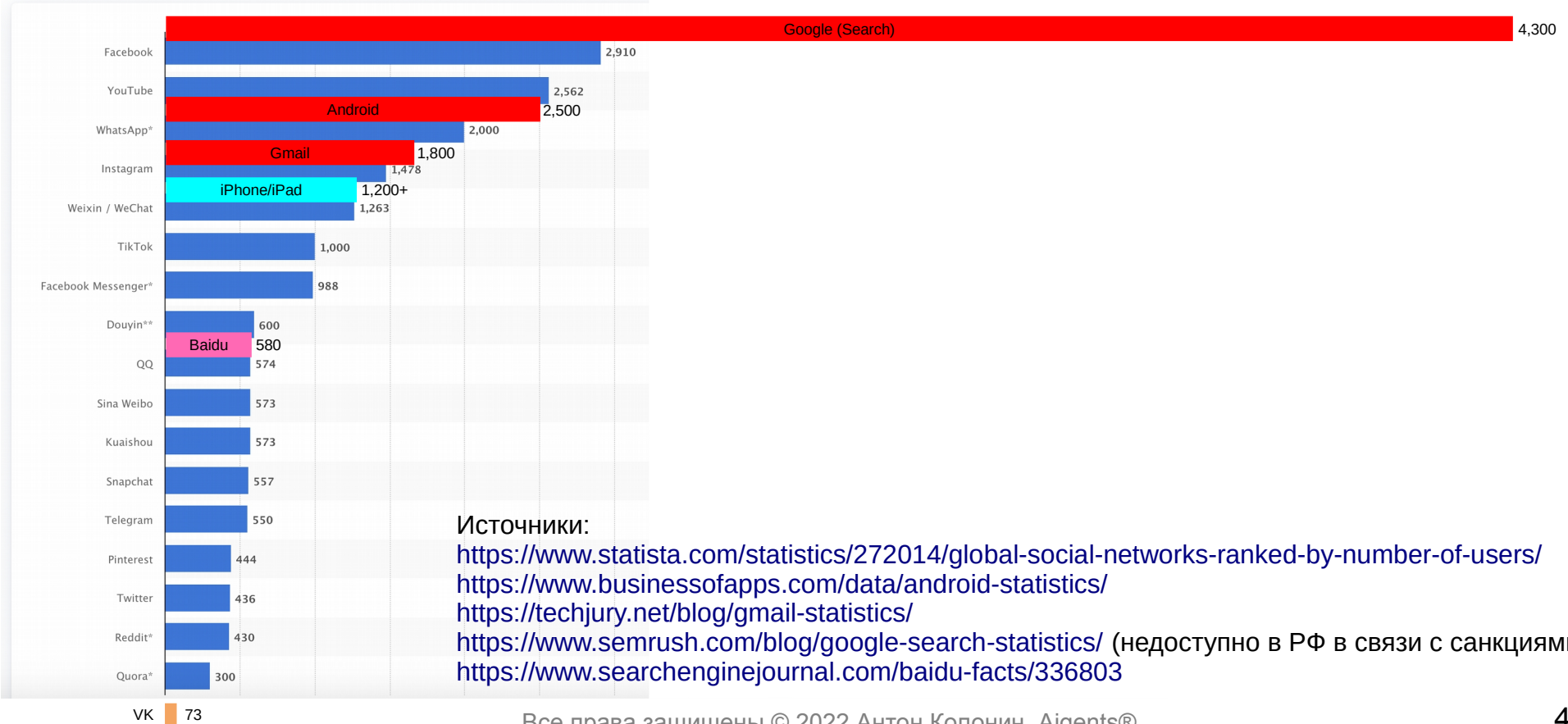
Google+Facebook, Telegram, TicTok+Baidu+WeChat



Масштабы распространения в миллионах

Most popular social networks worldwide as of January 2022, ranked by number of monthly active users

(in millions)

