TCER Working Paper Series

Vaccine Uptake - Geographic Psychology or the Information Field?

Peter Romero
Eisaku Daniel Tanaka
Yuki Mikiya
Shinya Yoshino
Atsushi Oshio
Teruo Nakatsuma

December 2023

Working Paper E-191
https://www.tcer.or.jp/wp/pdf/e191.pdf

TOKYO CENTER FOR ECONOMIC RESEARCH
1-7-10-703 Iidabashi, Chiyoda-ku, Tokyo 102-0072, Japan

Abstract

Rapid vaccine uptake is a crucial component of public health, and contributes towards a stable economy. While previous research shows influences from spatial distribution of personality, and temporal influences of the information field, we integrate both by help of a suggested framework. Partial evidence for the framework is delivered by a subsequent Japan-wide analysis of the influence of spatial personality and spatiotemporal changes in the information field. More concretely, we analyse 25,614,106 hyperlocal Tweets from 2019 to 2021 that cover all prefectures of Japan using J-LIWC2015, 14,418 responses to the TIPI-J collected between 2012 and 2019, 6,266 responses to the Japanese version of NEO-FFI and a COVID-19-vaccine-related questionnaire that covers cognitive, affective, and behavioural items. We offer three models that predict mid-term vaccine uptake, long-term vaccine uptake, and abidance by governmental measures. Results indicate that vaccine uptake speed is predicted by temporal distribution of the information field, geospatial distribution of agent and contextual personality (Extraversion), presence of severe COVID-19 cases, and agent belief systems. More concretely, relevant language (negative emotions, affected language, anxiety, risk-related language) that implies close proximity (family-related language), the presence of severe COVID-19 cases, contextual and agent Extraversion, as well as agent beliefs that vaccines are justified, predict vaccine uptake speed and abidance by governmental measures. For analysis, we suggest a semi-manual statistical feature reduction approach that allows injection of theoretical consideration by chaining traditional steps of statistics and statistical learning with human selection of final predictors. We then discuss possibilities to include our findings for enhancing vaccine acceptance, shaping better public health behaviors, customising and precisely targeting government communications to counter misinformation, fostering a healthier and more resilient society, as well as a more stable economy.

Peter Romero
Keio University
Graduate School of Economics
2-15-45 Mita, Minato-ku, Tokyo, 108-8345, Japan
rp@keio.jp

Eisaku Daniel Tanaka
Keio University
Faculty of Economics
2-15-45 Mita, Minato-ku, Tokyo, 108-8345, Japan
eisakut179@keio.jp

Yuki Mikiya
Keio University
Graduate School of Law
2-15-45 Mita, Minato-ku, Tokyo, 108-8345, Japan
yukimikiya@keio.jp

Shinya Yoshino
Waseda University
Faculty of Letters, Arts and Sciences
1-24-1 Toyama, Shinjuku-ku, Tokyo, 162-8644, Japan
shinyosh8.5@gmail.com

Atsushi Oshio
Waseda University
Faculty of Letters, Arts and Sciences
1-24-1 Toyama, Shinjuku-ku, Tokyo, 162-8644, Japan
oshio.at@waseda.jp

Teruo Nakatsuma
TCER
and
Keio University
Faculty of Economics
2-15-45 Mita, Minato-ku, Tokyo, 108-8345, Japan
nakatuma@econ.keio.ac.jp

Vaccine Uptake - Geographic Psychology or the Information Field?

Peter Romero[1,2*], Eisaku Daniel Tanaka[3*], Yuki Mikiya[4*], Shinya Yoshino[5*], Atsushi Oshio[5*] and Teruo Nakatsuma[3*]

[1*]Graduate School of Economics, Keio University, 2-15-45 Mita, Minato-ku, Tokyo, 108-8345, Japan.
[2*]The Psychometrics Centre, University of Cambridge, Trumpington Street, Cambridge, CB2 1AG, Cambridgeshire, United Kingdom.
[3*]Faculty of Economics, Keio University, 2-15-45 Mita, Minato-ku, Tokyo, 108-8345, Japan.
[4*]Graduate School of Law, Keio University, 2-15-45 Mita, Minato-ku, Tokyo, 108-8345, Japan.
[5*]Faculty of Letters, Arts and Sciences, Waseda University, 1-24-1 Toyama, Shinjuku-ku, Tokyo, 162-8644, Japan.

*Corresponding author(s). E-mail(s): rp@keio.jp; eisakut179@keio.jp; yukimikiya@keio.jp; shinyosh8.5@gmail.com; oshio.at@waseda.jp; nakatuma@econ.keio.ac.jp;

Abstract

Rapid vaccine uptake is a crucial component of public health, and contributes towards a stable economy. While previous research shows influences from spatial distribution of personality, and temporal influences of the information field, we integrate both by help of a suggested framework. Partial evidence for the framework is delivered by a subsequent Japan-wide analysis of the influence of spatial personality and spatiotemporal changes in the information field. More concretely, we analyse 25,614,106 hyperlocal Tweets from 2019 to 2021 that cover all prefectures of Japan using J-LIWC2015, 14,418 responses to the TIPI-J collected between 2012 and 2019, 6,266 responses to the Japanese version of NEO-FFI and

a COVID-19-vaccine-related questionnaire that covers cognitive, affective, and behavioural items. We offer three models that predict mid-term vaccine uptake, long-term vaccine uptake, and abidance by governmental measures. Results indicate that vaccine uptake speed is predicted by temporal distribution of the information field, geospatial distribution of agent and contextual personality (Extraversion), presence of severe COVID-19 cases, and agent belief systems. More concretely, relevant language (negative emotions, affected language, anxiety, risk-related language) that implies close proximity (family-related language), the presence of severe COVID-19 cases, contextual and agent Extraversion, as well as agent beliefs that vaccines are justified, predict vaccine uptake speed and abidance by governmental measures. For analysis, we suggest a semi-manual statistical feature reduction approach that allows injection of theoretical consideration by chaining traditional steps of statistics and statistical learning with human selection of final predictors. We then discuss possibilities to include our findings for enhancing vaccine acceptance, shaping better public health behaviors, customising and precisely targeting government communications to counter misinformation, fostering a healthier and more resilient society, as well as a more stable economy.

Keywords: behavioural economics, COVID-19, psycholinguistics, psychometrics, vaccine uptake

JEL Classification: C21 , D91 , I12

1 Introduction

The rapid uptake of vaccines is a critical determinant in controlling pandemic outbreaks, since it affects herd immunity levels and thus can mitigate the spread of viruses like COVID-19 [1, 2]. Beyond availability and direct or indirect measures to assert conformity, willingness or hesitancy of a population to get vaccinated determines uptake speed [3]. This willingness or hesitancy in turn is based on individual-psychological, normative-social, and cultural factors, as well as the information field in which an agent is embedded. Cultural factors include culture-specific norms and values that influence health behaviours, including vaccine acceptance [4]. Normative-social factors consist of social norms, beliefs, peer behaviour, and group dynamics on individual decision-making processes, which include vaccination decisions [5]. Individual psychological aspects like risk perception [6], dark triad [6, 7], conspiracy beliefs [6, 8–10], and in particular personality traits, are associated with health behaviour, affecting the responsiveness and compliance to vaccination campaigns [11]. The information field is the collective informational ecosystem from media reports, news broadcasts, governmental campaigns, and social networking services that shape cognition, affective perception, and behaviour towards specific issues. In this information field, narratives battle for attention, and the spread of misinformation or the lack of clear and relevant counter-information

can lead to increased vaccine hesitancy, while the dissemination of accurate and nudging, persuasive information can encourage uptake [12–14]. Hence, tailored communication strategies that address vaccine hesitancy, promote its acceptance and faster uptake, are essential for effectively mitigating public health crises, and could become at one point as crucial as the development of the vaccines themselves.

The strength the information field depends on valence and personal relevance of the information, which is, beyond personal connections with the life spaces of individuals, largely determined by geospatial and temporal proximity [15]. The geospatial distribution of psychological phenomena is well researched, however research on temporal distribution of psychological phenomena is sparse [16].

While each of these components has been studied in separate, their interaction has not been researched sufficiently. Especially underlying causal mechanisms are unclear, wherefore literature partially contradicts itself. Therefore, deeper and more rigorous research is needed to tailor interventions for enhancing the effectiveness of communication in vaccination campaigns and thus increase uptake, general health awareness, and immunity against adversarial influence and misinformation campaigns.

2 Relevant Work: Relevance of Time and Space for Psycholinguistic Measures

There exist a plethora of literature that covers geospatial, psychological and temporal aspects of the COVID-19 pandemic or vaccine uptake, and the shape of the information field. However, only a few paper that cover such aspects in combination to gather a more complex view of the situation and the interaction of these components.

