Vector-based pedestrian navigation in cities

(Article begins on next page)

28 April 2024

Vector-based Pedestrian Navigation in Cities

Christian Bongiorno[1,2], Yulun Zhou[1,3], Marta Kryven[4], David Theurel[5], Alessandro Rizzo[6], Paolo Santi[1,7], Joshua Tenenbaum[4] & Carlo Ratti[1]

[1] *Senseable City Lab, Massachusetts Institute of Technology, 77 Mass Av, 02139, Cambridge, MA, US.*

[2] *Université Paris-Saclay, CentraleSupélec, Mathématiques et Informatique pour la Complexité et les Systèmes, 91190, Gif-sur-Yvette, France.*

[3] *Department of Urban Planning and Design, Faculty of Architecture, The University of Hong Kong, Pokfulam, Hong Kong, China*

[4] *Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 77 Mass Av, 02139, Cambridge, MA, US*

[5] *Department of Physics, Massachusetts Institute of Technology, 77 Mass Av, 02139, Cambridge, MA, US*

[6] *Dipartimento di Elettronica e Telecomunicazioni, Politecnico di Torino, 10128 Torino TO, Italy Office of Innovation, New York University Tandon School of Engineering, Six MetroTech Center, Brooklyn NY 11201, US.*

[7] *Istituto di Informatica e Telematica del CNR, Via G. Moruzzi 1, 56124, Pisa, Italy.*

How do pedestrians choose their paths within city street networks? Researchers have tried to shed light on this matter through strictly controlled experiments, but an ultimate answer based on real-world mobility data is still lacking. Here, we analyze salient features of human

path planning through a statistical analysis of a massive dataset of GPS traces, which reveals that (1) people increasingly deviate from the shortest path when the distance between origin and destination increases, and (2) chosen paths are statistically different when origin and destination are swapped. We posit that direction to goal is a main driver of path planning and develop a vector-based navigation model that is a statistically better predictor of human paths than a model based on minimizing distance with stochastic effects. Our findings generalize across two major US cities with different street networks, hinting to the fact that vector-based navigation might be a universal property of human path planning.

1 Introduction

Although path planning is one of the hardest problems to solve computationally, [1] humans plan remarkably efficient paths when navigating cities, and do so at various scales. While human path planning can be near-optimal, it also exhibits systematic divergences from the shortest available path, [2] [3,4] which are still not well understood. We hypothesise that a contribution to such divergences arise from a common mental computational mechanism, which is shared between humans, can be modelled in precise quantitative terms, and generalises across urban environments. Describing this mechanism by a formal computational account can help explain mental computations in the brain that support human mobility, inform the design of real-time planning tools that can better couple human and machine intelligence, and improve urban planning.

In the laboratory setting humans often rely on the *approximate rationality principle*, meaning that humans use approximate planning heuristics to maximise their goals, while limiting subjec-

tive costs. [5–9] In the real world these costs may be a combination of mental and physical effort [10]; for example, the mental cost of planning a route comes with the physical cost of travel. Multiple studies have investigated aggregate mobility flows [11–21] and cognitive abilities that support navigation. [22–29] Quantitative models of how humans may plan their routes in real cities have also been proposed, although limited to a small neighborhood, in which the planning problem could be optimally solved with exhaustive route enumeration by Breadth First Search. [4] However, generalizable computational models, that can generate precise quantitative predictions in large-scale city environments, are still lacking. Thus, it is still unclear which subjective costs and planning heuristics may explain human routes in real urban environments.

We investigate this question by analyzing a large dataset of GPS traces of 552,478 pseudo-anonymised human paths undertaken by 14,380 pedestrians in two major US cities – Boston and San Francisco (see Methods). The original dataset was retrieved from a company used to run one of the largest mobile apps for mobility tracking. With full consent from users, the app records high-resolution trajectories of human movements in complex urban environments in their daily life. Activity types have been labeled by the company using proprietary machine learning tools, and only activities labeled as "walking path" were analyzed in this work. Compared to previous studies of aggregate human mobility, which mostly rely on sparse location sampling – such as the serial numbers of US banknotes, [12] surveys ,[18,30] commuting trips, [20] mobile communication records, [13] social-media check-ins, [31,32] or location history [15,17] – we study high resolution routes of individual pedestrians, reconstructed from their GPS traces. Thanks to pseudo-anonymisation, which assigns a unique anonymous ID to each individual in the data set, we could also associate

multiple trajectories to the same individual, allowing the study of individual-level properties of pedestrian routes – see Supplementary, Section 5. Importantly, the majority of human trajectories in our dataset were substantially different from routes suggested by Google Maps (see Supplementary Tables 1, 2, 3 and Supplementary Figure 3), indicating a minimal bias introduced by machine-generated trajectories on our data. This property, along with recordings of multiple trips by the same individual, and between the same locations, enabled us to fit and evaluate alternative quantitative models of pedestrian navigation. In particular, we restricted our attention to simple geometric, computational models that can be used to at least partially explain human navigation ability.

2   Results

Evaluating Paths Based on Distance As a first approximation, consider a simple way to formalise the cost of a path as equal to its distance:

$$C_{dist}[\mathcal{P}] = \sum_{S_i \in \mathcal{P}} l_i \ .$$

Here, $\mathcal{P}$ denotes a path connecting the origin to the destination, composed of a list of street segments $S_1, S_2, \ldots$, and l_i denotes the walking distance along segment S_i. While this simple model captures the motivation to reduce distance, which is central to human planning [33,34], several studies show that humans often deviate from the shortest paths [2,11,26,34,35].

Indeed, Figure 1-A shows that human paths recorded in our dataset were consistently longer than the shortest distance path, computed using the standard Dijkstra algorithm [36]. The tendency to