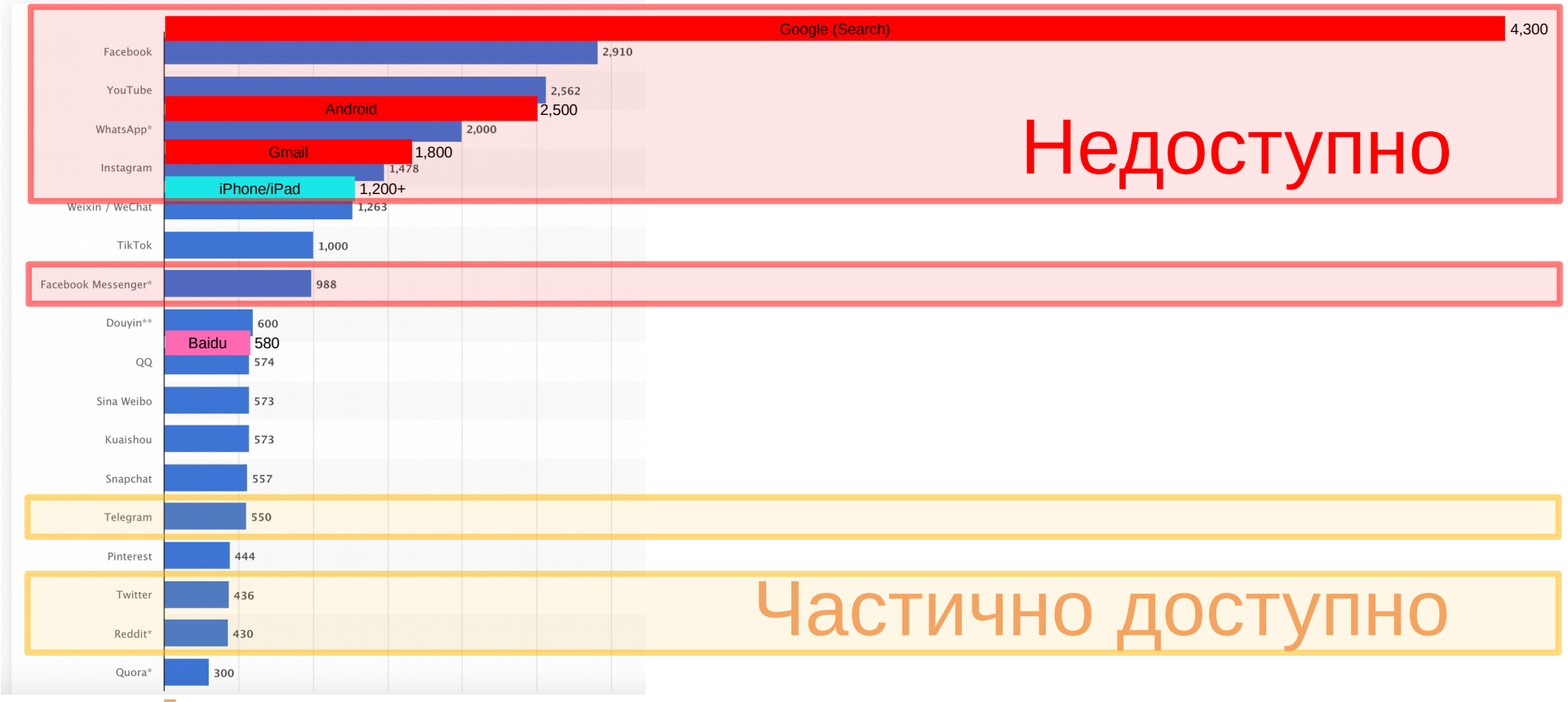


Источники:
https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/
https://www.businessofapps.com/data/android-statistics/
https://techjury.net/blog/gmail-statistics/
https://www.semrush.com/blog/google-search-statistics/ (недоступно в РФ в связи с санкциями)
https://www.searchenginejournal.com/baidu-facts/336803

Доступность для (легальных) исследований

Most popular social networks worldwide as of January 2022, ranked by number of monthly active users

(in millions)



Социальные медиа (сети) как “поле боя”?

U.S. Army: PSYCHOLOGICAL OPERATIONS LEADERS PLANNING GUIDE
https://irp.fas.org/doddir/army/psyopplan.pdf

U.S. Army: Psychological Operations
https://irp.fas.org/doddir/army/fm3-05-30.pdf

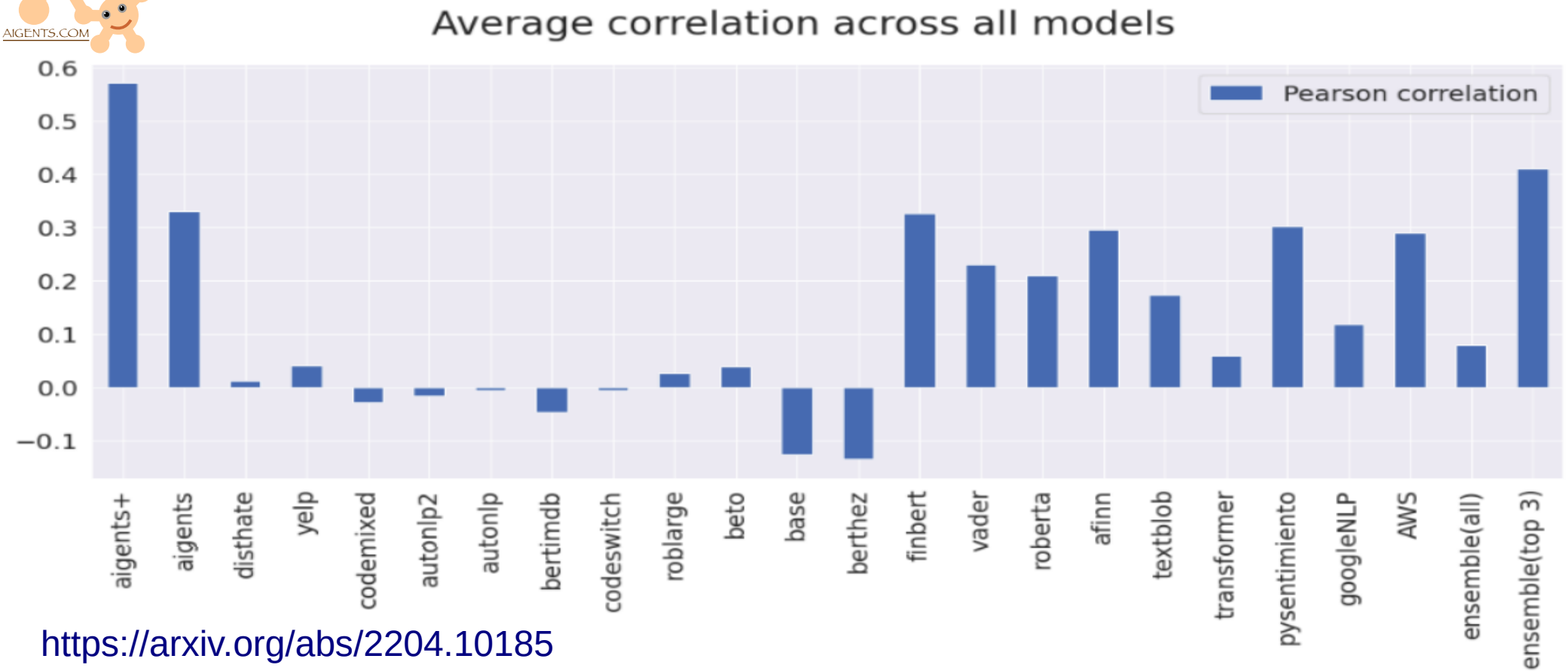
US Army PSYOP Book 1 - Psychological Operations Handbook
https://www.amazon.com/Army-PSYOP-Book-Psychological-Fundamentals/dp/1949117081