Neff et al. (2023) [17] offer an excellent overview of 20 years of research literature on vaccine hesitancy in online spaces, and analyse over 100 papers for that. They find that "levels of confidence and hesitancy" towards vaccines "might differ across conditions and vaccines, geographical areas, and platforms, or how they might change over time." (p.1.) and identify gaps for necessary research: focus on disciplinary actions, vaccine specifics, conditions, disease focus, involved stakeholders, implications, methodology, and geographical coverage. While not explicitly, they open the discussion about time and space related issues of vaccine hesitancy. Peters et al (2023) [18] combine data from self-reported personality traits of 3.5 million people and mobility observations of 29 million people in the United States and Germany to better understand both regional differences, and movement patterns that lead to viral spread. Their results show that regional compliance behaviour and personality differences, particularly Openness and Neuroticism, significantly influence the early spread of COVID-19, even after adjusting for socio-demographic, economic, and pandemic-related factors, while also revealing variations across countries, over time, and compared to individual-level effects. More concretely, they show

that in the early stages of the pandemic, Openness was a risk factor, whereas Neuroticism rather acted as protective influence. However, they find vast differences in terms of country-level Extraversion, temporal-level Openness and individual-level Agreeableness and Conscientiousness. Given the complexity of their findings, they warn about over-simplification. Also, one of the authors finds in a previous study that behavioural changes in the high-Neuroticism population may be externally triggered by the predominant narrative in the information field and thus rather mediate its influence [19]. The influence of Neuroticism, Openness, high Agreeableness, as well as dispositional greed on COVID-19-related hoarding behaviours are also shown by Yoshino and Oshio (2021) [20], who uses a similar approach, yet does not cover the information field. Finally, Mangalik et al. (2023) [21] model mental health in the USA through large-scale analysis of 1.2 billion Tweets from 2 million geo-located users to estimate changes in anxiety and depression on a granularity of weekly level time-wise and county-level geography-wise. They find moderate to large associations with mental health assessment and survey scores, and suggest this approach to economically or medically under-resourced communities that however have social media access.

However, these studies lack an overarching, connective framework, which allows scaling and comparison of research findings, and which informs potential avenues for simplification by abstraction. Hence, we suggest such a framework, and deploy it subsequently to simplify measures without giving up methodological rigour or theoretical foundations.

3 Method

3.1 Research Model

To account for geospatial influences, we conclude that more proximal influences are more important for agents than distal ones. In extension, and aligned with systems theory [22], we assume that this measure of proximity is ordered by systemic levels of individual (e.g., person), micro (e.g., family), meso (e.g., company), exo (e.g., industry), and macro system (e.g., society or, in extension, the world). This allows opening a spatial vector of influence $\vec{x} = [\alpha, \beta, \ldots, \omega]$, whereby α denotes the most individual systemic, and ω the most macrosystemic level.

To account for temporal influences, we conclude that when more recent events have a stronger effect on agents' perception, cognition, and behaviours, past and future effects are going to have a weaker effect. Thereby, we assume that the current moment t is the point of reference for an agent, and that each agent has backwards memory and, based on historic memory, forward-prediction capabilities of n time steps. For the sake of simplification and *ceteris paribus*, be assume that both directions consist of an equal amount of steps, hence forming nearly identical time intervals into the past and future. Thus at time t, two dynamic event horizons will arise; ϵ_{t-n} and ϵ_{t+n}, that shift with agent time at each time step in n.

Subsequently, we define the geospatial information field as a series of n spatial vectors $\vec{x}_z$ at current agent time t with $z \in [t-n \dots t+n]$. Each of these vectors $\vec{x}_z$ represents a spatial influence vector at a time z that represents systemic influences of varying strengths.

$$\vec{x}_{t-n} = \begin{pmatrix} \omega_{t-n} \\ \beta_{t-n} \\ \vdots \\ \alpha_{t-n} \end{pmatrix}, \dots, \vec{x} = \begin{pmatrix} \omega \\ \beta \\ \vdots \\ \alpha \end{pmatrix}, \dots, \vec{x}_{t+n} = \begin{pmatrix} \omega_{t+n} \\ \beta_{t+n} \\ \vdots \\ \alpha_{t+n} \end{pmatrix} \tag{1}$$

These vectors concatenate into a matrix X, which represents the information field of an agent a at time t, and which represents all informational effects that influence that agent to varying strengths depending on spatiotemporal proximity.

$$X = \left(\vec{x}_{t-n} \dots \vec{x} \dots \vec{x}_{t+n} \right) = \begin{pmatrix} \omega_{t-n} & \dots & \omega & \dots & \omega_{t+n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \alpha_{t-n} & \dots & \alpha & \dots & \alpha_{t+n} \end{pmatrix} \tag{2}$$

Figure 1 represents this information field in a more intuitive way.

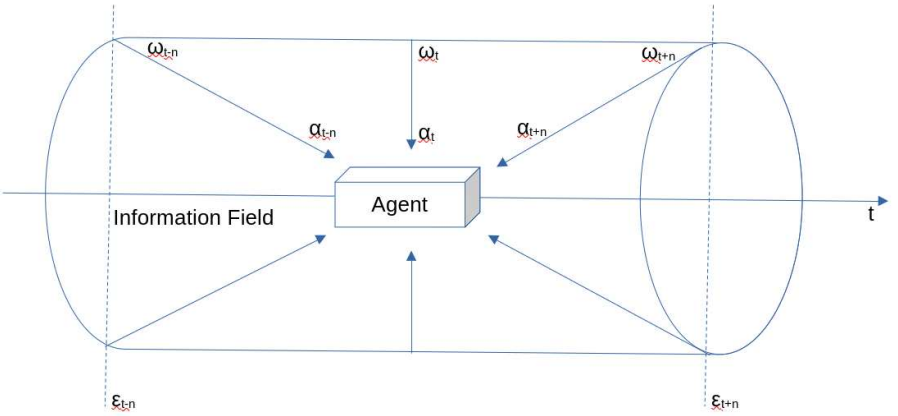

Fig. 1 Research Model: vector-based definition of the information field

While this represents a simplification and leaves quite a few questions unsolved, it is to the best of our knowledge the first attempt to operationalise the spatiotemporal nature of the information field by help of systems theory. Unsolved questions are, for example, whether both event horizons are the same number of time steps n apart from each other, or whether priming from events

at the current time t might generate memory effects that influence future prediction by selective memory retrieval or over-emphasis of specific memories.

Another question is where exactly the information field attaches to; at a systemic level above the agent, and from there through social norms and peer pressure, or as a sort of collective behaviour that only exists on collective level? However, peer pressure would rather attach on an individual systemic level, and could be counted as a part of the information field through informal communication. But then, it has to be asked, where in the psychological architecture of the individual system this attachment would take place, and which influence the relevant context of an agent has. For example, the information field could moderate the transfer of competency potential from psychological latent traits into behaviours. Or, it could moderate the transfer from relevant behaviours into respective outcomes. Finally, moderation - or even mediation - could take place at both transfer points, depending on the kind and strength of message and contextual embedding. Figure 2 depicts these potential moderation mechanisms at the transfer points within the individual-systemic level. For the sake of simplification, we excluded the neurophysiological level that resides underneath the other individual-systemic levels, which serves more fundamental functions.

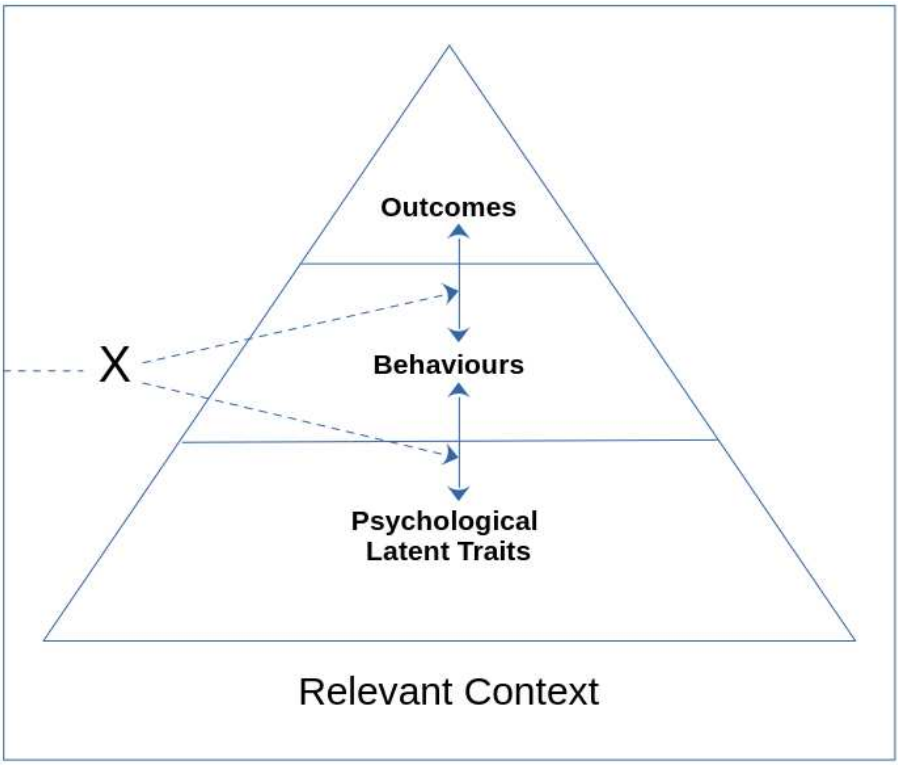

Fig. 2 Agent in contextual embedding

Since each of these questions demands much deeper discussion and further research, we assume that all mechanisms could be relevant, hence do not differentiate for any specific individual-systemic architecture. Also, since potential collective behaviours that cannot be measured on individual-systemic level are possible, we conclude that these have to be taken into consideration, as well.