4
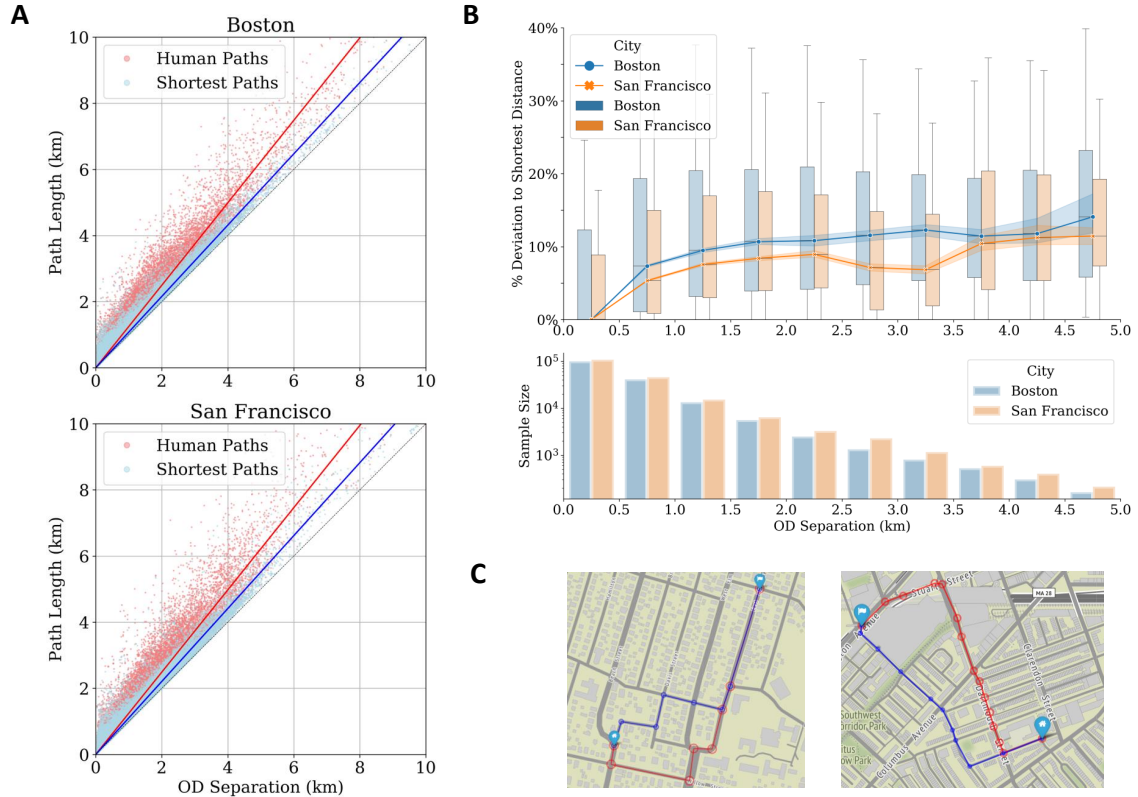
Figure 1: **Differences between human paths and shortest paths.** A. Aggregated comparisons between the path lengths of human and shortest paths (y-axis) in Boston and San Francisco, as a function of the Euclidean distance between origin and destination (x-axis); B. Relative differences between human and shortest path lengths (y-axis) as a function of the Euclidean distance between origin and destination (x-axis); C. Two examples of the difference between human paths (in red) and their corresponding shortest paths (in blue).

deviate from the shortest path increases with distance between origin and destination (see Figure 1-B), which could be due to the increasing complexity of evaluating relatively longer paths, in line with the approximate rationality principle. However, it is also interesting to observe that most of the relative deviation from shortest path is achieved by paths of length around 1 km, while only a modest further increasing deviation is observed for longer paths.

Stochastic Distance Minimization The increasing deviation from the shortest path observed in the data could arise from uncertainty about the lengths of street segments, which leads to an accumulation of errors over time. Formally, we can describe this process by an error term in the evaluation of street segment lengths:

$$l_i \rightarrow e^{\mathcal{N}(\log(l_i), \sigma)} = c(l_i) \ ,$$

and

$$C_{dist}[\mathcal{P}] = \sum_{S_i \in \mathcal{P}} c(l_i) \ , \tag{1}$$

where the new cost function $c(l_i)$ is obtained by applying a log-normally distributed random noise to the original street segment length l_i. The use of log-normal distribution to model uncertainty in street length estimation is motivated by the widely accepted Weber-Fechner law of just noticeable difference [37], which states that humans perceive measurable quantities on a logarithmic scale. We will refer to equation (1) as the *stochastic distance minimization* model. The number of street segments in the cost function (1) tends to increase with the Euclidean distance separating origin and destination. Hence, the deviation from the shortest distance tends to accumulate with increased separation between origin and destination, as we observed.

6

Importantly, in Methods we show that this model predicts path choices to be symmetrical – the available paths are ranked in the same order of preference regardless of travel direction, which may not be the case in human paths. Navigation asymmetry has been demonstrated in the laboratory [38][39–41], and empirically noted – although never statistically tested – in pedestrian flows [35] and driving routes [34]. A well-established theory of navigation by line of sight predicts asymmetric routes by suggesting that people travel along straight lines of sight in the desired direction, and when their view is obstructed, establish a new line of sight [11]. Empirical findings in neuroscience [42,43] and psychology [25,38] further suggest that neural and mental representations of space and direction could lead to asymmetry in two ways. Landmark-based navigation, extensively documented in humans [22,23,25,29] and animals, [44] may lead to asymmetric paths that depend on the distribution of views in the environment. Further, many animals, such as rodents, [45] bats, [46,47] and cats [48] rely on direction [42,43,45,49,50] for vector navigation[50,51]. Human subjects in the laboratory studies likewise rely on direction for navigation, taking a route to the first destination that begins in the direction of subsequent destinations[52], and rely on neural representations of both Euclidian distance, and direction to destination[42]. Several small-scale studies found that human routes exhibit systematic biases - such as bias toward paths that begin with a longer initial segment [39–41], or a southern bias [53], which may contribute to asymmetries. However, such biases have not yet been computationally modeled, nor quantitatively evaluated, in large-scare urban environments.

If present in human paths, asymmetric routes would falsify the stochastic distance minimization model as the only main contributor to pedestrian path formation. We tested for asymmetry as follows. First, we tested for asymmetric path choices within the same individual repeatedly vis-

iting any given two locations in either order. If present, such individual-level asymmetries could arise from a cognitive cost heuristic used to evaluate trajectories, which depends on direction, or from stochasticity in distance evaluation, followed by memorising the planned routes. Second, to disambiguate these two scenarios, we tested for asymmetries in repeated paths aggregated over different pedestrians. Such an asymmetry would indicate a persistent quality of human path planning, which cannot be explained by stochastic route learning, and thus may be attributed to a common direction-dependent cost heuristic. We have performed extensive statistical asymmetry tests at both individual and aggregate level (see Methods for details). The results of the test, reported in Figure 2, showed a statistically significant asymmetry in individual walking trajectories, as well as in aggregate trajectories, suggesting that asymmetries are a persistent quality of pedestrian navigation in urban street networks (see Methods for details).

Vector-based Navigation Model Other geometric models, such as the Initial Straightest Segment (ISS) strategy proposed by Bailenson et al. in [40], could be used to explain the observed asymmetry in human paths. If humans manifested a preference for the straightest first segment, we would expect to observe relatively longer first road segments in the human path than in the shortest path. However, human paths have consistently shorter initial segments than shortest paths, hinting to the fact that ISS strategy is likely not a cause of the observed asymmetry (see Supplementary Section 5.2).

Then, inspired by the evidence of asymmetry in human paths, and the prevalence of vector-navigation in animal models, we hypothesised that humans use direction when planning their paths.