NATO: ALLIED JOINT DOCTRINE FOR
PSYCHOLOGICAL OPERATIONS
https://info.publicintelligence.net/NATO-PSYOPS.pdf

Royal Military College of Canada: Influence Techniques Using Social Media
https://cradpdf.drdc-rddc.gc.ca/PDFS/unc365/p807750_A1b.pdf

Как определять тональность текста?

Метрики: Тональность, Позитив, Негатив, Противоречивость



https://arxiv.org/abs/2204.10185

Что можно узнать из текста о психике социума?

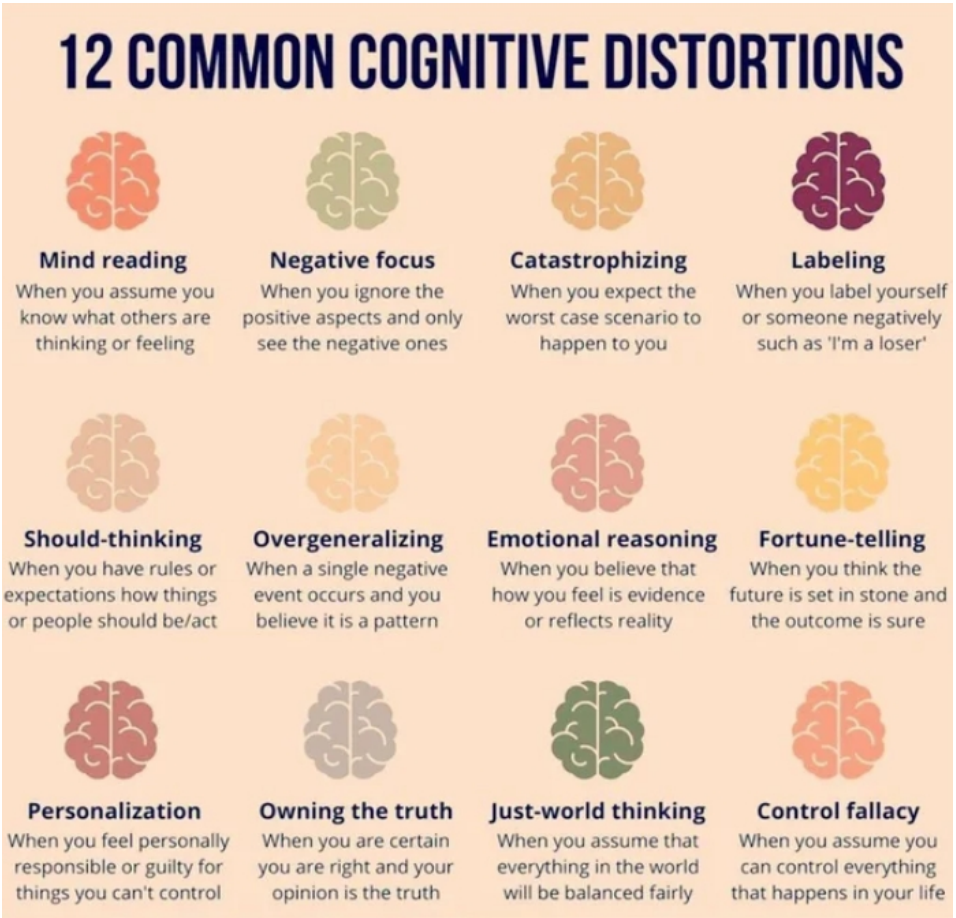
Когнитивные искажения (когнитивно-поведенческая терапия)

Catastrophizing: Exaggerating the importance of negative events (“that’s won’t end good”, “game over”...)

Fortune-telling: Making predictions, usually negative ones, about the future (“I will not”, “we will not”...)

Mind reading: Believing you know what others are thinking (“everyone knows”, “everyone thinks”...)