Finally, to operationalise uptake speed, we need to take various variables into account that are of stochastic nature and about which deeper information is partially inaccessible. For example, on-the-ground-truth on fluctuations in the local availability of vaccines, inflexibilities of the medical infrastructure, and local decisions on preferential treatments of age groups is mostly not documented.

Hence, we decide to use the first inflection point i of overall aggregated vaccine uptakes per initial vaccine and booster shots, which follow sigmoid functions with a clear slow growth at the beginning, followed by exponential growth, then an inflection point and a stabilisation at the upper asymptote m. This kind of curve is used in a variety of fields to model natural phenomena, from difficulty of psychological test scores in item response theory [23] to wildlife growth [24] to market saturation [25] to modelling future COVID-19 vaccine uptake over time and space [26]. The advantage of that approach is that we do not need to make any assumption about the functional form over and above its sigmoid nature. Also, the interpretation is intuitive — the lower asymptote m_0 starts at the origin and represents first day of vaccine availability, the slow growth at the beginning can be interpreted as covering a plethora of aforementioned stochastic influences that are undocumented, the subsequent exponential growth can be interpreted as the kernel of the *populus* that is willing or hesitant towards vaccines, the inflection point i cuts the curve in halves, thus represents the point when other mechanisms take over. The exponential decrease after the inflection point i could be interpreted as stochastic processes during that change (e.g., logistical challenges), and the slow decrease until the upper asymptote m represents the kernel of the population that takes the vaccine for other reasons like social pressure, logistical challenges, lack of health awareness or sense of urgency. Finally, the upper asymptote m represents the total number of people that accepted the vaccine for either reason or ability to receive it. The inflection point i indicates change in the concavity of the function $\sigma_m(x)$, which is detected by sign-change of its second derivative $\sigma''_m(x)$ and occurs at $x = 0$ irrespective of the value of m.

Thereby, we bound the sigmoid function $\sigma_m(x)$ to the upper asymptote m:

$$\sigma_m(x) = \frac{m}{1 + e^{-x}} \tag{3}$$

Its first derivative displays the instantaneous rate of change at each specific point x on $\sigma_m(x)$:

$$\sigma'_m(x) = \frac{d}{dx}\left(\frac{m}{1+e^{-x}}\right) = m \cdot \frac{e^{-x}}{(1+e^{-x})^2} \tag{4}$$

And its second derivative displays the instantaneous rate of change at each specific point x on $\sigma'_m(x)$, whereby the sign indicates the slope of the tangent line to increase or decrease:

$$\sigma''_m(x) = \frac{d}{dx}\left(m \cdot \frac{e^{-x}}{(1+e^{-x})^2}\right) \tag{5}$$

We denote i_1 as the inflection point of the first vaccine dose, where $x = 0$, which happens at time step n_{i_1}. Hence, we identify the overall uptake vaccine uptake speed v between the first shot and the x^{th} booster shot with:

$$v = n_{i_x} - n_{i_1} \tag{6}$$

While the first dose and first booster shot of the vaccine were available in short distance to each other, the second booster shot was only available after a significant time gap, as were subsequent booster shots, which were also limited to certain age and high risk groups. We denote each injection with $x \in \mathbb{N}$, thus the first dose as 1, the first booster shot as 2, and so forth. Hence, we interpret that $v = n_2 - n_1$ as having taken place in too short distance from each other to being able to exclude coercion or immediate fear thus obfuscating the true intrinsic vaccine acceptance. However, $v = n_3 - n_1$ took place after sufficient time for habituation passed, hence we use it as proxy for measuring the true intrinsic vaccine acceptance, and we denote it as v_{medium}. Finally, $v = n_5 - n_1$, denoted as v_{long}, displays the long-term vaccine acceptance; with the restriction of just representing a subgroup, hence it is not directly comparable to v_{medium}. However, both v_{medium} and v_{long} can be compared geographically, thus displaying local differences in vaccine acceptance. For an industrialised country like Japan, we can assume that infrastructural conditions are mostly homogeneous, hence *ceteris paribus*, we can exclude regional economic inequalities.

Hence, we define the outcome variables as:

$$v_{medium} = n_3 - n_1 \tag{7}$$

$$v_{long} = n_5 - n_1 \tag{8}$$

3.2 Research Design

We conduct a non-experiment by collecting behaviour artefacts and survey data at various data points before and during the COVID-19 pandemic.

For that, we represent the model in the following way: we use tweets to identify the information field that are temporally sorted, analysed for proximity to the agent, and of which we know the approximate geographic location on city level of the sender. The temporal sorting is realised through daily tweet collection, the approximate location is realised through identifying and scraping the tweets of followers of hyperlocal entities like police stations, local sports teams, or city mascots – following the hypothesis that nobody else but locals would have a reason to do so. While some authors approximate physical distance by narrative strength [27], this approach relies on available commercial tools in a target language, and does not differentiate on a system-theoretic level. Therefore, we identify approximate agent-proximity in a simple and more intuitive way by LIWC categories [28], whereby we hypothesise that LIWC categories like 'Affective processes' represent agent-internal language, 'Family','Friends', 'Home', and 'Perceptual processes' rather indicate language with close agent proximity, whereas 'Work' rather indicates language from more distal systemic layers.

We augment these with local data on COVID-19-related severe cases, hospitalisation, and deaths, as well as vaccination numbers to capture local norms, peer pressure, or imitation effects. Furthermore, we use personality questionnaires taken from participants all over Japan at four different time steps before and during the pandemic to identify geographic distribution of personality.

To understand the difference between agent and contextual properties, we conduct a Japan-wide survey covering COVID-19-related attitudes, about the severity of the situation, abidance by governmental measures, and cooperation with others on pandemic-related issues. To control against geographic personality, participants in the survey are asked to fill out a personality questionnaire, as well.

Since most of the data is not specifically collected for COVID-19-related research, the temporal granularity of the data points is uneven - ranging from individual points in time (e.g., each survey taken) over monthly data (e.g., certain COVID-19 statistics) to daily data (e.g., tweets, or vaccine doses administered). Hence, we decide to aggregate the data in the predictor space based on waves of COVID-19. Based on media research and official announcements, we identify eight waves, of which five are relevant for the phenomena observed:

1. < 16.01.20 – 0 baseline
2. 16.01.20 - 26.03.20 – 1st wave
3. 26.03.20 - 30.06.20 – 2nd wave
4. 01.07.20 - 31.07.20 – 3rd wave
5. 01.08.20 - 24.09.20 – 4th wave
6. 24.09.20 - 31.12.20 – 5th wave
7. 2021 – 6th wave (not relevant for the vaccine-related observations)
8. 2022 – 7th & 8th wave (continuing until 2023; not relevant for the vaccine-related observations)

Since the coarsest geographic differentiation is prefecture level, we furthermore aggregate all geographic data on thereon. Therefore, the level of analysis is by wave and prefecture.

Finally, we define the outcome space o as duration in days with $o \in \mathbb{N}$, with $(o \in v_{long} \wedge o \notin v_{medium}) \vee (o \notin v_{long} \wedge o \in v_{medium})$, thus creating quasi-continuous outcome measures.