A

i) < 250m

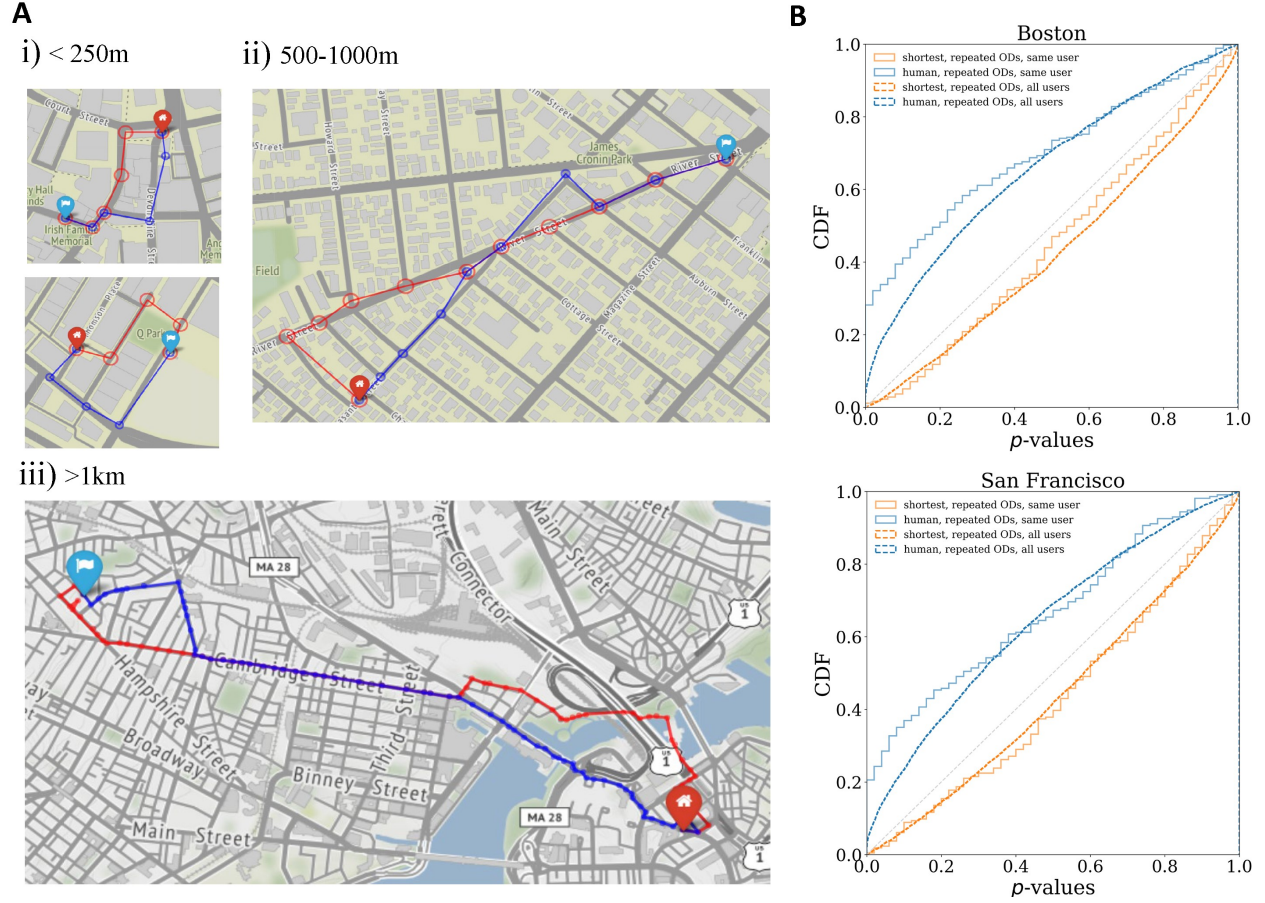ii) 500-1000m

iii) >1km

B

Figure 2: **Asymmetric human paths.** A. Examples of asymmetric human paths in street networks. Red paths start at red markers, blue paths start at blue markers. B. Cumulative Distribution Function (CDF) of the one-tail p-values for the asymmetry test in Boston and San Francisco (method section). p-values being consistently above the diagonal dashed line indicates a statistically significant deviation from symmetric paths. Solid lines refer to individual-level tests; dotted lines refer to aggregate-level tests. Shortest distance paths used in the tests are obtained by simulating the stochastic distance minimization model.

9

We formalise the *vector-based navigation model* as a cost that depends on the angular deviation of the street segment from the destination:

$$C_{dir}[\mathcal{P}] = \sum_{S_i \in \mathcal{P}} c(\theta_i, l_i) \,, \tag{2}$$

where

$$c(\theta_i, l_i) = e^{\mathcal{N}(\log(|\theta_i|l_i), \sigma)} \,,$$

and $\theta_i \in [-\pi, \pi]$ represents the angle between the tangent to the path at street segment S_i and the straight line to the destination (see Figure 3). In the example reported in Figure 3-a, the stochastic distance minimization model using Equation (1) slightly prefers the blue path (for which $\sum l_i = 935$m) over the red path ($\sum l_i = 938$m). The vector-based navigation model prefers the leanest red path ($\sum |\theta_i| l_i = 400$m·rad) over the blue path ($\sum |\theta_i| l_i = 516$m·rad). Since the angular deviations of street segments in this model depend on direction, the cost estimate of a path may change if its direction is reversed, implying that vector-based navigation could explain asymmetries observed in human trajectories (see Methods for a detailed proof.)

Model comparison Having formally defined the two models, we compare their explanatory power relative to each other. We aggregated all paths by OD distance separation, in steps of 50m, and estimated the most likely parameters for the stochastic distance minimization and for the vector-based model in each bin (see Methods). We measure the fraction of paths in each bin for which the vector-based model has a higher point-wise likelihood than the stochastic distance minimization model – we call this the *directional prevalence fraction* (DPF). A value higher than 50% would imply a higher probability that a class of paths is explained through the vector-based model than the

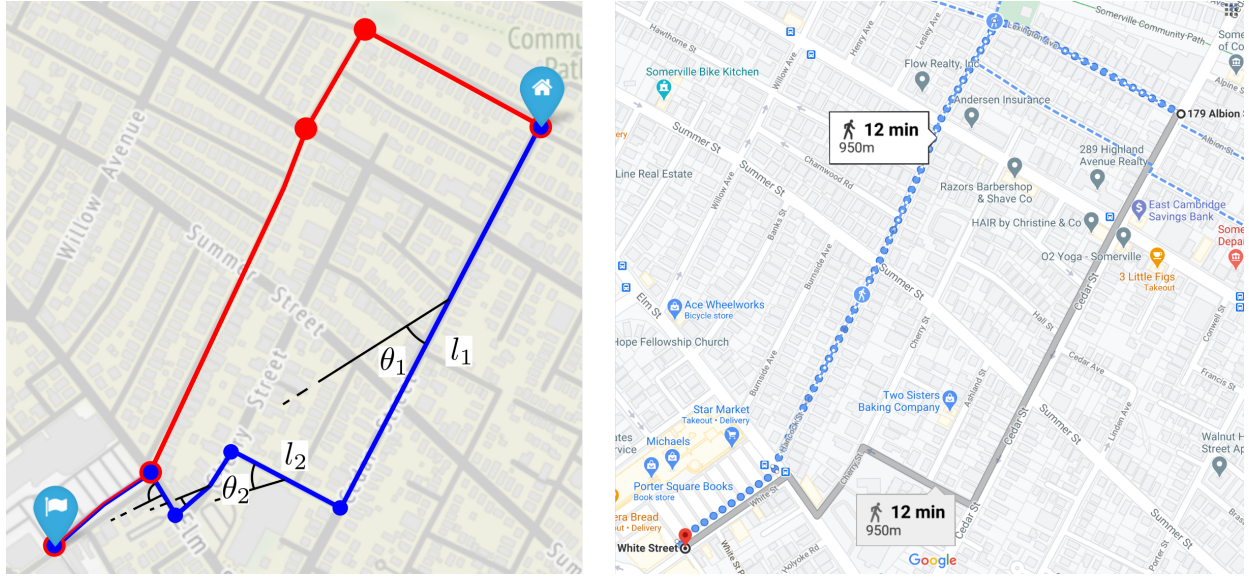Figure 3: **Model design.** Example illustrating the calculation of the vector-based cost approximated by the Equation (2). Left: the human path is shown in red; the shortest path is shown in blue. The quantities l_i, θ_i appearing in Equations (1, 2) are shown for the first street segment of the blue path. Right: path alternatives produced by Google maps. (Note that Google reports both path lengths as 950m.)

