Multivariate Behavioral Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hmbr20

Binding the Person-Specific Approach to Modern AI in the Human Screenome Project: Moving past Generalizability to Transferability

Nilam Ram, Nick Haber, Thomas N. Robinson & Byron Reeves

To cite this article: Nilam Ram, Nick Haber, Thomas N. Robinson & Byron Reeves (2023): Binding the Person-Specific Approach to Modern AI in the Human Screenome Project: Moving past Generalizability to Transferability, Multivariate Behavioral Research, DOI: 10.1080/00273171.2023.2229305

To link to this article: https://doi.org/10.1080/00273171.2023.2229305

Published online: 13 Jul 2023.

Submit your article to this journal

Article views: 50

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=hmbr20

MULTIVARIATE BEHAVIORAL RESEARCH
https://doi.org/10.1080/00273171.2023.2229305

Routledge
Taylor & Francis Group

Binding the Person-Specific Approach to Modern AI in the Human Screenome Project: Moving past Generalizability to Transferability

Nilam Ram[a], Nick Haber[b], Thomas N. Robinson[c], and Byron Reeves[d]

[a]Department of Psychology & Department of Communication, Stanford University; [b]Graduate School of Education, Stanford University; [c]Department of Pediatrics, Stanford University; [d]Department of Communication, Stanford University

ABSTRACT

Advances in ability to comprehensively record individuals' digital lives and in AI modeling of those data facilitate new possibilities for describing, predicting, and generating a wide variety of behavioral processes. In this paper, we consider these advances from a person-specific perspective, including whether the pervasive concerns about *generalizability* of results might be productively reframed with respect to *transferability* of models, and how self-supervision and new deep neural network architectures that facilitate transfer learning can be applied in a person-specific way to the super-intensive longitudinal data arriving in the Human Screenome Project. In developing the possibilities, we suggest Molenaar add a statement to the person-specific Manifesto – "In short, the concerns about *generalizability* commonly leveled at the person-specific paradigm are unfounded and can be fully and completely replaced with discussion and demonstrations of *transferability*."

KEYWORDS
Transfer learning; experience sampling; ecological momentary assessment; foundation models

When NR arrived at Penn State, Peter Molenaar was excitedly pushing forward from his recently published manifesto on the need for scientific study of the individual and the structure of their *intraindividual variability* (Molenaar, 2004). During that period, Peter avoided and sometimes flatly refused to discuss the topic of between-person differences. Instead, he deeply explored how theory and methods are constructed when working with $N = 1$ cases—*one* person, *one* historical record, *one* earth (e.g., as is done in study of history and geology). For the rest of us in the group, psychologists trained on literature based in study of interindividual differences, engagement with the $N = 1$ Gedankenexperiment was both tortuous and truly inspiring!

Person-specific time-series modeling

While we did eventually figure out how to generate theoretical explanations of psychological phenomena without reference to second or third individuals, the theoretical work remains quite difficult! In contrast, parallel efforts to generate and codify a variety of $N = 1$ data analysis pipelines for P-technique, dynamic factor analysis, broader classes of state-space models, and behavioral landscapes flowed more quickly. These person-specific methods provided interesting and useful *description* of how an individual's or a dyad's behavior change over time, *prediction* of time-varying outputs, and *generation* of person-specific dynamic control strategies (e.g., Molenaar et al., 2009; Molenaar & Ram, 2010; Ram et al., 2013). The $N = 1$ modeling work clearly demonstrated the important opportunity that emerges from the person-specific paradigm—"each person is a possibly unique system of interacting dynamic processes, the unfolding of which give rise to an individual life trajectory in a high-dimensional psychological space" (Molenaar, 2004, p. 202).

After a while, Peter tolerated conversation about between-person differences, and eventually he actually integrated between-person differences into the analysis pipelines. Pure $N = 1$ extensions of state-space models in the unified structural equation modeling framework (uSEM; Gates et al., 2010) were expanded into a $N =$ many, group iterative multiple model estimation (GIMME; Gates & Molenaar, 2012) approach where the best fitting model of multiple individuals' time-series data was obtained using an iterative model building algorithm. The "bottom-up" search procedures embedded into GIMME immediately showed promise for modeling the abundance of replicated time-series

CONTACT Nilam Ram ✉ nilamram@stanford.edu Department of Psychology & Department of Communication, Stanford University, 300-A Building 120, 450 Jane Stanford Way, Stanford, CA 94305, USA.