3.3 Data Set

We use the following data sets in our analysis to represent the different aspects of the framework:

- Tweets: daily hyperlocalised tweets from before and during the pandemic (on a daily granularity, aggregated and converted to all 60 LIWC features)
- COVID-19 Data: official local vaccination numbers, death numbers, hospitalisation numbers (daily granularity of each number)
- Ground truth data: geospatially distributed personality questionnaires at three different times before COVID-19 (used as constant value of five Big Five scores for personality before COVID-19)
- Questionnaire Data: covering demographics, psychological questions, economic questions, and attitudes towards the COVID-19 situation and governmental measures, as well as willingness to abide by those and cooperate with others on pandemic-mitigation-related issues taken all over Japan (used as constant value of agent status for during the pandemic)
- Questionnaire Personality: Personality tests of these survey participants (used as constant value for personality during the pandemic; being comprised of five Big Five values and constituent 60 facets)

3.3.1 Twitter Sample

The dataset is comprised of hyperlocalised tweets generated from January 1st, 2019 to April 1st, 2021 from at least two cities of all 47 Japanese prefectures. It is comprised of 25,614,106 (SD = 44,924.94) tweets, with on average 189,734 extracted tweets from every city. The cities are chosen based on parameters like population size, but also spatial separation, based on official numbers [29]. Due to the size of Hokkaido with its sub-prefectures, at least two cities per sub-prefecture are chosen. Given the metropolitan status of Tokyo as one of the largest urban centres on the planet, those cities and special wards with the most population and spatial separation are carefully selected. This results in 1,646 accounts, on average 35 per prefecture, with a maximum of Hokkaido with 235 and a minimum of Kumamoto-ken with 11 twitter accounts. Excluding Hokkaido, the average number of accounts per prefecture is 30. The minimum number of tweets for a city is 70,425 tweets, and maximum number is 244,331 tweets. All tweets are harvested from 107,873 followers of 1,648 local city representative accounts like police stations and city mascots. On average, 799 (SD = 46.16) follower accounts are harvested for each city; the minimum number of accounts for a city is 596 and the maximum is 822.

After removal of language features like retweet identifiers and emojis, this data subsequently is analysed with Linguistic Inquiry Word Count (LIWC) [28] to extract theory-driven, dictionary-based, hard-coded features. In specific, the 2015 version of LIWC [30], and the Japanese dictionary and tokenisation method introduced by Igarashi, Okuda, and Sasahara (2021) [31], are used, text is preprocessed by their latest Japanese psycholinguistic tokenisation dictionary, the MeCab/IPADIC [32] python library, and their morphological analysis (word segmentation) and part of speech analysis (POS) code. This results in 60 category-by-category features based on word frequency analysis, featuring words and word stems, including standard language categories like pronouns, psychological processes like emotions, and six sub-scores: insight, causation, discrepancy, tentativeness, certainty, and differentiation [28]. Finally, all tweets from the same city are treated as one document, and daily LIWC results per prefectures are aggregated.

3.3.2 Survey

In January 2022, we conduct a survey to deeper understand regional connection between agent personality, agent cognitive and affective aspects about COVID-19, agent congruency with government and science, agent social synergy, and agent synergy with the narrative, thus about the information field, direct systemic embedding, and about agent-internal aspects.

This survey is comprised of demographic questions (age, gender, income, household income, number of children, family status), psychological questions (number of siblings, sibling order), economic questions (household income, income, times eating out per week, profession), and COVID-19-related questions. The specific COVID-19-related questions are:

- "The measures of the government are justified" - cognitive item to capture the level of perceived justification of governmental measures and implicitly the level of congruence of participants with governmental measures
- "I believe in vaccines" - cognitive item to capture the degree of trust in science of a participant and implicitly the degree of acceptance of governmental narrative
- "The COVID-19 situation is dangerous" - affective item to capture the emotional sense of danger as well as congruence with official health narrative
- "I cooperate with those around me to deal with the pandemic" - social-behavioural item to capture horizontal synergy of participants and implicitly abidance with norms of the direct contextual embedding
- "I abide by governmental measures" - behavioural item to capture the vertical synergy of participants and implicitly abidance and congruence with broader societal and cultural norms

Overall, 6,266 (prefecture mean = 133.32; SD = 114.92, min = 35.00 max = 564.00) persons participate in the survey, of which on average 48% (SD = 3%, min = 42%, max = 55%) per prefecture are male. The average age over prefectures is 45.56 (SD = 1.67; min = 41.71, max = 51.36).

Furthermore, we ask each agent to fill-out the NEO-FFI, a high quality, well established personality questionnaire. Enabling comprehensive insights into human personality, the NEO-FFI is a seminal instrument of psychometrics that is well-documented, developed on sound science, and has been a stable in countless international studies since the 1980s. It uses a five-point Likert scale and offers broad applicability in different use cases including professional assessment, clinical psychology, and research. For example, it is deployed in the assessment of personality disorders and for deriving optimal treatment strategies [33]. In research, it is used to study associations of personality with various psychological and behavioural phenomena like the influence of personality traits on life outcomes [34, 35]. It is based on the Five-Factor Model of personality [36]; the Big Five Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. The development of the NEO-FFI is a major advancement for personality psychology, since it provides a robust, valid, and reliable of the Big Five model [37], with meticulously researched facets. It is comprised of 60 items among the five latent traits of the Big Five model, each composed of further six 6 facets; asking two items per facet, of which about the half are reversely scored for being able to detect attention, cheating, and overall answering consistency. Despite being proprietary, we deploy the Japanese version of the NEO-FFI for our study, since it offers a robust framework, cross-cultural validity, well-proven psychometric properties, and still a decent answering time (10-15 min; which is much less than 45-60 min for the NEO-PI-R [33]) that allows its co-deployment with other surveys or questionnaires.

3.3.3 Ground truth data

For Ground truth, we use the open-sourced geospatial personality distribution data collected from Yoshino and Oshio (2021) [38], who uses the Japanese version [39] of the Ten Item Personality Inventory (TIPI) [40], a well-established psychometric instrument that exists in 27 languages and is used in 9,167 peer-reviewed papers. Meant for mass-deployment and large-scale studies, it uses a seven-point Likert scale, is comprised of only ten items; two per Big Five factor, of which one is reversed. “Although somewhat inferior to standard multi-item instruments” (p.504) [40], its outcomes for self-ratings, external ratings, and peer ratings vastly overlap with other established Big Five instruments, it displays high congruence between self-ratings and observer ratings, its test-retest reliability is high, and the levels of external correlates are on par with other studies reported in research.

Data is collected in three iterations; first between January and March 2012 (n = 4469, prefecture mean 95.09; SD = 85.95, min = 14.00, max = 388.00, 46% male; SD = 6%, min = 25% and max = 58%), the second iteration in January 2017 (n = 5619, prefecture mean 119.55; SD = 13.99, min 87.00, max 149.00, 60% male; SD = 5%, min = 50 %, and max = 71%), and the last iteration was in January 2019 (n = 4330, prefecture mean = 92.13; SD = 14.34, min = 58.00, max = 127.00, 66% male; SD = 6%, min = 53%, and max = 80%),

and overall n = 14,418, prefecture mean = 306.77; SD = 101.63, min = 161.00 max = 648.00, 57 % male; SD = 4%, min = 46%, and max max = 65%).

3.3.4 COVID-19-related data

We use official COVID-19 data provided by the Japanese government [41] that we cross-check with data from the World Health Organisation [42] to ensure their correctness. Our main focus lies on severe cases/ hospitalisations, and death numbers, since these not only have a significant economic influence and detrimental impact on the health system, but foremost since these are psychological markers that effect the perception, cognition, emotion, and behaviour of people.

3.4 Analysis

Survey data is cleaned manually for identifying potential repetitive and inattentive answering behaviour, and Twitter data as described in section . To represent the waves of COVID-19 in Japan explicated above, data is aggregated by these time windows. Since most COVID-19-related data is reported by official statistics on prefecture level, granulated city data is missing. Also, psychological data is collected in single points of time, wherefore no time series model can be used for further analysis. Unfortunately, this results in an unbalanced sample, with only 47 prefectures as cases, and 711 final features as predictors. Hence, application of normal machine learning methods would immediately burn all degrees of freedom, and yield no results. On the other hand, there are too many predictors to do serious theory-driven predictor selection. Therefore, we use a semi-manual method that simulates feature selection and dimensionality reduction by chaining traditional steps of statistics and statistical learning, which we designate as *statistical feature reduction*. Figure 3 depicts this approach.

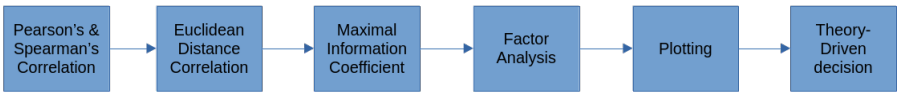

Fig. 3 Analysis flow for semi-manual *statistical feature reduction*

Concretely, we first pre-select features based on their broad association with the outcome measure by Pearson's r and Spearman's ρ. Then, we further test the association by Euclidean Distance Correlation and the Maximal Information Coefficient. Next, we reduce dimensionality by factor analysis, which is as vulnerable to unbalanced samples as other machine learning techniques, but sometimes captures latent traits and therefore can be more powerful in a selection task than a principal component analysis (PCA) [43], especially with psychological latent traits that are known to be intercorrelated and best to be explored with non-orthogonal rotations like "oblimin" [23]. Subsequently, we

plot the selected predictors against the outcomes to make an informed decision on the functional form. Finally, we select all predictors that occur in more than one selection method, and such that are aligned with theory, and create various manual iterations of regression models, until we find the one that yields the highest significance and R^2.