stochastic distance minimization model. The results reported in Figure 4 show that for each bin of OD distance separation with enough samples to reach statistical significance (see Supplementary Section 3) the DPF is consistently above 50%, suggesting the higher explanatory power of the vector-based model. The decreasing trend of the DPF function as OD distance separation increases can be consistently observed both in Boston and in San Francisco. Moreover, the value of the DPF function is very similar across the entire range of OD distance separation, ranging from a peak of 68% at 150 m separation – which corresponds to 35% better predictive power of vector-based vs. stochastic shortest distance model– to 53% at 1 Km separation. For the sake of completeness, the σ parameters on both models reach overall different values, which are 0.44 and 1.06 for the stochastic vector-based and shortest path model in Boston, and 0.42 and 1.1 in San Francisco. More details on the calibration process and the detailed shape of the σ function are reported in the Supplementary Figures 4 and 5.

The decreasing trend of the DPF function suggests a metacognitive mental computational mechanism, which trades off mental and physical costs during navigation. As the length of the planned path increases, humans likely shift from the relatively easier direction optimization strategy toward optimizing distance, in order to save physical effort and travel time. This hypothesis is also supported by the observed deviation of human trajectories from shorter paths, which tends to level-off for paths longer than 1 km – recall Figure 1. Observing this trend in two cities with drastically different street network topology suggests that this mental process generalises across city environments.
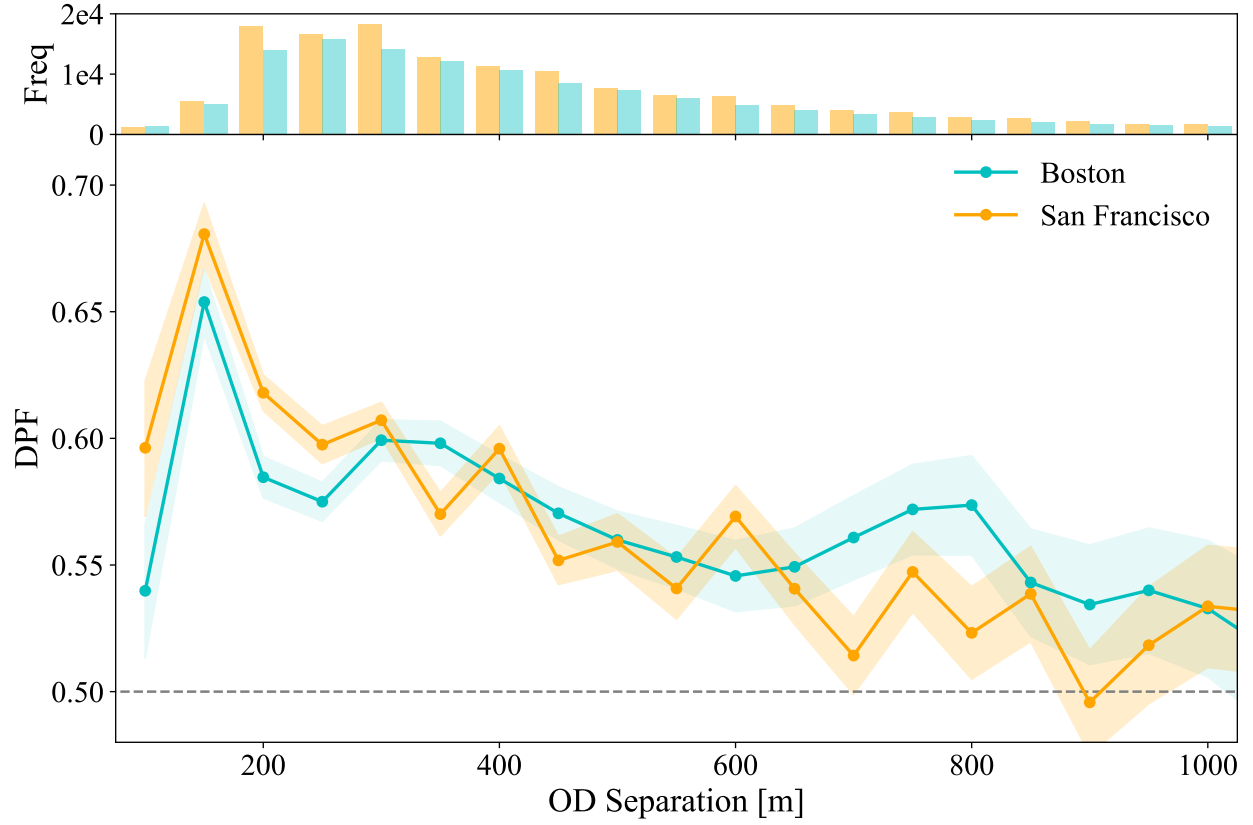
Figure 4: **Directional Prevalence Fraction (DPF) for different OD separations.** Top: marginal frequency distributions of paths in Boston and San Francisco. Bottom: Similar DPF patterns in Boston and San Francisco. Error bars are obtained with the 95% Wilson-score interval [54].

3 Discussion

The data-driven approach used in this paper illustrates how large-scale observations of behaviour in natural settings can be used to derive quantitative models of complex cognitive and physical tasks, and complements studies of cognition in the laboratory by providing unique insight into how the different cognitive faculties work together in real life. Despite all the challenges of an uncontrolled real city environment, as well as lack of information about users and their activities that limits the range of applicability of our findings, we have discovered behavioural trends which generalise across cities, attesting to the potential of big-data in the fields of psychology and cognitive science.