Copyright © 2023 Society of Multivariate Experimental Psychology

data generated in fMRI studies. Those procedures, when accompanied by network graph representations of the underlying uSEMs, now generously inform theoretical explanations of how neural activations and cognitive/emotional/interpersonal behaviors unfold in multi-dimensional space. The wider community of modelers are now, as indicated by the wealth of work in this special issue and elsewhere, systematically filling in a variety of "top-down" versions of these models (e.g., multilevel and mixture state-space models, Fischer et al., this issue; Hunter et al., this issue; Oravecz & Vandekerckhove, this issue).

Person-specific machine learning

Parallel to the $N = 1$ to N = many modeling extensions, we also demonstrated how traditional machine learning (ML) models could be used, in accordance with the person-specific paradigm, to model the intensive longitudinal data arriving from smartphone-based experience sampling studies (Tuarob et al., 2017). Engaging a pure $N = 1$ approach, the relatively long ($T = 425+$) multivariate time-series data from each of 150 participants in an experience sampling study (wherein individuals completed short questionnaires 7+ times per day for up to 9 wk, Ram et al., 2014) were modeled separately using a variety of supervised learning models. Each individual's data were split into training/test data and modeled using regression (e.g., Vector Auto-Regression, VAR), function based (e.g., support vector machines, multi-layer preceptrons), tree based (M5), and lazy learning based (K-nearest neighbor, locally weighted learning) methods. Generally, the two best performing prediction models were person-specific random forests and person-specific radial basis function neural networks (with Gaussian kernels), each of which has known advantages for obtaining interpretable explanations of the data (e.g., *via* probing of feature importance; see e.g., Brick et al., 2017) or adjusting online dynamic control (see e.g., Yu et al., 2011). The $N = 1$ deployments highlighted how person-specific ML paradigms can support clinicians' efforts to obtain personalized prediction models and deliver precision interventions.

Although we demonstrated that person-specific ML methods could locate hints of signal in what many consider to be extremely noisy data, two things were obvious. First, the volume, velocity, veracity of experience sampling and ecological momentary assessment paradigms would never support entry into the "big data" space. Any real attempts to combine the person-specific paradigm with the exciting innovations in ML and AI would require a lot more data than any person would ever provide *via* the intrusive experience sampling methods. We needed a different way to get data. Second, the perennial idiographic vs. nomothetic debate could not be avoided by switching to ML and AI. The same familiar set of conceptual and practical questions emerges even when we prioritize prediction (over explanation) and even when there are lots of data. *Why on earth would we only ingest $N = 1$ data when we can ingest the $N =$ many data? And, seriously, what about generalization?*

In the remainder of this paper, we attempt to answer these questions while outlining an agenda and infrastructure for deploying person-specific AI on the big data collected in the Human Screenome Project (HSP; Reeves et al., 2020). While surfacing the possibilities, we point toward a possible resolution of "the generalizability problem"—or at least a re-considered position and a promising way forward.

The generalizability problem

A pervasive concern with the person-specific paradigm is that the models and findings obtained through analysis of $N = 1$ data will not *generalize*—that is, a person-specific model will not provide accurate representations or predictions when applied to other data (other persons, other variables, other stimuli, or other occasions). Some researchers think this deems the person-specific paradigm useless—because the person-specific distribution is viewed as a localized, non-representative subset of a larger population distribution. Findings derived from an individual's idiosyncratic distribution will, at best, provide a biased representation of the population model, and at worst will be so far *out-of-distribution* that knowledge of the behavioral processes for any one person is counterproductive for science (e.g., as in the practice of removing outliers). Other researchers view this "lack of generalizability" as the very reason the person-specific paradigm is useful. When individuals are unique dynamic systems, the patterns of behavior they produce will be idiosyncratic. Other individuals' data and behavior are, by definition, *out-of-distribution*. Each person is unique. When the model for Person A is used to deploy interventions for Person B, we likely did them a disservice. For example, when the individual development plan (IDP) organizing the average child's education is used without edit to direct a specific child's education, that child's development was almost certainly compromised. Every child you know is not the average child. They are unique and deserve better! Our recent engagement with AI *foundation models* (i.e, large language models; LLMs) suggests that there may be a way around "the generalizability problem"—to a world where diversity and uniqueness are celebrated and acknowledged.