Источник:
https://www.pnas.org/doi/pdf/10.1073/pnas.2102061118



Что можно узнать из текста о психике социума?

Когнитивные искажения (когнитивно-поведенческая терапия)

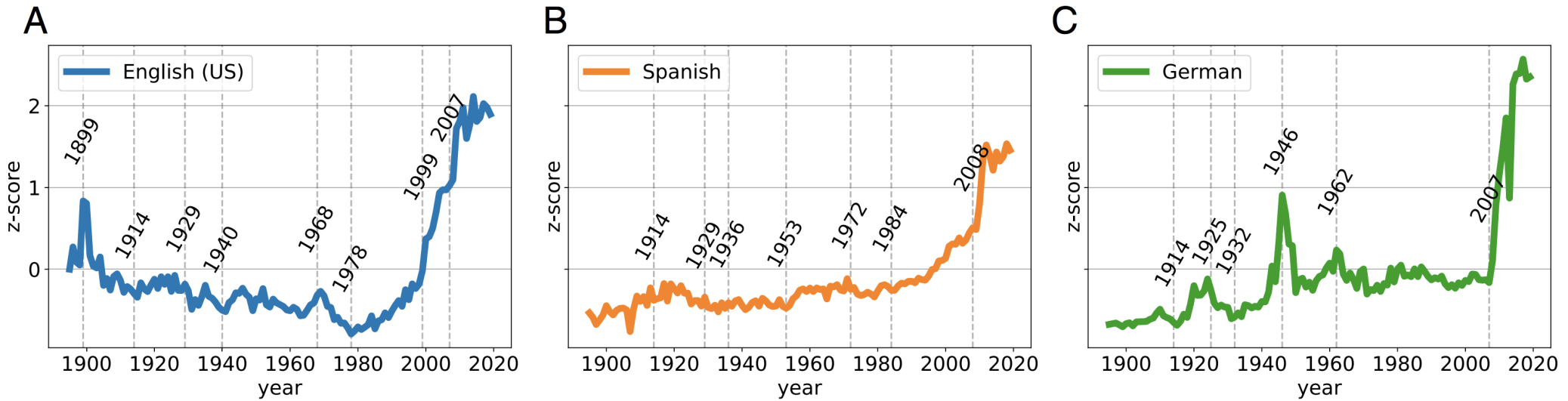


Fig. 2. (*A*–*C*) Median z scores of time series of CDS n-gram prevalence from 1855 to 2020 (125 y) in US English (*A*), Spanish (*B*), and German (*C*) with year markers added for major historical events. All time series reveal stable or declining levels for most of the 20th century followed by a sharp surge of cognitive distortions in the past three decades. US English shows declining levels from 1899 to 1978, with minor peaks around 1914 and 1940 (World War I and World War II) and notably 1968. This decline is followed by a surge of CDS prevalence starting in 1978 that continues to 2019. For Spanish we find stable levels from 1895 to the early 1980s at which point a trend occurs toward higher CDS prevalence levels above any of those previously observed. German shows stable CDS prevalence levels, with the exception of strong peaks around and after World War I and World War II, until 2007 at which point a sudden surge occurs.

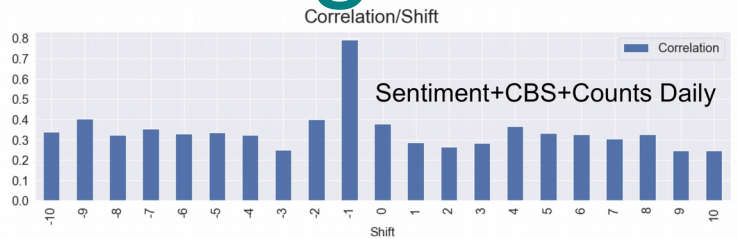
Источник:
https://www.pnas.org/doi/pdf/10.1073/pnas.2102061118

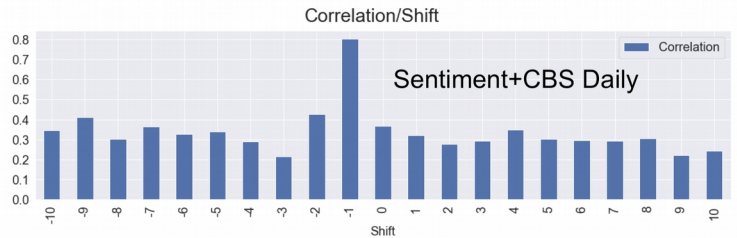
Как манипуляции в тексте влияют на поведение?