Our results suggest that vector navigation may be a common property of human route planning in cities, and establish a direct connection between the study of human mobility, human cognitive psychology, and neuroscience. The empirical support of the vector-navigation model by real human behaviour provides evidence in favour of theoretical claims that human minds and brains represent both Euclidean maps and graphs of street networks, as suggested in prior work[55]. The quantitative cost models described in this work suggest a possible computational mechanism by which route planning may be implemented in the brain. However, more work is needed to overcome the limitations of our study – lack of information about users and their activity – and understand, for instance, the effect of individual differences, the extent to which people pre-plan their routes, use hierarchical map representations, develop routing habits, and diversify path alternatives as a result of map learning.

The requirement to implement precise falsifiable models has limited our study to factors that

could be measured objectively and exactly – such as direction and distance – as well as a number of simplifying assumptions. While our models assume that pedestrians have accurate map knowledge, future work should consider how human path planning may depend on mental representations of the street networks, which may include larger streets and omit secondary ones[56]. Future work should also consider other factors that can influence navigation decisions, to the extent that these factors can be quantitatively tracked – such as day of the week, sunlight, weather, trees, attractions, presence of crowds, fatigue, time of the day, neighborhood safety, and elevation gradient – as well as individual differences in the responsiveness to these factors.

Our results extend our understanding of human route planning, and could have significant implications for those areas where route planning is a basic mechanism, such as transportation, cognitive maps, and real-time planning algorithms. We expect our methodology to become of even greater use with increasingly accurate 5G and 6G tracking data. The latter should allow monitoring which side of the street someone is walking, the time spent at the intersections, forward facing direction, and speed, supporting a more detailed investigation of the internal planning mechanisms. Future models should account for walking speed, tendency to change direction at a red light, as well as tendency to consistently orient toward specific visual landmarks. We hope that our findings will stimulate new research exploring these connections.

Methods

Dataset description and preparation The full dataset comprises 579,231 pseudo-anonymised human paths produced by 14,380 pedestrians, 5,590 in Boston and 8,790 in San Francisco, recorded by an always-on pedestrian tracking smartphone application over a time-period of one year. As customary in research based on data sets acquired through mobile applications, we do not have information about the demographics of the population of app users, which is then not necessarily representing an unbiased sample of the population in Boston and San Francisco. The two cities differ in street network geometry (see Supplementary Figure 1). San Francisco street network is designed as a grid, while Boston street network is highly irregular for historical reasons. The application was continuously recording individuals' movements throughout the day. Thus, the dataset consists of raw high-quality GPS traces, which include a range of pedestrian activities. The analysis reported in the paper focuses on activities that has been labeled as "walking activity" by the app using proprietary machine learning algorithms. Also, due to privacy limitations, we have no information about user profiles, such whether they are resident of the city, level of familiarity with the environment, etc. The average recording gap is 15s, and the positioning accuracy is within 10m. The GPS traces were segmented to individual paths based on tracking continuity - a path is considered to end at a destination if the walking activity was paused for longer than 5 minutes. To preserve privacy, the origin and destination of each trip have been randomly relocated within a 100 meter radius centered at the original location. To remove any possible bias resulting from this randomisation procedure, we trimmed the beginning and the end of each path within this range. All paths were map-matched to the open street map network available at `www.openstreetmap.org` us-

ing a Hidden-Markov-Chain algorithm [57]. Furthermore, we screened all map-matched paths using the Douglas-Peucker (DP) algorithm [58] to exclude any detours that could be caused by the GPS jitters, that were not fully removed by the trimming.

We analyzed a subset of walking paths selected from the full dataset, that met the following criteria: (1) no straight line paths, and (2) a paths was not more than 80% longer than the shortest possible path connecting the origin and destination. The 80% cutoff was taken as the 95% quantile. The first criterion removed straightforward (from a navigational perspective) paths. The second criterion was implemented to exclude multi-purpose trips, sight-seeing, and exercise. We also excluded paths having their shortest network distance smaller than 200 meters. After this pre-processing 165,645 trajectories by 4,879 pedestrians from Boston, and 189,075 trajectories by 7,372 pedestrians from San Francisco remained in the analysis, comprised of about 60% of paths from the original dataset.

The paths in Boston has a mean length of $856.0m$ ($SD = 843.6m$), while the paths in San Francisco had a mean length of $868.1m$ ($SD = 912.1m$). Detailed aggregate statistics of the human paths included in the study are reported in the Supplementary Table 1.

Street Networks We retrieved the street network of the city of Boston and San Francisco and their surrounding areas from the Open Street Map (https://www.openstreetmap.org) (see Supplementary Figure 1). All walkable street segments are included to form the walkable street network that is used consistently for all the following calculation and analysis, including but not limited to the map-matching of human path, calculation of shortest path, and random walk paths.

17

We simplified the retrieved street networks to speed up the calculation by cleaning up redundant nodes and edges around intersections. Specifically, we grouped adjacent intersections into one by applying a hierarchical clustering using the complete linkage and network distance among nodes. We selected 30m as the threshold of the diameter of the clusters through repeated experiments. This simplification eliminated unnecessary details around intersections while preserving network topology. The street network of Boston and San Francisco used in our data analysis is illustrated in Supplementary Figure 1.

Walking Paths The GPS trajectories of human paths were map-matched to the walking street network from the Open Street Map (Supplementary Section 1.1) using a Hidden Markov Chain algorithm (Newson and Krumm, 2009). The algorithm was implemented using the map matching API of Graphhoper (https://graphhopper.com/api/1/docs/map-matching/). A small number of paths that were not successfully matched were eliminated from the data-set. The map-matched human paths were projected onto the simplified network utilizing the link table between nodes in the original network and the simplified network (Supplementary Section 1.1). The summary statistics of the shortest paths in both Boston and San Francisco are shown in the Supplementary Table 1.

For each human path, the corresponding shortest distance path was calculated based on the street network from the Open Street Map (Supplementary Section 1.1) using the classic Dijkstra's algorithm implemented in the python package of igraph-python (http://igraph.org/python/). The map-matched shortest paths were projected onto the simplified network utilizing the link table between nodes in the original network and the simplified network. The summary statistics of the

shortest paths in both Boston and San Francisco are shown in the Supplementary Table 1. The trip velocity distribution is reported in Supplementary Figure 2.

The Google paths were retrieved from the Google Map Routing API (https://cloud.google.com/maps-platform/routes/) and map-matched to the original street network using the identical HMM algorithm as applied to the human paths. Due to resource limitations, we randomly extracted a subset of paths for the comparison between human and google paths in Boston and San Francisco respectively. We retrieved Google paths for 9,254 origin-destination pairs and 1,492 origin-destination pairs in San Francisco. The comparison was restricted to the human paths for which we could obtain a corresponding Google path between the given origin and destination. If multiple human paths existed for one OD pair, all human paths were compared separately to the Google path, and average similarity performances were taken for the comparison. All retrieved Google trajectories were map-matched to the original street networks before being projected to the simplified network.

To examine the probability of pedestrians following app-planned paths, we compared human paths to paths planned by the most widely used routing app - Google Map. Results show that Google paths are significantly different from human paths. First, the length distribution of Google path is significantly different from that of human paths in both cities (see Supplementary Figure 3). Metrics comparing geometries of Google paths with geometries of human paths also demonstrate a significant difference in Jaccard Similarity measured as exact overlaps, and in geometric similarity measured using Hausdorff distance (see Supplementary Tables 2 and 3).