Modern artificial intelligence and foundation models

Deep neural network models, as a class of models used in modern AI, extend from familiar models. Similar to many structural equation models, inputs (predictor variables **X**) are connected to outputs (outcome variables **Y**) through a series of "hidden" layers that consist of relations among latent variables that capture specific features of the multivariate input vectors and provide optimal prediction of the output vector. For instance, raw images (pixel arrays) of animals might be connected to ground-truth content labels (e.g., cat, not-cat) through linear and nonlinear relations among latent variables that capture visual features like the pointiness of ears and the roundness of eyes that were learned directly from the data. Notably, when deep neural network models were being developed, the content and interpretability of the latent variables was purposively deprioritized in favor of prediction accuracy. The modeling enterprise is singularly focused on doing prediction really well, without any need to explain how or why the inputs are connected to the outputs. This approach sits in stark contrast to how latent variable models are typically used in the social sciences—where the models are considered and evaluated as diagrammatic representations of the hypothesized causal processes that produced the observed data. The need for theoretical explanation keeps the models relatively small and simple, and almost exclusively comprised of linear relations. In contrast, because deep neural network models are unburdened from the need to provide parsimonious and theoretically-meaningful explanation, these models can capture complex nonlinear mappings of the inputs and outputs using millions, or in some cases, many billions (Brown et al., 2020) or trillions of parameters.

In the last decade, advances in computational hardware (e.g., graphical processing units, GPUs, capable of doing matrix operations quickly) supported experiments with deep neural network model architectures that simply abandoned all notions of parsimony. Massive model expansions obtained when motivated modelers got access to bigger computational resources led to the discovery that models with billions of parameters are tractable and actually perform better than the asymptotic limits on performance implied by traditional statistical theory about model parsimony-accuracy tradeoffs (Belkin et al., 2019). In the last few years, availability of large-scale training data (e.g., everything on the internet) and access to the super-massive computational resources needed to bind the models to the data, has spurred progression of deep neural network architectures. Particularly exciting is discovery of the *transformer* architecture (Vaswani et al., 2017)—where initially obtained embeddings (i.e., latent representations) of input sequences (e.g., words) are supplemented with self-attention vectors that aggregate information from all of the surrounding data (surrounding words and sentences) to generate enriched representations that are informed by context.

The transformer architecture's ability to capture long-range dependencies (e.g., word meanings that follow from prior sentences), and the possibility to train the models on broad data using *self-supervision* (supervision where the output variables "come for free" with the data, as opposed to supervision where production of the output variables requires intensive human labeling, e.g. image classification) has prompted a full-on paradigm shift into what are now being referred to as *foundation models and large language models* (Bommasani et al., 2021). For example, self-supervised learning in the language domain—where models are trained to "fill in" missing words from sentences where one or more words have been masked (i.e., to predict text from previous, next, or surrounding text)—has facilitated rich encoding of the English language (and many other languages too). Foundation models such as Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2018) and Generative Pre-trained Transformer 3 (GPT-3; Brown et al., 2020) that were trained on minimally-curated data (e.g. the massive amounts of text that can be found on the internet) now facilitate AI systems that are able to summarize text, make predictions from text, and generate text in real-time conversation. Similar advances follow from models trained on hundreds of millions of image-text pairs collected from the Internet. For example, Contrastive Language-Image Pre-training (CLIP; Radford et al., 2021) maps images and text into a high-dimensional embedding space, providing locations of specific texts and images that facilitate quick retrieval of images *via* text query; and DALL-E 2 (Ramesh et al., 2022) can generate new photo-realistic images from user-provided text prompts (e.g., "photo of an avocado chair"). Following these innovations, research domains with an abundance of large-scale training data and access to massive computation resources are leveraging transformer model architectures and self-supervision to make speedy progress on problems that were previously considered decades away or impossible (Economist, 2022).

Moving past generalizability to transferability

Foundation models are, by definition, models that are trained on massive amounts of unlabeled data (using self-supervised learning) and can be easily adapted to perform a wide range of downstream tasks they were not explicitly trained for (Bommasani et al., 2021). These models are replacing task-specific models by pre-training an "over-parameterized" general-purpose model architecture on a broad array of data to obtain a rich latent embedding of the input data. The big excitement for AI stems from unexpected speed and range of transfer to new tasks—specifically, the possibility for "zero-shot or few-shot learning" where state-of-the art performance on new tasks is available immediately or is obtained quickly with relatively little training data. For example, the pre-trained foundation model in IBM's Watson NLP learns to score emotional sentiment of sentences in new languages using only a few thousand sentences of training data (orders of magnitude less data than was needed with previous models). In contrast to the psychological literature, where discussions of *generalization* are primarily concerned with whether results obtained from a specific sample of persons or stimuli will generalize to a population of persons or stimuli (e.g., Judd et al., 2012; Yarkoni, 2020 and associated commentaries), discussions in the AI literature are currently focused on how foundation models support speedy *transfer* to a broad range of tasks.