Step 1 - Correlation

In essence, Pearson's r is the preferred choice for analysing data with light-tailed distributions and linear relationships but requires normally distributed, continuous data, whereas Spearman's ρ is more flexible, and better suited for assessing monotonic relationships, accepts non-parametric data, including ordinal variables, and can be deployed to assess non-linear relationships or when the strict assumptions of Pearson's r are violated. It is ideal for heavy-tailed distributions or in cases where outliers are prevalent, a common scenario in psychological studies [44]. The occurrence of disparate results between these two methods, under the assumption of an identical underlying Gaussian distribution, typically signals potential issues with the data, most likely affecting the predictive outcomes. On the other hand, including both provides insights about the association between outcome and features and thus provides information about the best functional form at the same time.

Step 2 - Euclidean Distance Correlation

The Euclidean distance correlation serves as a statistical metric quantifying the dependence between two variable sets, denoted as X and Y. This measure effectively captures both linear and nonlinear relationships. It is grounded in the principle of distance covariance, which in turn expands the classic concept of covariance to accommodate multivariate and nonlinear scenarios.

For any two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, the Euclidean distance correlation, symbolized as $\mathcal{R}(X, Y)$, is defined by the following equation:

$$\mathcal{R}(X, Y) = \frac{\mathcal{V}^2(X, Y)}{\sqrt{\mathcal{V}^2(X) \times \mathcal{V}^2(Y)}} \tag{9}$$

Here, $\mathcal{V}^2(X, Y)$ represents the distance covariance between X and Y, while $\mathcal{V}^2(X)$ and $\mathcal{V}^2(Y)$ are the respective distance variances of X and Y.

The computation of distance covariance, $\mathcal{V}^2(X, Y)$, is given by:

$$\mathcal{V}^2(X, Y) = \frac{1}{n^2} \sum_{k,l=1}^{n} A_{kl} \cdot B_{kl} \tag{10}$$

In this context, A_{kl} and B_{kl} denote double-centered matrices derived from the pairwise Euclidean distances among the elements of X and Y, and n signifies the total number of samples.

The Euclidean distance correlation is constrained between 0 and 1, where $\mathcal{R}(X, Y) = 0$ implies a state of independence (subject to certain conditions) between X and Y, and $\mathcal{R}(X, Y) = 1$ reflects a perfect functional correspondence [45].

Step 3 - Maximal Information Coefficient

The Maximal Information Coefficient (MIC), as introduced by Reshef et al. (2011) [46], serves as a robust measure for quantifying the strength of the most pronounced linear or nonlinear associations between two variables within a dataset. Rooted in information theory, and diverging Pearson's r, which is limited to linear correlations, MIC excels in identifying a broad spectrum of associations, better encompassing nonlinear relationships than Spearman's ρ.

MIC strives to unveil any underlying patterns in data by aligning the greatest mutual information value with a grid-like partitioning on the x-y plane. For any given variables X and Y, the formal definition of MIC is as follows:

$$MIC(X,Y) = \max_{xy} \left(\frac{I(X,Y)}{\log(\min(x,y))} \right) \tag{11}$$

In this equation, $I(X,Y)$ denotes the mutual information shared between X and Y, with the maximization process spanning over the number of bins x and y utilized in segmenting the dataset.

MIC is constrained between 0 and 1, and finds extensive applications across diverse scientific domains, including bioinformatics, neuroscience, and environmental sciences, playing a pivotal role in revealing complex relationships within substantial datasets [47]. For example, Chauhan and Choi (2023) [48] use it to classify Alzheimer's Disease, Lazarsfeld, Johnson, and Adéníran (2022) [49] for ensuring differential privacy, and just like us, Zhou et al. (2022) [50] for feature selection.

Step 4 - Factor Analysis

Factor Analysis, a widely-utilized statistical method, aims to uncover latent variables, or factors, that elucidate the correlation patterns among a collection of observed variables. This technique simplifies the complexity of observed variables into a smaller number of unobserved variables, leveraging their correlations. The fundamental premise of this method is the direct correlation of each observed variable with any of the factors [51].

Considering $X = (X_1, X_2, \ldots, X_n)$ as a vector representing observed variables, Factor Analysis formulates X as:

$$X = \mu + \Lambda F + \epsilon \tag{12}$$

where:

- μ represents the vector of means.
- Λ is a matrix detailing the factor loadings on the variables.
- F comprises the vector of common factors.
- ϵ corresponds to the vector of unique factors, or error terms.

The primary objective of Factor Analysis is to ascertain the factor loadings Λ that optimally account for the observed correlations in the dataset.

It is widely used in research, e.g., to deduce the covariance structure from diverse data sources [52], for investigating temporal and spatial variations in patterns [53] as we do in this paper, or to reduce items in the construction of personality questionnaires [23]. In psychometrics, and behavioural economics,

many latent traits tend to be inter-correlated, which can be captured by non-orthogonal rotations in factor analysis, like "oblimin" or "promax".

In summary, each step of this process is increasingly more able to capture non-linear, higher dimensional aspects of the feature-outcome associations. However, in **Step 5 - Plotting** and **Step 6 - Theory-Driven decision-making**, we inject human perspective, expert-knowledge, and theory into the process again. In many ways, this even captures higher complexity, since it allows to step away from a pure data-driven process, and align all steps with both research perspective, and strategic alignment.

3.4.1 Software Used

Data manipulations have been conducted with Python 3.8.9 [54], Pandas 2.1.3 [55], and calculations have been conducted in SciPy 1.11.4 [56], numpy 1.26.0 [57], and statsmodels 0.14.0 [58]. All graphs have been plotted with Matplotlib 3.8.2 [59], GeoPandas 0.14.1 [60], and seaborn [61].

4 Results

4.1 Outcome Measure

As discussed above, outcome measures are:

1. First phase = difference in days between the inflection points of vaccine uptake of the first and the third dose to represent the general psychological readiness of the population to take the vaccine (intrinsic motivation)
2. Second phase = difference in days between the inflection points of vaccine uptake of the first and the fifth dose to represent the effect of secondary measures (e.g., persuasion/ extrinsic motivation)
3. Agent readiness = survey answers to the question "I abide by governmental measures", since that encompasses getting a vaccine.

As expected, all vaccine uptake curves follow a sigmoid shape with clear inflection points. Figure 4 displays a random example to illustrate population vaccine uptake behaviour.

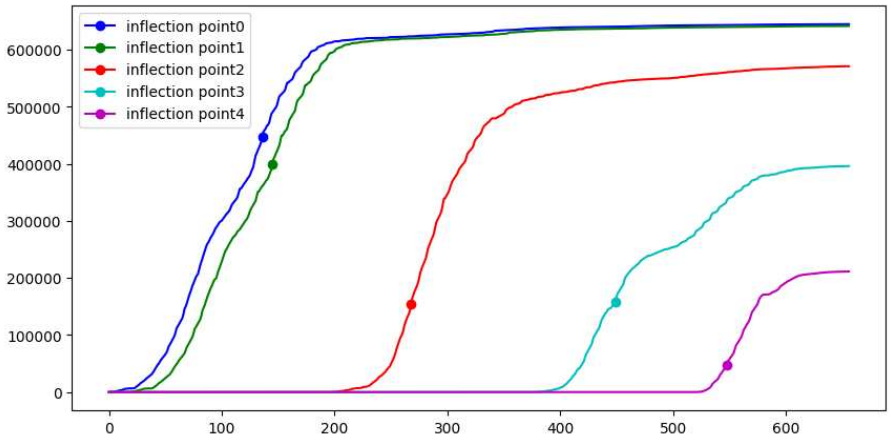

Fig. 4 Inflection points of vaccine uptake curves Okinawa

Also, in accordance with theory, there are clear differences in the uptake time between the fist and the third, and between the first and the fifth dose. These differences display clear variance across prefectures, as depicted in figure 5.

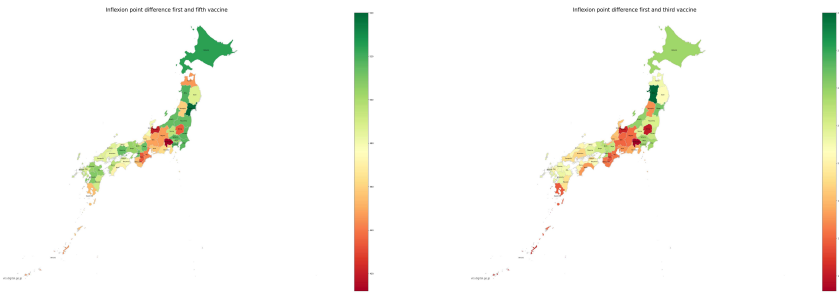

Fig. 5 Comparison uptake times 1st to 5th and 1st to 3rd vaccine

Since our inflection-point-based approach already captures various probabilistic issues like different infrastructure or vaccine availability, *ceteris paribus* this pattern can best be explained by the effect of geospatial personality distribution or spatiotemporal fluctuations of the information field.

4.2 Statistical Feature Reduction Pipeline and Regression Results

We use the **Statistical Feature Reduction Pipeline** described above to manually "simulate" automated feature selection. For that, we extract those variables that have both correlation coefficients r_p and r_s above .7 or below -.7, an Euclidean Distance Correlation of above .7, and a MIC of above .7.