Detection of Asymmetries To test whether the sample probability distribution of paths chosen between any specific origin-destination (OD) pair depends on the direction we need to aggregate trips across OD pairs. However, since the sample size obtained by considering the original OD pair is limited, we also consider intermediate points in a trajectory as possible OD pairs. To achieve a sufficiently large sample size we consider only OD pairs within an Euclidean separation distance between 200m and 250m, with at least 50 recorded paths, including at least 20 paths in each direction, that do not lie on a straight line path. We define a straight line path between an OD if there exists a path that simplifies to two points with the DP algorithm [58], with a cutoff of 30m. Note that this approach is conservative, since a symmetric model must necessarily be symmetric at all scales.

For individual-level asymmetry test this criterion is met by 235 OD pairs generated by 73 pedestrians in San Francisco, and by 316 OD pairs generated by 101 pedestrians in Boston. For the aggregate asymmetry analysis we also include any OD pairs that have a sufficient number of paths combined between different pedestrians, which gives 2,865 and 3,000 origin-destination pairs in San Francisco and Boston, respectively. Note that a single pedestrian path could be accounted multiple times by considering different sub-parts of the original trajectory.

Aggregate-level analysis is done as follows. For a specific OD pair, we consider the universe of all paths followed by humans in each direction, and represent them as an unordered set of street intersection points. For each intersection point, we record the occurrence of outbound $OD^{\rightarrow} = \{n_1^{\rightarrow}, \cdots, n_m^{\rightarrow}\}$, and in-bound $OD^{\leftarrow} = \{n_1^{\leftarrow}, \cdots, n_m^{\leftarrow}\}$ paths. Individual-level

analysis is done as follows. We group all paths taken by an individual between a specific OD pair $OD = \{\mathcal{P}_1, \ldots, \mathcal{P}_m\}$, to determine the universe of observed paths in each direction. We count the number of occurrences of these paths in the forward and reverse direction, defined as $OD^{\rightarrow} = \{n_1^{\rightarrow}, \ldots, n_m^{\rightarrow}\}$ and $OD^{\leftarrow} = \{n_1^{\leftarrow}, \ldots, n_m^{\leftarrow}\}$, and establish an analogy with the process of extracting with replacement marbles from an urn containing m different types of marbles – one for each possible path. More specifically, the urn contains $n_i^{\rightarrow} + n_i^{\leftarrow}$ marbles of type i, for each i. Given this analogy, the observed group of paths in one direction, say, $OD^{\rightarrow}$, can be seen as an instance of a random extraction of $q^{\rightarrow} = \sum_i n_i^{\rightarrow}$ marbles from the urn. If paths were to be symmetric – null hypothesis, the extraction would obey a multinomial distribution. Thus, we can apply standard statistical hypothesis test to the null hypothesis that the observations from sets $OD^{\rightarrow}$ and $OD^{\leftarrow}$ follows such distribution. Results of the test, reported in Figure 2, show that the null hypothesis can be rejected with a 0.05 significance threshold for 32% and 24% of the origin-destination pairs for Boston and San Francisco respectively, and that the cumulative distribution of the resulting p-values is considerably skewed towards lower values than those expected under the null hypothesis. Thus, we find a statistically significant asymmetry in individual human paths.

The proposed statistical test, which applies to both individual- and aggregate-level analysis, builds on the analogy with the process of independently extracting m marbles from an urn. If the null hypothesis that paths are symmetric is true, there exists a set of probabilities $\{p^1, \cdots, p^m\}$ with $\sum_{i=1}^{m} p^i = 1$, associated to each path that are direction independent. In such a case, the best estimates for such probabilities can be obtained as

$$p^i = \frac{n_i^{\leftarrow} + n_i^{\rightarrow}}{q^{\leftarrow} + q^{\rightarrow}}, \tag{3}$$

with $q^{\leftarrow} = \sum_{i=1}^{m} n_i^{\leftarrow}$ and $q^{\rightarrow} = \sum_{i=1}^{m} n_i^{\rightarrow}$.

To assess the validity of the null hypothesis, we test whether the sample occurrences $OD^{\rightarrow}$ and $OD^{\leftarrow}$ are statistically compatible with the process of repeatedly and independently extracting $q^{\rightarrow} + q^{\leftarrow}$ elements from a urn with m marbles (with replacement). The resulting distribution is multinomial and defined by the probabilities p_i. The exact p-value for the test can not be obtained analytically; however, the Likelihood-Ratio test [59] provides an asymptotic distribution in the case the null hypothesis holds. Specifically, the maximum likelihood estimate is

$$\log \mathcal{L}_M = \sum_{i=1}^{m} n_i^{\leftarrow} \log \left(\frac{n_i^{\leftarrow}}{q^{\leftarrow}} \right) + \sum_{i=1}^{m} n_i^{\rightarrow} \log \left(\frac{n_i^{\rightarrow}}{q^{\rightarrow}} \right) \tag{4}$$

and the alternative model is

$$\log \mathcal{L}_A = \sum_{i=1}^{m} n_i^{\leftarrow} \log(p_i) + \sum_{i=1}^{m} n_i^{\rightarrow} \log(p_i). \tag{5}$$

Finally, according to Wilks' theorem [59], the Likelihood-Ratio statistic should converge to a chi square with $m - 1$ degrees of freedom

$$2(\log \mathcal{L}_M - \log \mathcal{L}_A) \xrightarrow{d} \chi^2(m - 1). \tag{6}$$

The results of the test are reported in the main text. Since the underlying distribution on which performing the statistical test is unknown, we proceed with a more conservative test based on a null sampled distribution. This implies that the real null expectation should be slightly below the first bisector, as depicted by the shortest distance p-value distribution in Figure 2-B.

Asymmetry Proofs To prove that the stochastic distance minimization model cannot explain asymmetry, consider the (random) cost $\mathcal{C}_{dist}(OD^{\rightarrow})$ of a path between a certain origin-destination

pair OD, and the cost $\mathcal{C}_{dist}(OD^{\leftarrow})$ of the reverse path. Since both costs are obtained as sum of independent random variables, and the random variables considered in the summation are the same (except for their order), both $\mathcal{C}_{dist}(OD^{\rightarrow})$ and $\mathcal{C}_{dist}(OD^{\leftarrow})$ have the same probability distribution, which contradicts our empirical observations of asymmetric human paths.

To prove that the vector-based navigation model can explain asymmetry, notice that, for a given path between O and D the cost of the path in the forward and reverse direction, $C_{dir}(OD^{\rightarrow})$ and $C_{dir}(OD^{\leftarrow})$, are obtained from the summation of *different* random variables, since the value $\theta_i^{\rightarrow}$ for street segment S_i in the forward direction is different from that of $\theta_i^{\leftarrow}$ in the reverse direction – see Figure 3. Thus, the probability distribution of $C_{dir}(OD^{\rightarrow})$ is different from that of $C_{dir}(OD^{\leftarrow})$, which implies that the stochastic distance minimization model can support the hypothesis of asymmetric human navigation.