In brief, *transfer learning* occurs when a model that was trained on one task or type of data is then used to accomplish another task or analyze a new type of data. Consider a scenario where we (pre-) train a model **M0** to obtain a rich (latent) representation of dataset **A**, and use that representation (i.e., high-dimensional embedding) to describe, predict, and generate data in accordance with the distribution of **A**. We then engage *transfer* by fine-tuning specific parts of model **M0** to accommodate additional embeddings needed to represent dataset **B** and use the resulting model **M1** for description, prediction, and generation tasks with distribution **B**. Again, we can engage *transfer* by fine-tuning specific parts of model **M1** (or **M0**) to accommodate additional embeddings needed to represent dataset **C**. All along the way we keep track of how much fine-tuning is needed—measures of *transferability* include the number of parameters that were updated, how much they changed, how many new instances of data or update iterations were needed to achieve good fit. When there are no additional data and/or no fine-tuning is needed, we have *zero-shot generalization* (Palatucci et al., 2009). This is equivalent to the classic notion of generalization used in psychology, where it is expected that Model **M0** can be used, *as is*, for datasets **B** and **C,** without ever even looking at those data. When some fine-tuning is needed but it is relatively quick, we have *one-shot* or *few-shot learning*. Models **M1** and **M2** can be trained to accuracy by drawing just a few instances from datasets **B** and **C**. In the *many-shot learning* case (i.e., classic transfer learning), more than a few instances from datasets **B** and **C** are needed for the fine-tuning process. Consideration of these three types of transfer learning informs our new perspective on how person-specific analyses should be structured.

Acknowledging there is not yet good theory about how transferability of foundation models is actually structured, let us assume that transferability is related to both how far out-of-distribution **B** and **C** are relative to **A** and how useful the learned representations of **M0** are for **M1** and **M2**. When distributions **B** and **C** are redundant with distribution **A**, the transfer learning task is wholly uninteresting (although it may be somewhat interesting to consider how redundancy expectations may contribute to the replication crisis in social science). Are there data collection or validity threat scenarios (*a la* Campbell & Stanley) where new data **B** and **C** are truly expected to be redundant with data **A**?

When distributions **B** and **C** are not redundant with distribution **A**, zero-shot and few-shot learning might emerge when the representations of **M0** provide the possibility to "impute" or "interpolate" relations that exist in gaps/holes that are inside or are just outside the boundaries of distribution **A**. Outside a few simplistic scenarios, more expressive and flexible models generally fill in the gaps better than parsimonious and rigid models. Indeed, much of the current work on zero- and few-shot learning explores how to ensure that the embeddings obtained by **M0** sustain transfer beyond dataset **A**. We speculate that when the latent representations are especially rich, **M0** might be able to stretch the distance to out-of-distribution datasets **B** and **C** to achieve zero-shot generalization, or at least support fast and easy fine-tuning of **M1** and **M2** with few-shot learning. As the distances to datasets **B** and **C** increase, fine-tuning requires more iterations and data, and eventually turns into a major re-engineering that might even include changing the architecture of the model. At the extreme, many-shot transfer learning extends to the case where **M2** is simply trained directly from all the data **A** + **B** + **C** (or *via* cross-validation procedures that iterate through random subsets of the combined data).

The implications of our short consideration of transferability are twofold. First, the concerns about *generalizability* often leveled at $N = 1$ person-specific analyses are limited in scope. As noted above, the classical notion of generalizability is simply a special case of a larger concern about *transferability* of models. Pushing beyond the zero-shot learning case, the on-going discussions and work on transferability provide a cohesive framework for organizing modeling goals. We boldly suggest that, even beyond our interest in transfer and fine-tuning of person-specific models, the field might progress faster when replication and reproducibility are reconsidered in terms of transferability.

Second, zero-shot generalization is rare, even for humans. When it does emerge, for example in humans or AI that can identify category instances (e.g., pictures of cats) even when they have never been exposed to that category, the new category is usually in a nearby domain (e.g., the training data included pictures of other mammals). Molenaar's expositions on ergodicity, the third source of developmental differences, and dynamic systems suggest extreme heterogeneity in both the content and structure of individuals' behavior (Molenaar, 2004; Molenaar et al., 1993). That heterogeneity might be accommodated using transfer and fine-tuning. Molenaar and colleagues' GIMME approach, where a relatively flexible and expressive modeling architecture—uSEM—is trained using all the replicated multivariate time-series collected in an fMRI or experience sampling study, is already moving in this direction. Without going into the technical details, iterative model fitting procedures are used to obtain a "group model" that applies to everyone's data *and* "person-specific variants" that accommodate idiosyncratic features of any given individual's data. In transferability terms, the group model might be considered the general-purpose model that is transferred to and fine-tuned for each individual.