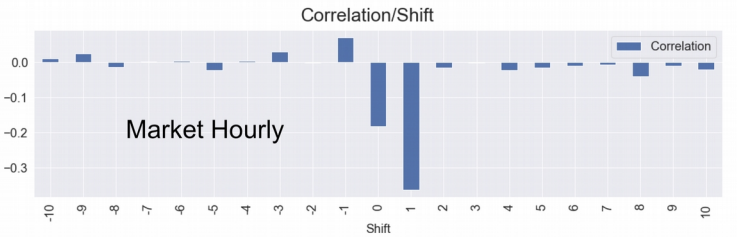
Связь когнитивных искажений с ценой криптовалют и объемом торгов

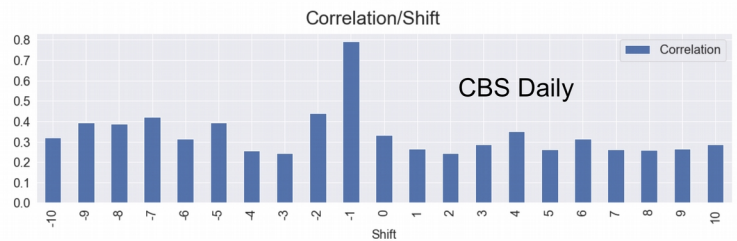
Источник:
https://arxiv.org/abs/2204.12928
https://link.springer.com/chapter/10.1007/978-3-031-19907-3_4

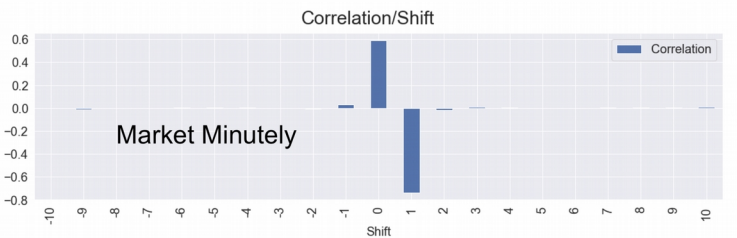
Searching for Compound Causation

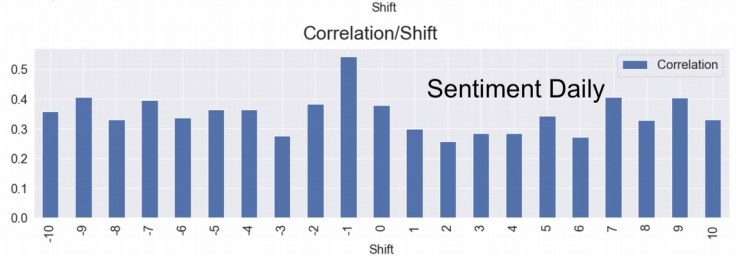


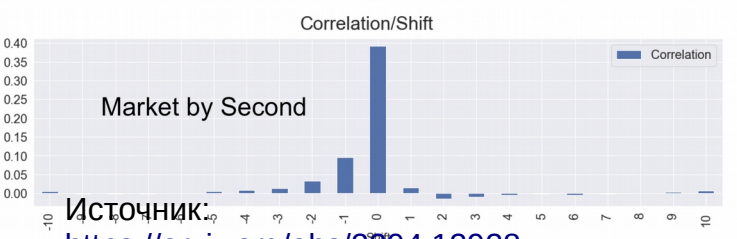












Источник:
https://arxiv.org/abs/2204.12928
https://link.springer.com/chapter/10.1007/978-3-031-19907-3_4

Платформа социальных вычислений Aigents®

Ethereum

PayPal

Steemit

WWW

Golos.id

Google

VKontakte

Facebook

Messenger

Slack

Telegram

RSS

Discourse

Friends

Twitter

Reddit

News

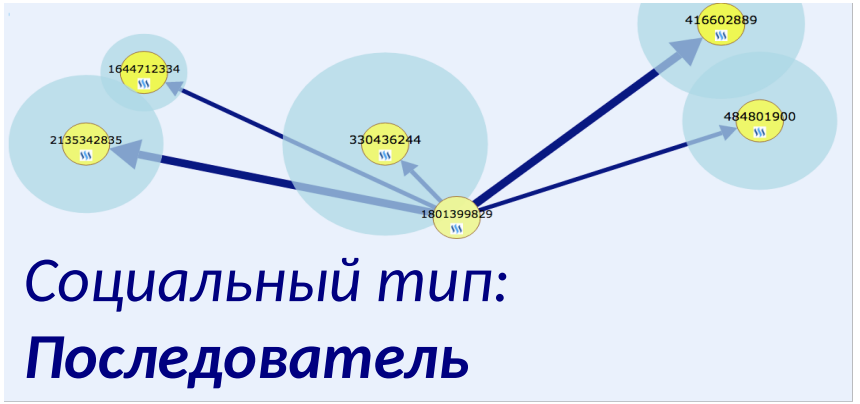
Aigents

Me

Платформа социальных вычислений Aigents®

https://link.springer.com/chapter/10.1007/978-3-030-03402-3_25
https://ieeexplore.ieee.org/document/8109887
https://www.sciencedirect.com/science/article/pii/S1877050918300383
https://ieeexplore.ieee.org/document/8071965
https://ieeexplore.ieee.org/document/7587680

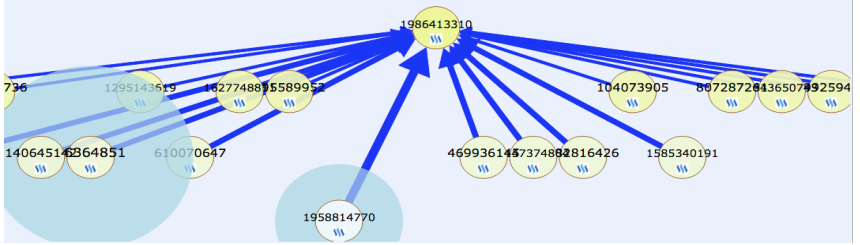
Применение 1: Выявление и поиск социо-коммуникационных типов, профилей и структур, кругов общения и аудиторий, лидеров мнений и каналов распространения информации