As expected, the factor analysis does not result in any meaningful result due to the unbalanced sample, with 46 factors of an Eigenvalue above 1 in the exploratory run, thus neither interpretable nor useful. This resulted in about 30 useful predictors for each relevant outcome variable, which we then used for manual, theory-driven feature selection. Most plots appear to have a linear relationship with the outcome measures, wherefore we decide for a linear OLS regression analysis.

We designate the outcome of the **first model: predicting mid-term vaccine uptake** as "Cluster of the Lower Maslow Pyramid", since it is comprised of the following variables: the number of severe cases during the third phase of the pandemic, LIWC scores on "risk" before the onset of the COVID-19 pandemic and during the first wave, "negative emotions" during the first wave, "feelings" during the second wave, and "family" during the second and third wave. We interpret this as communication patterns that make people aware of risks, trigger interest and sense of urgency and existential fear through preceding negative emotions in the onset of COVID-19, and about micro-systemic elements during the second and third phase, when uncertainty and fear during the pandemic peaked. Interestingly, no personality results are significant for this most important measure, and it is purely based on predictors from the information field, which are both proximal and relevant to primeval fears of survival on an individual- and micro-systemic level.

Table 1 OLS Regression Results "Vaccine Uptake Mid-Term"

Statistic	**Value**	**Statistic**	**Value**
Dep. Variable:	Vaccine Uptake Mid-Term	R-squared:	0.592
Model:	OLS	Adj. R-squared:	0.519
Method:	Least Squares	F-statistic:	8.090
Date:	Tue, 28 Mar 2023	Prob (F-statistic):	4.66e-06
Time:	00:14:52	Log-Likelihood:	-209.70
No. Observations:	47	AIC:	435.4
Df Residuals:	39	BIC:	450.2
Df Model:	7		
Covariance Type:	nonrobust		

Variable	**coef**	**std err**	**t**	**P>t**	**[0.025**	**0.975]**
const	-368.5625	116.788	-3.156	0.003	-604.788	-132.337
$risk_0$	1337.4154	410.808	3.256	0.002	506.477	2168.354
$risk_1$	-1654.1748	421.570	-3.924	0.000	-2506.880	-801.470
$negemo_1$	410.0607	99.189	4.134	0.000	209.432	610.689
$feel_2$	1254.5321	254.711	4.925	0.000	739.331	1769.733
$family_2$	-700.7044	228.158	-3.071	0.004	-1162.198	-239.211
$family_3$	392.4999	172.348	2.277	0.028	43.893	741.107
$severe_3$	7.3231	1.551	4.723	0.000	4.187	10.459

Omnibus:	3.147	Durbin-Watson:	1.811
Prob(Omnibus):	0.207	Jarque-Bera (JB):	2.244
Skew:	0.273	Prob(JB):	0.326
Kurtosis:	3.921	Cond. No.:	469

Notes: [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

We designate the **second model: predicting long-term vaccine uptake** as "Cluster of Social Learning Effects", since it is comprised of the following LIWC scores: "negative emotions" before the onset of the COVID-19 pandemic and during its first wave, "anxiety" during the fourth wave, plus agent Extraversion, and Extraversion of the contextual embedding. It shows that such people take the fifth dose that are primed normatively towards negative emotions, then are exposed in the fourth wave with messages of anxiety, and who are very extraverted, in extraverted environments. This is interesting, since it indicates that regional extraversion improves social learning from others and the creation of normative believes, plus, synergetic with agent extraversion, fosters a climate of mutual exchange through human contact. It has to be stated though that the information field is orders of magnitude stronger than personality factors. However, results give indication for the correctness of the assumed event horizon, since it covers the first half of the time from the onset of the pandemic to the first availability of vaccines.

Table 2 OLS Regression Results "Vaccine Uptake Long-Term"

Statistic	**Value**	**Statistic**	**Value**
Dep. Variable:	Vaccine Uptake Long-Term	R-squared:	0.431
Model:	OLS	Adj. R-squared:	0.362
Method:	Least Squares	F-statistic:	6.214
Date:	Tue, 28 Mar 2023	Prob (F-statistic):	0.000225
Time:	00:30:58	Log-Likelihood:	-214.01
No. Observations:	47	AIC:	440.0
Df Residuals:	41	BIC:	451.1
Df Model:	5		
Covariance Type:	nonrobust		

Variable	**coef**	**std err**	**t**	**P>t**	**[0.025**	**0.975]**
const	-361.8265	251.280	-1.440	0.157	-869.297	145.644
$negemo_0$	1393.6929	327.229	4.259	0.000	732.840	2054.546
$negemo_1$	-1426.4957	338.818	-4.210	0.000	-2110.752	-742.239
anx_4	789.6031	296.970	2.659	0.011	189.859	1389.347
$E_{context}$	86.9427	40.665	2.138	0.039	4.818	169.068
E_{agent}	133.0699	51.547	2.582	0.014	28.969	237.171

Omnibus:	0.144	Durbin-Watson:	2.444
Prob(Omnibus):	0.930	Jarque-Bera (JB):	0.341
Skew:	-0.073	Prob(JB):	0.843
Kurtosis:	2.609	Cond. No.:	718

Notes: [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Finally, we designate the **third model: comply with governmental measures** as "Anatomy of Extraverted Agents", since it is comprised of the following LIWC scores: "affect" before the onset of the pandemic, during the second and fourth wave, "negative emotion" before the onset, during the first wave, and "anxiety" during the fourth wave. Furthermore, as with the second model, agent and contextual Extraversion are significant predictors, as well as survey results that indicate that agents find governmental measures justified. A further similarity is that strongest effects are negative emotions during the

early onset of the pandemic, however, the difference is not as strong as with the second model. In many ways, this third model is most interesting, since it predict survey results with both agent context and agent-internal variables, and thus offers unique insight into agents and not only an overview over mass behaviour. As expected, its explanatory power is much higher ($R^2 = 0.836$), and it offers a better temporal granulation, thus indicating that strong initial communication is imperative for reaching strongest health management results. Furthermore, it also provides partial evidence for our proposed framework, since it shows that steady, relevant information, plus proper contextual embedding and agent cognition promotes vaccine uptake; however, initial information seems to yield the strongest influence, which contradicts framework assumptions of current events being more influential to an agent.

Table 3 OLS Regression Results "Abiding by governmental measures"

Statistic	**Value**	**Statistic**	**Value**
Dep. Variable:	Abiding	R-squared:	0.836
Model:	OLS	Adj. R-squared:	0.796
Method:	Least Squares	F-statistic:	20.96
Date:	Tue, 28 Mar 2023	Prob (F-statistic):	5.74e-12
Time:	00:31:01	Log-Likelihood:	85.719
No. Observations:	47	AIC:	-151.4
Df Residuals:	37	BIC:	-132.9
Df Model:	9		
Covariance Type:	nonrobust		

Variable	**coef**	**std err**	**t**	**P>t**	**[0.025**	**0.975]**
const	-1.1935	0.507	-2.354	0.024	-2.221	-0.166
affect_0	0.2880	0.087	3.319	0.002	0.112	0.464
negemo_0	2.3093	0.741	3.115	0.004	0.807	3.812
negemo_1	-2.4781	0.758	-3.269	0.002	-4.014	-0.942
affect_2	0.3836	0.104	3.673	0.001	0.172	0.595
affect_4	-0.3099	0.083	-3.731	0.001	-0.478	-0.142
anx_4	1.8305	0.543	3.371	0.002	0.730	2.931
$\text{E}_{context}$	0.2420	0.075	3.238	0.003	0.091	0.393
E_{agent}	0.2268	0.096	2.374	0.023	0.033	0.420
justified	0.4680	0.058	8.039	0.000	0.350	0.586

Omnibus:	1.115	Durbin-Watson:	2.028
Prob(Omnibus):	0.573	Jarque-Bera (JB):	0.938
Skew:	0.071	Prob(JB):	0.626
Kurtosis:	2.322	Cond. No.:	1.56e+03

Notes: [1] Standard Errors assume that the covariance matrix of the errors is correctly specified. [2] The condition number is large, 1.56e+03. This might indicate that there are strong multicollinearity or other numerical problems.

5 Discussion

We demonstrate that spatiotemporal measures can can be successfully used to explain and predict economic behaviour like population abidance with health regulations and governmental measures. For that, we use data from SNS to model the information field based on considerations from systems theory and

brain sciences. We furthermore use inflection point differences of sigmoid-shaped vaccine uptake curves as a simple mean to identify the point when the first momentum is exhausted and psychological and normative processes kick-in. This formulates a distribution of speed differences over Japan, which look dramatically different between the first and the fifth than between the first and the third dose. These differences can only be explained by psychological factors, given the homogeneous, highly efficient, and industrialised infrastructure of Japan.