When better *transferability* is an explicit modeling goal, a simple next step in the existing modeling approaches (e.g., GIMME) might be to check whether all the data are actually needed when fine-tuning the person-specific models, and figure out how to reduce the amount of individual data needed for fine-tuning. Our initial hunch extends from the work on transportability of models and data fusion; that is, the piecing together of multiple datasets collected under heterogeneous conditions (e.g., different individuals, tasks, occasions) in the structured causal model framework (Bareinboim & Pearl, 2016; Pearl & Bareinboim, 2014). We propose that creative re-ordering and titration of how much of each participant's data are used during estimation and absolution from the parsimony-based model selection criteria used in the iterative model fitting will produce a more transferable model. Paralleling that work, our goal is to leverage the flexible and expressive modeling architectures of AI foundation models to obtain models that support description, prediction, and generation of human behavior. We forward the Human Screenome Project as an exemplar forum for exploring these possibilities.

Human Screenome Project and AI foundation models

Inspired by Molenaar's (2004) call for $N = 1$ study of individuals as unique complex systems, we use the possibility to passively collect rich behavioral data from smartphones into a *big data* repository that we think has enough volume, velocity, and veracity for a merging of the person-specific paradigm with the exciting innovations emerging from AI foundation models. Specifically, the Human Screenome Project passively records and analyzes *everything* study participants see and do on their screens. Screenshots of individuals' smartphones and laptops are obtained continuously (at 5 s intervals) as they go about their daily lives (see Reeves et al., 2021 for general overview). The 400+ US adults and adolescents participating in the most recently completed and on-going studies have and are providing recordings of their *in situ* smartphone behavior for up to a full year. These long and "super-intensive" longitudinal data track with great precision, alongside fortnightly self-reports of mental/physical states, the full spectrum of human behaviors that each individual engages as they interact with an unbounded (and completely idiosyncratic) digital environment.

Models for description

Similar to how a larger number of manifest variables are projected onto a smaller number of latent factors within structural equation models, foundation models (i.e., LLMs) have leveraged self-supervised learning on massive amounts of data (e.g., all the words and pictures on the internet) to obtain richly descriptive high-dimensional embedding spaces. For example, the currently quintessential RoBERTa model projects English language text into a 768-dimensional space (Liu et al., 2019). Similarly, the CLIP model projects multimodal data (text and pictures) into a 1024-dimensional embedding space (Radford et al., 2021). Following from our discussion of transferability, these models can be applied to the HSP data in a variety of ways.

When using *zero-shot generalization*, the foundation models are used in an off-the-shelf manner to map each screenshot into the high-dimensional space that was learned by the model. As is done when calculating principal component scores on new data, the parameters of the model are used to calculate vectors that indicate where each screenshot is located in relation to the many "sign-posts" attached to the textual and graphical content that were ingested during model training. With CLIP, for instance, text queries such as "photo of a cat," "picture of mountains," or "expression of joy" return screenshots that are located near any of the many topics, objects, concepts and locations that were baked into the embedding space. The collection of vectors attached to one individual's screenshot sequence supports richly detailed descriptions of *intraindividual change*—the digital life trajectory that unfolds as an individual moves through a 1024-dimensional space. For example, we can describe when and how often individuals move toward and away from images of cats and people, topics related to climate change or music, content that they typically visit or new content, as well as how all these various changes in content are sequenced with respect to clock-time and each other. In sum, application of AI foundation models in accordance with zero-shot generalization immediately supports rich descriptions of complex (i.e., nonlinear) person-specific trajectories.