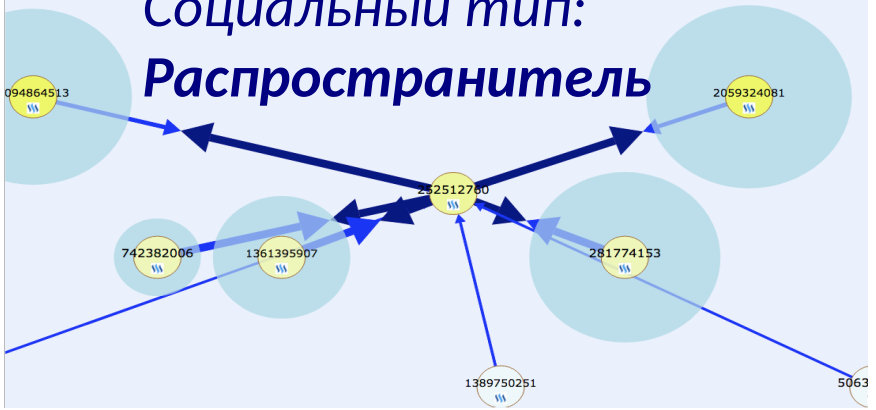
Социальный тип:
Последователь

Социальный тип:
Друг и соратник

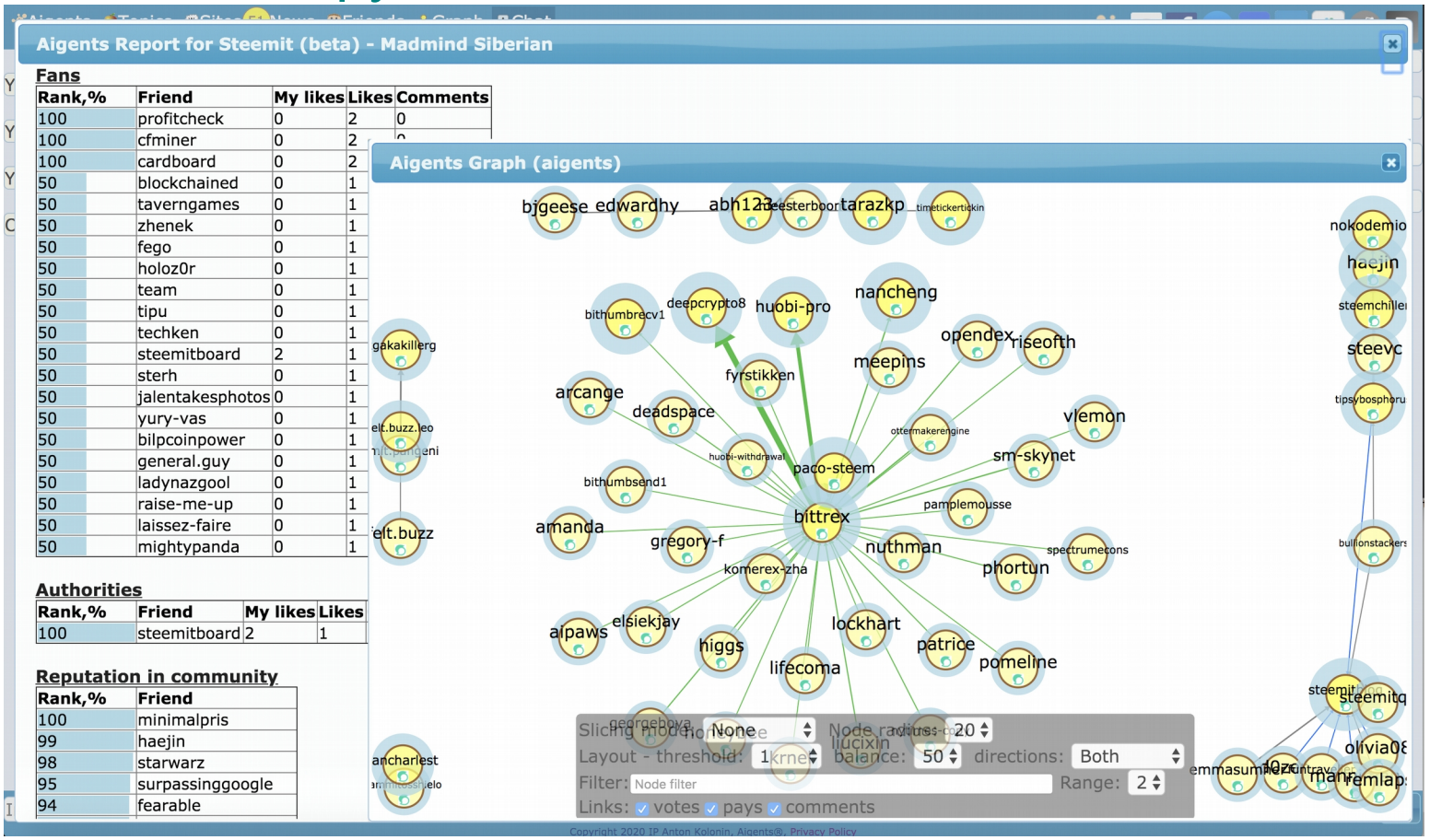
Социальный тип:
Лидер



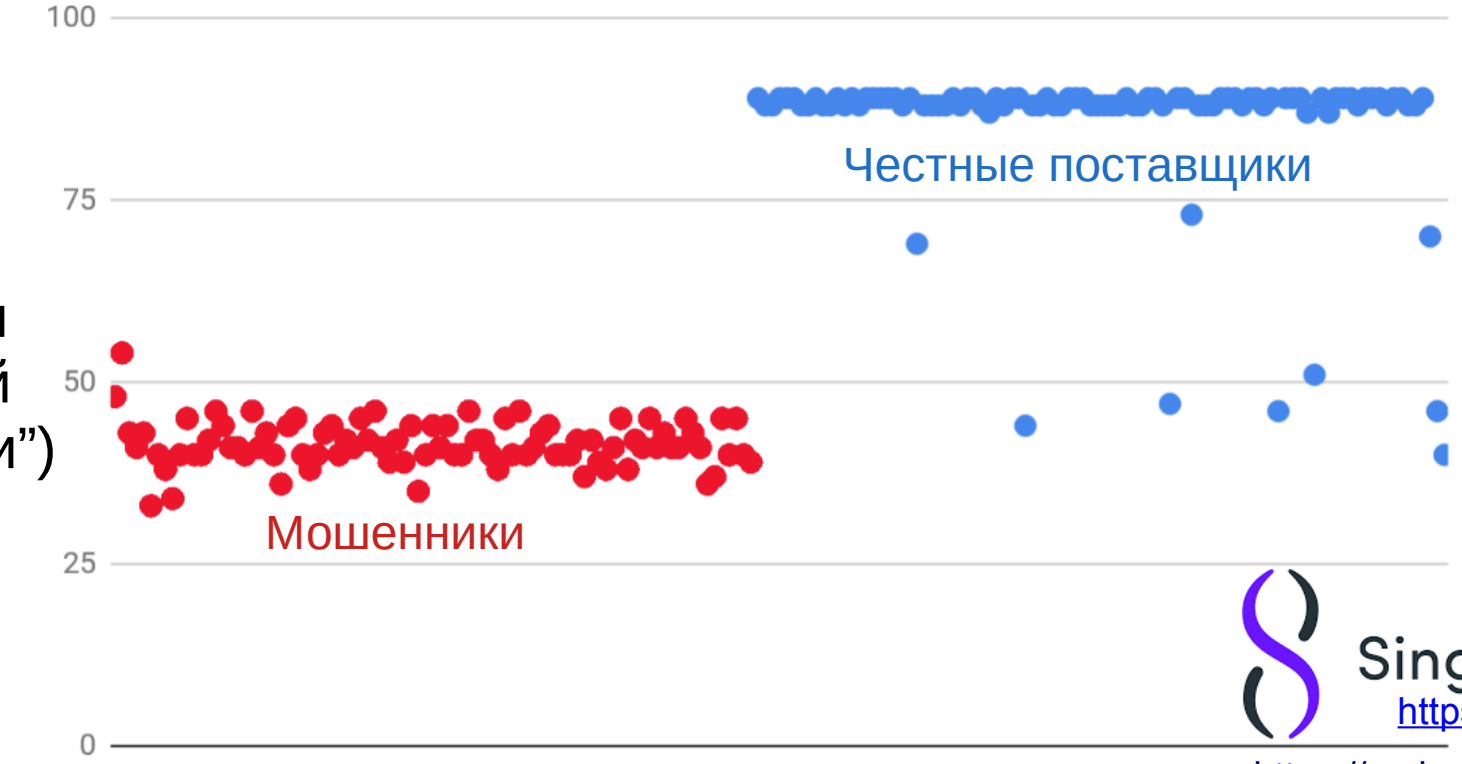
Социальный тип:
Распространитель



Применение 2: Количественное определение уровней репутации и доверия, а также – характера, количества и качества межличностных отношений в крупно-масштабных сетевых сообществах



Применение 3: Разделение поставщиков в системе продаж онлайн на честных и мошенников (алгоритм “текучей репутации”)



Уровень доверия (“текучей репутации”)

SingularityNET
https://singularitynet.io

https://arxiv.org/abs/1811.08149
https://arxiv.org/abs/2108.03542

Ranks of Suppliers, dishonest Supplier (including alias) in red and honest suppliers in blue