Model Fitting and Comparison We compare the explanatory power of the two models by performing a set of 1000 simulations in Boston and San Francisco to optimally tune the error parameter σ for both the stochastic distance minimization and the vector-based model – details about the exploration ranges for σ can be found in the Supplementary Section 3. Given a human path $OD^{(h)\rightarrow}$ from O to D, we run both models on that origin destination pair for 1000 simulations, and recorded the number of times each of the two models selected exactly path $OD^{(h)\rightarrow}$. To determine which of the two models performed statistically better in this task, we compared their respective

23

likelihood values. The likelihood of a model is obtained from

$$\log \mathcal{L}^{(x)} = \sum_{i=1}^{N} \log P(OD^{(h)\rightarrow}|OD_i, C_{(x)}, \sigma_x),$$

that is, the sum over all N paths of the logarithm of the probability

$$P(OD^{(h)\rightarrow}|OD_i, C_{(x)}, \sigma_x)$$

of selecting the human path $OD^{(h)\rightarrow}$ for origin-destination pair OD_i, computed using the cost function of the specific model $x = \{dist, dir\}$ with the optimal value σ_x for model x of the error parameter.

We aggregated paths into bins by OD separation, in steps of 50m. For each bin we calibrated σ and tested model performances with a leave-one-out cross-validation [60]. So, we found the value of σ that maximises the likelihood of $N_s - 1$ paths associated with the bin s, and measured the likelihood of the out-of-sample path. We repeated this procedure by leaving out another path until we covered the whole set of paths in each bin. We also included a cutoff parameter c to account for paths with zero sample-probability, which would cause divergence of the likelihood metric. Such parameter is set to $c = 1/N_s = 0.001$ which is the expected minimal detectable probability with $N_s = 1000$ simulations. A detailed exploration of the dependency of c is reported in the Supplementary Section 3, however we did not observe substantial change of our results with different values of c.

Data Availability Statement

Due to privacy constrains policies and signed Data Usage Agreement, we are not allowed to share the full GPS tracks considered in this work. For this reason, we generated a small sample of 100 trajectories for Boston. We also make available the pre-processed pedestrian street networks for Boston and San Francisco. The sample data set and street network data can be accessed at Zenodo [61].

Source data for figures 1, 2, and 4 is also available with this manuscript. Figure 1c, 2A, and 3 (left) used basemap from Open Street Map (https://www.openstreetmap.org) under the Open Database License (https://www.openstreetmap.org/copyright). Figure 3 (right) uses Map data ©2021 Google under the fair-use guidelines (https://about.google/brand-resource-center/products-and-services/geo-guidelines/♯general-guidelines-copyright-fair-use).

Code Availability Statement

The version of PedNav package used in this study and a guide to reproducing the results is available through GitHub under the GNU GPL-3.0 License (https://github.com/cbongiorno/pednav). The specific version of the package used to generate the results in the current study can be found at Zenodo [61].

A pseudo-code description of the algorithms used for human navigation based on stochastic distance minimization and vector navigation is reported in the Supplementary Section 4.

4 References

1. Newell, A., Shaw, J. C. & Simon, H. A. Elements of a theory of human problem solving. *Psychological review* **65**, 151 (1958).

2. Zhu, S. & Levinson, D. Do people use the shortest path? an empirical test of wardrop's first principle. *PloS one* **10** (2015).

3. Lima, A., Stanojevic, R., Papagiannaki, D., Rodriguez, P. & González, M. C. Understanding individual routing behaviour. *Journal of The Royal Society Interface* **13**, 20160021 (2016).

4. Javadi, A.-H. *et al.* Hippocampal and prefrontal processing of network topology to simulate the future. *Nature communications* **8**, 1–11 (2017).

5. Griffiths, T. L., Lieder, F. & Goodman, N. D. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science* **7**, 217–229 (2015).

6. Huys, Q. J. *et al.* Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences* **112**, 3098–3103 (2015).

7. Gershman, S. J., Horvitz, E. J. & Tenenbaum, J. B. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science* **349**, 273–278 (2015).

8. Baker, C. L., Jara-Ettinger, J., Saxe, R. & Tenenbaum, J. B. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour* **1**, 1–10 (2017).

9. Liu, S., Ullman, T. D., Tenenbaum, J. B. & Spelke, E. S. Ten-month-old infants infer the value of goals from the costs of actions. *Science* **358**, 1038–1041 (2017).

10. Gershman, S. J. Origin of perseveration in the trade-off between reward and complexity. *bioRxiv* (2020).

11. Hillier, B. & Iida, S. Network and psychological effects in urban movement. In *International Conference on Spatial Information Theory*, 475–490 (Springer, 2005).

12. Brockmann, D., Hufnagel, L. & Geisel, T. The scaling laws of human travel. *Nature* **439**, 462–465 (2006).

13. Gonzalez, M. C., Hidalgo, C. A. & Barabasi, A.-L. Understanding individual human mobility patterns. *nature* **453**, 779–782 (2008).

14. Simini, F., González, M. C., Maritan, A. & Barabási, A.-L. A universal model for mobility and migration patterns. *Nature* **484**, 96–100 (2012).

15. Alessandretti, L., Sapiezynski, P., Sekara, V., Lehmann, S. & Baronchelli, A. Evidence for a conserved quantity in human mobility. *Nature Human Behaviour* **2**, 485–491 (2018).

16. Hamedmoghadam, H., Ramezani, M. & Saberi, M. Revealing latent characteristics of mobility networks with coarse-graining. *Scientific reports* **9**, 1–10 (2019).

17. Kraemer, M. U. *et al.* Mapping global variation in human mobility. *Nature Human Behaviour* 1–11 (2020).

18. Verbavatz, V. & Barthelemy, M. The growth equation of cities. *Nature* **587**, 397–401 (2020).

19. Alessandretti, L., Aslak, U. & Lehmann, S. The scales of human mobility. *Nature* **587**, 402–407 (2020).

20. Er-Jian, L. & Xiao-Yong, Y. A universal opportunity model for human mobility. *Scientific Reports (Nature Publisher Group)* **10** (2020).

21. Gallotti, R., Bazzani, A., Rambaldi, S. & Barthelemy, M. A stochastic model of randomly accelerated walkers for human mobility. *Nature communications* **7**, 1–7 (2016).

22. Gillner, S. & Mallot, H. A. Navigation and acquisition of spatial knowledge in a virtual maze. *Journal of cognitive neuroscience* **10**, 445–463 (1998).

23. Foo, P., Warren, W. H., Duchon, A. & Tarr, M. J. Do humans integrate routes into a cognitive map? map-versus landmark-based navigation of novel shortcuts. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **31**, 195 (2005).

24. Norman, J. F., Crabtree, C. E., Clayton, A. M. & Norman, H. F. The perception of distances and spatial relationships in natural outdoor environments. *Perception* **34**, 1315–1324 (2005).