We furthermore suggest a dynamic, spatiotemporal framework for better understanding and operationalisation of the information field, and enabling future research to be based on a solid foundation. Also, this framework enables us to discuss potential attachment points of the information field to individual-psychological architectures of agents. Furthermore, it has the advantage over models solely based on time or space that it is able to model ripple effects in the information field over space and time and explain congruent behaviour changes. To facilitate ease of operationalisation of this framework, we suggest a new way to interpret LIWC scores through a system-theoretic perspective. This allows us to simplify the data structure we operate on, however still leads to an unbalanced sample, which we tackle with a semi-manual statistical feature reduction pipeline.

Finally, we show that agent and embedding personality are small but important factors to predict both individual and collective behaviour. Extraversion emerges as one of the leading factors of health behaviour, which we interpret as behaviour of actively seeking both the proximity and opinion of others, and thus leads to more spread of information and a synchronisation of ideas. Unfortunately, aligned with findings from media psychology [15], we also find that negative and relevant information in terms of proximity and personal involvement, as well as priming effects play a role in vaccine uptake speed. However, on top of that we find, that strong communication at the onset of the pandemic is the biggest contributing factor to abiding with governmental measures and vaccine uptake. Those findings offer new pathways to tailor messaging from the government and health authorities to better mobilise a population towards health behaviour and thus may contribute to a safer society. However, nowhere else than in public communication is the line between help and manipulation thinner, and nowhere else can wrong approaches undermine public trust. Hence, those findings are powerful but need more rigorous confirmatory research, and a strong ethics debate, should one want to use it for public relations.

6 Limitation and Outlook

The strength of this study – the introduction of various simple yet effective methods to represent a framework for information field effectiveness – is also its biggest weakness, since several assumptions need to be done for that. Each of these assumptions (e.g., that LIWC categories represent geospatial distance)

would need further research to hold. Hence, we cannot say that the *ceteris paribus* approach is fulfilled, and further research on each sub-component of the framework and methods needs to be conducted to stress test causality, robustness and generalisability. This is especially important for our finding that information at the onset of the pandemic has a stronger effect than such at later stages, which contradicts framework assumptions, but could be explained by priming effects since the pandemic represents a strong exogenous shock. Also, due to the unbalanced sample and theory-driven approach, some effects might be invisible. Hence further, more automated, and more extended analysis on those parts of the data that exists in deeper granularity (e.g., survey results on city level) should be conducted to inject more data-driven information into the outcomes. Also, more research on the forecasting aspects of the model should be conducted, which either uses a different method like topic modelling for identifying future-oriented language (which does not exist in the Japanese version of LIWC), or this research should be repeated in another language where those missing LIWC features do exist.

7 Conclusion

We introduce and find evidence for a framework for understanding the effect of the information field on both individual agents as on aggregate groups of agents, based on information theory, systems theory, psychometrics, and behavioural economics. Furthermore, we introduce a novel way to allocate physical proximity by LIWC word categories, and use that effectively to find evidence for the framework. Unfortunately, future prediction is not possible to explore since J-LIWC2015 does not provide time categories; partially due to peculiarities in the Japanese language. We also introduce a simplified pipeline for manual statistical feature reduction for unbalanced samples, or when theory needs to be injected and results need to be aligned with strategic imperative. Finally, we use that framework and the methodology successfully to identify such individual-psychological factors in agents, aggregate regional psychology, and the information field that positively influence vaccine uptake. This enables novel, more precise and effective ways to encourage vaccine uptake, health behaviour, tailor and precision-apply governmental messages to combat fake news, and contribute towards a more healthy and robust society.

Acknowledgments. We thank Shino Takishita and Rieko Okada, who helped managing the survey company. Further, we thank KC Chen from Manzanita Intelligent Marketing for initial data science advice for SNS scraping, and Stephen Fitz for participating in initial discussions.

The main author of this research was supported by the Keio University Academic Development Funds for Individual Research, and The Keio University Ushioda Memorial funds.

This paper was made possible through the research grant provided by the Tokyo Center for Economic Research.

Declarations

Authors' contributions

Peter Romero: conceptualisation (psychometrics, NLP, and formal analysis), methodology, formal analysis, data curation, writing, visualisation, team management, and overall project management. Eisaku Tanaka collected tweets, cleaned and analysed SNS data. Yuki Mikiya helped managing the survey and translating the items. Atsushi Oshio and Shinya Yoshino provided the data set for psychological ground truth data. Teruo Nakatsuma supervised the project.

Inclusion and Ethics Statement

All contributors that fulfill all authorship criteria are included as authors. All others are listed in the Acknowledgements section. No local researchers have been involved. Roles and responsibilities were agreed amongst collaborators ahead of the research. No ethical guidelines in the setting of the authors or contributors exist that would severely restrict or prohibit this research, hence no ethics review committee needed to be consulted. Local and regional research was taken into account.

Conflict of interest/Competing interests

The authors have no monetary or organisational connection with IBM, or any other provider of psychometric assessments or software deployed in this paper.

References

[1] Forman, R., Shah, S., Jeurissen, P., Jit, M., Mossialos, E.: Covid-19 vaccine challenges: What have we learned so far and what remains to be done? Health policy **125**(5), 553–567 (2021)

[2] MacIntyre, C.R.: Increasing the uptake of vaccination against infectious diseases. Medical Journal of Australia (2015)

[3] Nehal, K.R., Steendam, L.M., Campos Ponce, M., van der Hoeven, M., Smit, G.S.A.: Worldwide vaccination willingness for covid-19: a systematic review and meta-analysis. Vaccines **9**(10), 1071 (2021)

[4] Yoo, S., Gretzel, U.: The influence of culture on consumer sensitivity to health communication: a multilevel study of the impact of individualism-collectivism. PsyEcology **7**(3), 251–279 (2016)

[5] Betsch, C., *et al.*: Social norms and vaccination decisions. The Lancet **379**(9829), 1855–1856 (2012)

[6] Giancola, M., Palmiero, M., D'Amico, S.: Dark triad and COVID-19 vaccine hesitancy: the role of conspiracy beliefs and risk perception (2023). https://doi.org/10.1007/s12144-023-04609-x. Accessed 2023-11-12

[7] Howard, M.C.: The good, the bad, and the neutral: Vaccine hesitancy mediates the relations of psychological capital, the dark triad, and the big five with vaccination willingness and behaviors **190**, 111523 (2022). https://doi.org/10.1016/j.paid.2022.111523. Accessed 2023-11-12

[8] Douglas, K.M., Uscinski, J.E., Sutton, R.M., Cichocka, A., Nefes, T., Ang, C.S., Deravi, F.: Understanding conspiracy theories **40**, 3–35 (2019). https://doi.org/10.1111/pops.12568. Accessed 2023-11-12

[9] Oortwijn, R.: How openness to experience relates to conspiracy mentality and vaccine hesitancy. Economic Psychology Available online at: http://arno.uvt.nl/show.cgi (2020)

[10] Li, T.Y., de Girolamo, G., Zamparini, M., Malvezzi, M., Candini, V., Calamandrei, G., Starace, F., Zarbo, C., Götz, F.M.: Openness buffers the impact of belief in conspiracy theories on COVID-19 vaccine hesitancy: Evidence from a large, representative italian sample **208**, 112189 (2023). https://doi.org/10.1016/j.paid.2023.112189. Accessed 2023-11-12

[11] Sherman, R.A., Nave, C.S., Funder, D.C.: The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. Advances in experimental social psychology **52**, 71–115 (2016)

[12] Brennen, J.S., Simon, F.M., Howard, P.N., Nielsen, R.K.: Types, sources, and claims of covid-19 misinformation. Reuters Institute (2020)

[13] Loomba, S., de Figueiredo, A., Piatek, S.J., de Graaf, K., Larson, H.J.: Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. Nature Human Behaviour **5**, 337–348 (2021)

[14] Chou, W.-Y.S., Gaysynsky, A., Vanderpool, R.C.: Media and misinformation in the time of covid-19. Journal of Health Communication **25**(10), 760–763 (2020)

[15] Harcup, T., O'neill, D.: What is news? news values revisited (again). Journalism studies **18**(12), 1470–1488 (2017)

[16] Romero, P., Mikiya, Y., Nakatsuma, T., Fitz, S., Koch, T.: Modelling Personality Change During Extreme Exogenous Conditions. PsyArXiv (2021). https://doi.org/10.31234/osf.io/rtmjw. https://psyarxiv.com/rtmjw/ Accessed 2022-06-28

[17] Neff, T., Kaiser, J., Pasquetto, I., Jemielniak, D., Dimitrakopoulou, D., Grayson, S., Gyenes, N., Ricaurte, P., Ruiz-Soler, J., Zhang, A.: Vaccine hesitancy in online spaces: A scoping review of the research literature, 2000-2020. Harvard Kennedy School Misinformation Review (2021). https://doi.org/10.37016/mr-2020-82. Accessed 2023-11-29