Reputation System for Marketplaces

Scam Period	Reputation System	Loss to Scam (LTS)	Profit from Scam (PFS)	LTS Relative Decrease	PFS Relative Decrease
182	No	2.4%	44%		
182	Regular	2.7%	49%	-13%	-13%
182	Weighted	2.3%	42%	2%	3%
182	TOM-based	1.4%	30%	41%	31%
182	SOM-based	2.2%	40%	8%	7%
92	No	3.0%	54%		
92	Regular	3.5%	65%	-19%	-20%
92	Weighted	2.8%	52%	5%	4%
92	TOM-based	1.7%	36%	43%	33%
92	SOM-based	2.6%	47%	13%	12%
30	No	3.9%	73%		
30	Regular	4.7%	86%	-19%	-18%
30	Weighted	3.3%	59%	17%	19%
30	TOM-based	1.5%	31%	63%	58%
30	SOM-based	1.5%	27%	63%	63%
10	No	4.4%	81%		
10	Regular	4.7%	88%	-7%	-8%
10	Weighted	3.0%	54%	33%	33%
10	TOM-based	0.2%	3%	**96%**	**96%**
10	SOM-based	0.3%	6%	**93%**	**93%**

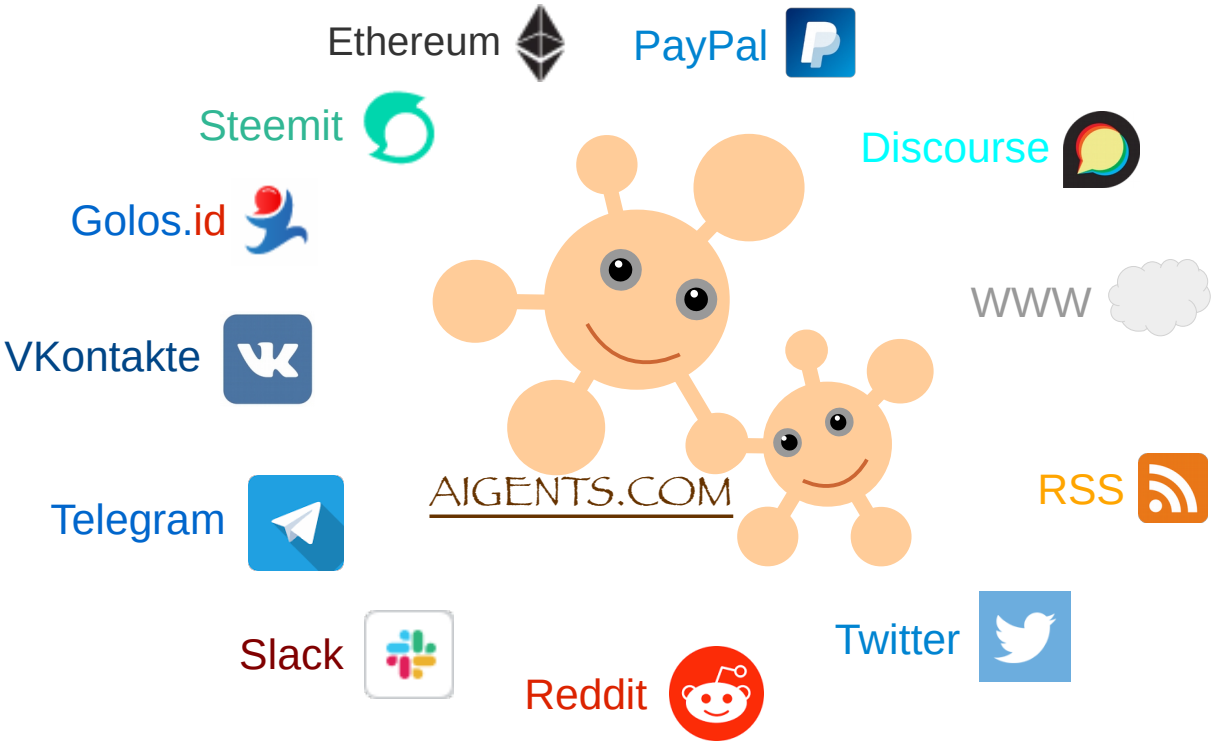
No reputation system: participants are making decisions relying only on their own memories and not referring to any reputation system.
Regular reputation system: standard version of reputation system. Does not take into account any factors other than values of ratings that consumers make to suppliers.
Weighted reputation system: When considering ratings as regular reputation system does, accounts to financial values of transactions between participants so that rating values are weighted by costs of transactions that are rated.
TOM-based reputation system: In addition to weighting ratings with financial values per-transaction, weights the ratings based on the rater's time on the market (TOM) as a "proof-of-time". That is, the raters (buyers) are implicitly rated based on how long have they been on the market. So, rating by buyer with a longer history influences reputation of a seller more than the one made by rater with shorter history.
SOM-based reputation system: In addition to weighting ratings with financial values per-transaction, weights the ratings based on rater's spendings on the market (SOM) as a "proof-of-burn" value. That is, the raters (buyers) are implicitly rated based on how much they spend on this market. So, rating by buyer with a lot of spendings influences reputation more than the one made by rater with smaller spendings.

SingularityNET

https://arxiv.org/abs/2210.02356

https://github.com/singnet/reputation

Доступность для платформы Aigents®



https://github.com/aigents/
https://aigents.com/

Спасибо за внимание!

Антон Колонин

akolonin@aigents.com

Telegram: akolonin

Новосибирский государственный университет

*НАСТОЯЩАЯ НАУКА

https://www.nsu.ru

AIGENTS

https://aigents.com

AGI IN RUSSIAN

https://agirussia.org