25. Sun, Y. & Wang, H. Perception of space by multiple intrinsic frames of reference. *PloS one* **5** (2010).

26. Weisberg, S. M. & Newcombe, N. S. How do (some) people make a cognitive map? routes, places, and working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **42**, 768 (2016).

27. Vuong, J., Fitzgibbon, A. W. & Glennerster, A. No single, stable 3d representation can explain pointing biases in a spatial updating task. *Scientific reports* **9**, 1–13 (2019).

28. Bécu, M. *et al.* Age-related preference for geometric spatial cues during real-world navigation. *Nature Human Behaviour* **4**, 88–99 (2020).

29. van der Ham, I. J., Claessen, M. H., Evers, A. W. & van der Kuil, M. N. Large-scale assessment of human navigation ability across the lifespan. *Scientific Reports* **10**, 1–12 (2020).

30. Marshall, J. M. *et al.* Mathematical models of human mobility of relevance to malaria transmission in africa. *Scientific reports* **8**, 1–12 (2018).

31. Yan, X.-Y., Wang, W.-X., Gao, Z.-Y. & Lai, Y.-C. Universal model of individual and population mobility on diverse spatial scales. *Nature communications* **8**, 1–9 (2017).

32. Yan, X.-Y. & Zhou, T. Destination choice game: A spatial interaction theory on human mobility. *Scientific reports* **9**, 1–9 (2019).

33. Coutrot, A. *et al.* Virtual navigation tested on a mobile app is predictive of real-world wayfinding navigation performance. *PloS one* **14**, e0213272 (2019).

34. Manley, E., Addison, J. & Cheng, T. Shortest path or anchor-based route choice: a large-scale empirical analysis of minicab routing in london. *Journal of transport geography* **43**, 123–139 (2015).

35. Malleson, N. *et al.* The characteristics of asymmetric pedestrian behavior: A preliminary study using passive smartphone location data. *Transactions in GIS* **22**, 616 (2018).

36. Dijkstra, E. A note on two problems in connexion with graphs. *Numerische Mathematik* **1**, 269 (1959).

37. Fechner, G. T. *Elements of Psychophysics* (Howes, D H; Boring, E G (eds.), 1860).

38. Newcombe, N., Huttenlocher, J., Sandberg, E., Lie, E. & Johnson, S. What do misestimations and asymmetries in spatial judgement indicate about spatial representation? *Journal of Experimental Psychology: Learning, Memory, and Cognition* **25**, 986 (1999).

39. Bailenson, J. N., Shum, M. S. & Uttal, D. H. Road climbing: Principles governing asymmetric route choices on maps. *Journal of Environmental Psychology* **18**, 251–264 (1998).

40. Bailenson, J. N., Shum, M. S. & Uttal, D. H. The initial segment strategy: A heuristic for route selection. *Memory & Cognition* **28**, 306–318 (2000).

41. Christenfeld, N. Choices from identical options. *Psychological Science* **6**, 50–55 (1995).

42. Howard, L. R. *et al.* The hippocampus and entorhinal cortex encode the path and euclidean distances to goals during navigation. *Current Biology* **24**, 1331–1340 (2014).

43. Marchette, S. A., Vass, L. K., Ryan, J. & Epstein, R. A. Anchoring the neural compass: coding of local spatial reference frames in human medial parietal lobe. *Nature neuroscience* **17**, 1598 (2014).

44. Collett, T. S. & Graham, P. Animal navigation: path integration, visual landmarks and cognitive maps. *Current Biology* **14**, R475–R477 (2004).

45. Hafting, T., Fyhn, M., Molden, S., Moser, M.-B. & Moser, E. I. Microstructure of a spatial map in the entorhinal cortex. *Nature* **436**, 801–806 (2005).

46. de Cothi, W. & Spiers, H. J. Spatial cognition: Goal-vector cells in the bat hippocampus. *Current Biology* **27**, R239–R241 (2017).

47. Toledo, S. *et al.* Cognitive map–based navigation in wild bats revealed by a new high-throughput tracking system. *Science* **369**, 188–193 (2020).

48. Poucet, B., Thinus-Blanc, C. & Chapuis, N. Route planning in cats, in relation to the visibility of the goal. *Animal Behaviour* **31**, 594–599 (1983).

49. Epstein, R. A., Patai, E. Z., Julian, J. B. & Spiers, H. J. The cognitive map in humans: spatial navigation and beyond. *Nature neuroscience* **20**, 1504 (2017).

50. Poulter, S., Lee, S. A., Dachtler, J., Wills, T. J. & Lever, C. Vector trace cells in the subiculum of the hippocampal formation. *Nature Neuroscience* 1–10 (2020).

51. Tolman, E. C. Cognitive maps in rats and men. *Psychological review* **55**, 189 (1948).

52. Fu, E., Bravo, M. & Roskos, B. Single-destination navigation in a multiple-destination environment: a new "later-destination attractor" bias in route choice. *Memory & cognition* **43**, 1043–1055 (2015).

53. Brunyé, T. T. *et al.* Planning routes around the world: International evidence for southern route preferences. *Journal of Environmental Psychology* **32**, 297–304 (2012).

54. Wilson, E. B. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22**, 209–212 (1927).

55. Peer, M., Brunec, I. K., Newcombe, N. S. & Epstein, R. A. Structuring knowledge with cognitive maps and cognitive graphs. *Trends in Cognitive Sciences* (2020).

56. Stern, E. & Leiser, D. Levels of spatial knowledge and urban travel modeling. *Geographical Analysis* **20**, 140–155 (1988).

57. Newson, P. & Krumm, J. Hidden markov map matching through noise and sparseness. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, 336–343 (Association for Computing Machinery, New York, NY, USA, 2009). URL `https://doi.org/10.1145/1653771.1653818`.

58. Douglas, D. H. & Peucker, T. K. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: the international journal for geographic information and geovisualization* **10**, 112–122 (1973).

59. Wilks, S. S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics* **9**, 60–62 (1938).

60. Zhang, P. Model selection via multifold cross validation. *The annals of statistics* 299–313 (1993).

61. Buongiorno, C. *et al.* Pednav (1.1). *Zenodo. https://doi.org/10.5281/zenodo.5187718* (2021).

Author Contributions   A.R. and P.S. conceived and supervised the research. C.B. and Y.Z. led and performed the exploratory data analysis. C.B. designed the asymmetry testing procedure, the stochastic modeling, the validation method and performed the simulations. Y.Z. processed the data, conducted statistical analyses, and drafted the paper. M.K. and D.T. carried out exploratory modeling converging on the presented model; M.K. drafted the introduction, content related to cognitive science and neuroscience, cognitive science research and methodology; C.R. contributed to conceptualise and design the research; J.T. provided conceptualization and feedback; all authors contributed to writing and revising the paper.

Competing Interests   The authors declare that they have no competing financial interests.

Correspondence   Correspondence and requests for materials should be addressed to Paolo Santi (email: psanti@mit.edu).