Models for prediction

The embeddings obtained from the foundation models' (i.e., LLMs') expressive and efficient architectures also provide a useful, compact starting point for a wide range of tasks. When using *few-shot learning*, the expensive ground truth data required to train new end-to-end prediction models is no longer needed. Rather, only small amounts of ground truth data are needed to fine tune the outputs derived from the high-dimensional representations for the new prediction task (i.e., only retraining the last few layers of the network). For example, following from how IBM's Watson NLP learns to score emotional sentiment of sentences in new languages, we can fine-tune image-text embeddings derived from CLIP into predictions of each screenshot's emotional valence using ground-truth data from just a few thousand hand-coded images (orders of magnitude less data than needed to train end-to-end models). In this way, propagation of new content labels speeds up and costs less. In the HSP, this means that any type of digital content and behavior of interest can be identified and studied in detail relatively quickly. Researchers simply need to produce relatively small batches of manually labeled screenshots and leverage transfer of all the implicit "knowledge" that was baked into the foundation model's embeddings during training. Now, rather than being constrained to the somewhat arbitrarily selected set of features that were of interest early on in HSP, new research questions about any specific domain or content can be identified and studied relatively quickly—with much lower cost than anticipated. Similarly, person-specific research questions about idiosyncratic expressions of behavior or content can be identified and studied at relatively low cost. For example, when an individual adopts a new social media platform, knowledge about their patterns of social media use can be transferred quickly to the new context. Researchers only need to identify and code a few instances of the new pattern. In sum, application of AI foundation models (i.e., LLMs) in accordance with few-shot learning greatly expands the range of predictable behaviors that can be examined using HSP data and the speed at which prediction models can be adapted to new content and new individuals.

Models for generation

Foundation models (i.e., LLMs) are deemed as foundational in part because the *many-shot learning* done during self-supervised training encodes the joint distribution, both $P(\mathbf{Y}|\mathbf{X})$ and $P(\mathbf{X})$, into the latent manifold of embeddings. Thus, the models are *generative models* that can be used to produce and simulate new data that looks very much like the training data (Dube, 2021). Employed within the context of the HSP, the flexibilities inherent in foundation model architectures open opportunities to obtain rich and expressive embeddings that both represent the multi-dimensional and multi-timescale complexity of individuals' digital lives *and* can be used to generate new human-like data that is similarly distributed and similarly complex.

In biology, newly trained transformer models are already being used to generate new molecules that may be useful for medical intervention (e.g., Dollar et al., 2021; Yang et al., 2021). In brief, a generative model trained on large databases of existing molecules is integrated with a reinforcement learning model trained on data-based *benchmarks* that indicate which molecules are likely to be useful for a particular purpose (e.g., bind to specific proteins). In essence, the reward structure becomes part of the latent embeddings (e.g., Decision Transformer; Chen et al., 2021), and can be used to fine-tune the generative model toward production of the most promising new molecules. The models thus provide a sandbox for drug discovery that puts many less humans at risk.

Analogously, the hundreds of millions of unstructured pixel arrays in the HSP data constitute a large library of humans' action possibilities—detailed observations of how individuals engage in a diverse array of complex tasks, how they switch among tasks, and how they simultaneously manage and mismanage multiple competing goals. Viewed as such, the data serve as an ideal platform for deriving *benchmarks of actual human behavior* (Raji et al., 2021) that might provide an aspirational standard for design of AI agents that mimic humans' ability to navigate through a nearly unbounded multimodal (text + image + icon) digital environment while balancing multiple competing goals and managing multiple interruptions and distractions. For example, a foundation model trained on HSP data ($N = 1$ or $N =$ many) could drive behavior of an autonomous AI agent that mimics how an individual with specific characteristics faces new "out-of-distribution" situations. Experiments with the AI agent might then lead to discovery of the specific constraints and interventions that might help individuals optimize their digital lives. In sum, the generative modeling capabilities that emerge when foundation models (LLMs) are trained on HSP data (in accordance with many-shot learning) can push forward development of dynamic, "just-in-time" interventions that can productively augment individuals' digital lives in real-time.

Cautions

First, *all* collections and uses of Screenome data + AI raise and require engagement with ethical issues and concerns. The data, and the models derived from them, are immensely personal. Thus, all the data inquires and modeling innovations are purposively directed through ethical frameworks that surface not only the privacy, respect, and beneficence issues, but also the fairness, representativeness and justice implications that inform and guide inclusion practices, algorithm development, and the actions that should be taken to mitigate potential dual use of AI augmentations in individuals' lives (Martinez-Martin et al., 2021). Second, we underscore that research on how the transformer architecture and foundation models work, when they fail, and how they can and should be applied is in a discovery mode. These models have challenged and are requiring revision of statistical theory and practice. For example, many deep learning models are not uniquely identified—meaning that different parameter values and perhaps even qualitatively different parameters might be obtained on re-fitting the model to the same data (see e.g., Khemakhem et al., 2020; Roeder, et al., 2021). Two instantiations of the same model might produce the same result—predictions—with vastly different model structures. This is both exciting and frightening. We are excited that these new models embody a reality parallel to how individuals (as unique dynamic systems) can solve the same problem using different neural structures. However, we are also somewhat frightened that the reproducibility and interpretability of model parameters that we often rely on for scientific explanations is being exchanged for black-box models that more accurately reproduce predictions and replicate human behavior (note, however, the recent resurgence of "explainable AI"). The main thesis of this paper is that in such cases, the model that has better transferability should be the preferred model, whether it is interpretable or not. Exciting directions forward might follow from recent advances in computational neuroscience where sophisticated measures for the predictivity of artificial to real human neural activity are being developed (Schrimpf et al., 2020; Yamins et al., 2014). While being buoyed by the opportunity to approach our data and science in new ways—*prioritizing transferability over generalizability*—we should remain somewhat cautious about how the models might be used to intervene in individuals' actual lives. Psychologists using these new models should develop precise goals regarding when iterative model formulations should maintain absolute consistency in internal structure and when they should maintain consistency in prediction performance and causal effect.