[18] Peters, H., Götz, F.M., Ebert, T., Müller, S.R., Rentfrow, P.J., Gosling, S.D., Obschonka, M., Ames, D., Potter, J., Matz, S.C.: Regional personality differences predict variation in early covid-19 infections and mobility patterns indicative of social distancing. Journal of Personality and Social Psychology **124**(4), 848 (2023)

[19] Obschonka, M., Stuetzer, M., Rentfrow, P.J., Lee, N., Potter, J., Gosling, S.D.: Fear, populism, and the geopolitical landscape: The "sleeper effect" of neurotic personality traits on regional voting behavior in the 2016 brexit and trump elections **9**(3), 285–298 (2018). https://doi.org/10.1177/1948550618755874. Accessed 2018-12-11

[20] Yoshino, S., Shimotsukasa, T., Hashimoto, Y., Oshio, A.: The association between personality traits and hoarding behavior during the covid-19 pandemic in japan. Personality and individual differences **179**, 110927 (2021)

[21] Mangalik, S., Eichstaedt, J.C., Giorgi, S., Mun, J., Ahmed, F., Gill, G., Ganesan, A.V., Subrahmanya, S., Soni, N., Clouston, S.A., et al.: Robust language-based mental health assessments in time and space through social media. arXiv preprint arXiv:2302.12952 (2023)

[22] Willke, H.: Systemtheorie 1. Grundlagen. UTB, ??? (2000)

[23] Rust, J., Kosinski, M., Stillwell, D.: Modern Psychometrics: The Science of Psychological Assessment, 4th edn. Routledge, Fourth edition. | Milton Park, Abingdon, Oxon ; New York, NY: Routledge, 2021. (2020). https://doi.org/10.4324/9781315637686. https://www.taylorfrancis.com/books/9781317268772 Accessed 2023-08-26

[24] Zullinger, E.M., Ricklefs, R.E., Redford, K.H., Mace, G.M.: Fitting sigmoidal equations to mammalian growth curves **65**(4), 607–636 (1984). https://doi.org/10.2307/1380844. Accessed 2023-11-28

[25] Minakov, V., Makarchuk, T., Minakova, T., Kostin, V., Lobanov, O.: The expansion of time series innovations in a series of sigmoid. International Journal of Applied Business and Economic Research **15**(18), 311–319 (2017)

[26] Wood, A.J., MacKintosh, A.M., Stead, M., Kao, R.R.: Predicting Future

Spatial Patterns in COVID-19 Booster Vaccine Uptak. https://doi.org/10.1101/2022.08.30.22279415. Type: article. https://europepmc.org/article/PPR/PPR539178 Accessed 2023-11-28

[27] Houghton, J., Siegel, M., Goldsmith, D.: Modeling the influence of narratives on collective behavior, 24 (2013)

[28] Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K.: The development and psychometric properties of liwc2015 (2015)

[29] Statistics Bureau, M.o.I.A., with collaboration of Ministries, C., Agencies: Regional Statistics Database (System of Social and Demographic Statistics). https://www.e-stat.go.jp/en/regional-statistics/ssdsview Accessed 2021-06-13

[30] Pennebaker, B.R.J.B.R.L..F.M.E. J. W.: Linguistic Inquiry and Word Count: LIWC 2015 [computer Software]

[31] Igarashi, T., Okuda, S., Sasahara, K.: Development of the japanese version of the linguistic inquiry and word count dictionary 2015 (j-LIWC2015) (2021). https://doi.org/10.31234/osf.io/5hq7d

[32] Kudo, T.: Mecab: Yet another part-of-speech and morphological analyzer. http://mecab. sourceforge. net/ (2005)

[33] Costa, P.T., McCrae, R.R.: The revised neo personality inventory (neo-pi-r). The SAGE handbook of personality theory and assessment **2**, 179–198 (2008)

[34] McCrae, R.R., *et al.*: Neo-pi-r data from 36 cultures: Further intercultural comparisons. In International perspectives on psychological science **2**, 105–125 (2004)

[35] McCrae, R.R., Costa, P.T.: Validation of the five-factor model of personality across instruments and observers. Journal of personality and social psychology **52**(1), 81 (1987)

[36] Costa, P.T., McCrae, R.R.: The neo personality inventory manual (1985)

[37] Costa, P.T. Jr., McCrae, R.R.: Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional Manual. Psychological Assessment Resources, Odessa, FL (1992). Psychological Assessment Resources

[38] Yoshino, S., Oshio, A.: Regional differences in Big Five personality traits in Japan. Japanese Journal of Environmental Psychology **9**(1), 19–33 (2021). https://doi.org/10.20703/jenvpsy.9.1_19

[39] Oshio, A., Abe, S., Cutrone, P.: Development, reliability, and validity of the japanese version of ten item personality inventory (TIPI-j). **21**(1), 40–52 (2012). Accessed 2022-06-01

[40] Gosling, S.D., Rentfrow, P.J., Swann, W.B.: A very brief measure of the big-five personality domains **37**(6), 504–528 (2003). https://doi.org/10.1016/S0092-6566(03)00046-1. Accessed 2022-06-01

[41] MHLW: Press Conference from the Ministry of Health, Labour and Welfare. https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/newpage_00032.html Accessed 2021-07-18

[42] WHO: Japan: World Health Organization Coronavirus Disease (COVID-19) Dashboard With Vaccination Data. https://covid19.who.int Accessed 2021-07-18

[43] Fávero, L.P., Belfiore, P., Souza, R.d.F.: Chapter 12 - principal component factor analysis. In: Fávero, L.P., Belfiore, P., Souza, R.d.F. (eds.) Data Science, Analytics and Machine Learning with R, pp. 203–214. Academic Press. https://doi.org/10.1016/B978-0-12-824271-1.00006-8. https://www.sciencedirect.com/science/article/pii/B9780128242711000068

[44] de Winter, J.C.F., Gosling, S.D., Potter, J.: Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data **21**(3), 273–290 (2016). https://doi.org/10.1037/met0000079. Place: US Publisher: American Psychological Association

[45] Székely, G.J., Rizzo, M.L., Bakirov, N.K.: Measuring and testing dependence by correlation of distances. Annals of Statistics **35**(6), 2769–2794 (2007). https://doi.org/10.1214/009053607000000505

[46] Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M., Sabeti, P.C.: Detecting novel associations in large datasets **334**(6062), 1518–1524. https://doi.org/10.1126/science.1205438. Accessed 2023-11-26

[47] Kinney, J.B., Atwal, G.S.: Equitability, mutual information, and the maximal information coefficient **111**(9), 3354–3359 (2014). https://doi.org/10.1073/pnas.1309933111. Publisher: Proceedings of the National Academy of Sciences. Accessed 2023-11-28

[48] Chauhan, N., Choi, B.-J.: Classification of alzheimer's disease using maximal information coefficient-based functional connectivity with an extreme learning machine. Brain Science **13**(7), 1046 (2023). https://doi.org/10.3390/brainsci13071046

[49] Lazarsfeld, J., Johnson, A., Adéníran, E.: Differentially private maximal information coefficients. In: International Conference on Machine Learning (2022). https://doi.org/10.48550/arXiv.2206.10685

[50] Zhou, P., Zhang, Y., Yan, Y.-T., Zhao, S.: Unknown type streaming feature selection via maximal information coefficient. IEEE (2022). https://doi.org/10.1109/ICDMW58026.2022.00089

[51] Mathai, A.M.: Factor analysis revisited. Journal of Statistical Theory and Practice (2021). https://doi.org/10.1007/S42519-020-00160-1

[52] Chandra, N.K., Dunson, D.B., Xu, J.: Inferring covariance structure from multiple data sources via subspace factor analysis. arXiv:2305.04113 (2023). https://doi.org/10.48550/arxiv.2305.04113

[53] Stegle, O., Rohan, T.M.: Identifying temporal and spatial patterns of variation from multimodal data using mefisto. Nature Methods (2022). https://doi.org/10.1038/s41592-021-01343-9

[54] Python Software Foundation: Python. https://www.python.org/

[55] pandas development team, T.: Pandas-dev/pandas: Pandas. https://doi.org/10.5281/zenodo.3509134. https://doi.org/10.5281/zenodo.3509134

[56] Virtanen, P., et al.: SciPy: Open Source Scientific Tools for Python

[57] Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E.: Array programming with NumPy. Nature **585**(7825), 357–362 (2020). https://doi.org/10.1038/s41586-020-2649-2

[58] Seabold, S., Perktold, J.: statsmodels: Econometric and statistical modeling with python. In: 9th Python in Science Conference (2010)

[59] Hunter, J.D., et al.: Matplotlib: Visualization with Python

[60] den Bossche, J.V., et al.: GeoPandas

[61] Waskom, M.L.: seaborn: statistical data visualization. Journal of Open Source Software **6**(60), 3021 (2021). https://doi.org/10.21105/joss.03021