Conclusion

Comprehensive recordings of individuals' digital lives now facilitate description and modeling of a wide variety of behavioral processes. In this paper, we forwarded the Human Screenome Project as an instantiation of Molenaar's (2004) person-specific paradigm and outlined how these data, when combined with recent advances in AI, are facilitating a variety of descriptive, predictive, and generative tasks that provide better understanding of individuals' digital lives and design of AI agents that might eventually productively augment individuals lives. We considered how AI foundation models might be integrated into and support a robust person-specific paradigm that stays true to its $N = 1$ roots. In developing that possibility, we suggest Molenaar add a statement to the Manifesto—"In short, the concerns about *generalizability* commonly leveled at the person-specific paradigm are unfounded and can be fully and completely replaced with discussion and demonstrations of *transferability*." We are convinced that focus on transferability will speed description and prediction of individual behavior, and the development of personalized interventions that augment and optimize individuals' digital lives, health and well-being. We are excited to push forward!

Article information

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was supported by Stanford HAI.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The authors would like to thank Prof. Peter Molenaar for inspiration, everlasting intellectual exchange, and enduring friendship. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institution is not intended and should not be inferred.

ORCID

Nilam Ram http://orcid.org/0000-0003-1671-5257

References

Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(27), 7345–7352. https://doi.org/10.1073/pnas.1510507113

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(32), 15849–15854. https://doi.org/10.1073/pnas.1903070116

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., & Brynjolfsson, E. (2021). On the opportunities and risks of foundation models. arXiv Preprint arXiv, 2108.07258.

Brick, T. R., Koffer, R. E., Gerstorf, D., & Ram, N. (2017). Feature selection methods for optimal design of studies for developmental inquiry. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, *73*(1), 113–123. https://doi.org/10.1093/geronb/gbx008

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & Agarwal, S. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., & Mordatch, I. (2021). Decision transformer: Reinforcement learning via sequence modeling. *Advances in Neural Information Processing Systems*, *34*, 15084–15097.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv Preprint arXiv:1810.04805.

Dollar, O., Joshi, N., Beck, D. A., & Pfaendtner, J. (2021). Attention-based generative models for de novo molecular design. *Chemical Science*, *12*(24), 8362–8372. https://doi.org/10.1039/d1sc01050f

Dube, S. (2021). *Intuitive exploration of artificial intelligence*. Springer International Publishing.

Economist. (2022). Huge "foundation models" are turbo-charging AI progress. The Economist. 10 June 2022, ISSN 0013-0613. Retrieved 10 June 2022.

Gates, K. M., & Molenaar, P. C. (2012). Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. *NeuroImage*, *63*(1), 310–319. https://doi.org/10.1016/j.neuroimage.2012.06.026

Gates, K. M., Molenaar, P. C. M., Hillary, F., Ram, N., & Rovine, M. (2010). Automatic search for fMRI connectivity mapping: An alternative to Granger causality testing using formal equivalences between SEM path modeling, VAR, and unified SEM. *NeuroImage*, *50*(3), 1118–1125. https://doi.org/10.1016/j.neuroimage.2009.12.117

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Khemakhem, I., Kingma, D., Monti, R., & Hyvarinen, A. (2020, June). Variational autoencoders and nonlinear ica:

A unifying framework. In *International Conference on Artificial Intelligence and Statistics* (pp. 2207–2217). Proceedings of Machine Learning Research, 108.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv Preprint arXiv:1907.11692.

Martinez-Martin, N., Luo, Z., Kaushal, A., Adeli, E., Haque, A., Kelly, S. S., Wieten, S., Cho, M. K., Magnus, D., Fei-Fei, L., Schulman, K., & Milstein, A. (2021). Ethical issues in using ambient intelligence in health-care settings. *The Lancet. Digital Health*, *3*(2), e115–e123. https://doi.org/10.1016/S2589-7500(20)30275-2

Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research & Perspective*, *2*(4), 201–218. https://doi.org/10.1207/s15366359mea0204_1

Molenaar, P., Boomsma, D. I., & Dolan, C. V. (1993). A third source of developmental differences. *Behavior Genetics*, *23*(6), 519–524. https://doi.org/10.1007/BF01068142

Molenaar, P. C. M., & Ram, N. (2010). Dynamic modeling and optimal control of intraindividual variation: A computational paradigm for nonergodic psychological processes. In S.-M. Chow, E. Ferrer, & F. Hsieh (Eds.), *Statistical methods for modeling human dynamics: An interdisciplinary dialogue* (pp. 13–37). Routledge/Taylor & Francis Group.

Molenaar, P. C. M., *Sinclair, K., Rovine, M. J., Ram, N., & Corneal, S. E. (2009). Analysis of developmental processes based on intra-individual variation by means of non-stationary time series modeling. *Developmental Psychology*, *45*, 260–271.

Palatucci, M., Pomerleau, D., Hinton, G. E., & Mitchell, T. M. (2009). Zero-shot learning with semantic output codes. *Advances in Neural Information Processing Systems*, *22*

Pearl, J., & Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, *29*(4), 579–595. https://doi.org/10.1214/14-STS486

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., & Krueger, G. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (pp. 8748–8763). PMLR.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the everything in the whole wide world benchmark. arXiv Preprint arXiv:2111.15366.

Ram, N., Brose, A., & Molenaar, P. C. M. (2013). Dynamic factor analysis: Modeling person-specific process. In T. Little (Ed.), *Oxford handbook of quantitative methods Volume 2 Statistical Analysis* (pp. 441–457). New York: Oxford University Press.

Ram, N., Conroy, D., Pincus, A. L., Lorek, A., Rebar, A. H., Roche, M. J., Morack, J., Coccia, M., Feldman, J., & Gerstorf, D. (2014). Examining the interplay of processes across multiple time-scales: Illustration with the Intraindividual Study of Affect, Health, and Interpersonal Behavior (iSAHIB). *Research in Human Development*, *11*, 142–160.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. arXiv Preprint arXiv:2204.06125.

Reeves, B., Ram, N., Robinson, T. N., Cummings, J. J., Giles, C. L., Pan, J., Chiatti, A., Cho, M. J., Roehrick, K., Yang, X., Gagneja, A., Brinberg, M., Muise, D., Lu, Y., Luo, M., Fitzgerald, A., & Yeykelis, L. (2021). Screenomics: A framework to capture and analyze personal life experiences and the ways that technology shapes them. *Human-Computer Interaction*, *36*(2), 150–201. https://doi.org/10.1080/07370024.2019.1578652

Reeves, B., Robinson, T. R., & Ram, N. (2020). Time for the human Screenome project. *Nature*, *577*(7790), 314–317. https://doi.org/10.1038/d41586-020-00032-5

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., & Schmidt, K. (2020). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007.

Srivastava, S., Li, C., Lingelbach, M., Martín-Martín, R., Xia, F., Vainio, K. E., Lian, Z., Gokmen, C., Buch, S., Liu, K., & Savarese, S. (2022). BEHAVIOR: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In *Conference on Robot Learning* (pp. 477–490). PMLR.

Tuarob, S., Tucker, C. S., Kumara, S., Giles, C. L., Pincus, A. L., Conroy, D. E., & Ram, N. (2017). How are you feeling? A personalized methodology for predicting mental states from temporally observable physical and behavioral information. *Journal of Biomedical Informatics*, *68*, 1–19. https://doi.org/10.1016/j.jbi.2017.02.010

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30).

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

Yang, L., Yang, G., Bing, Z., Tian, Y., Niu, Y., Huang, L., & Yang, L. (2021). Transformer-based generative model accelerating the development of novel BRAF inhibitors. *ACS Omega*, *6*(49), 33864–33873. https://doi.org/10.1021/acsomega.1c05145

Yarkoni, T. (2020). The generalizability crisis. *The Behavioral and Brain Sciences*, *45*, 1–78. https://doi.org/10.1017/S0140525X20001685

Yu, H., Xie, T., Paszczyñski, S., & Wilamowski, B. M. (2011). Advantages of radial basis function networks for dynamic system design. *IEEE Transactions on Industrial Electronics*, *58*(12), 5438–5450. https://doi.org/10.1109/TIE.2011.2164773