How to Build an Observatory for the Mind

Vencislav Popov (vencislav.popov@gmail.com) & Klaus Oberauer (k.oberauer@psychologie.uzh.ch)

Department of Psychology, University of Zurich

Abstract: Psychological theories are rarely explicitly rejected or gradually incorporated into a cumulative theoretical structure. Instead, they rise and fall with the whims of individual researchers. Formal modeling offers a potential solution to this crisis by strengthening the link between theories, hypotheses, and data, and are becoming increasingly popular in many subfields of the experimental behavioral sciences. However, evaluating these models poses its own set of challenges. Current methods often rely on small, ad-hoc experiments that lack the statistical power, generality and diversity needed to robustly test these models. To solve these issues, we propose to create an Observatory of the Mind – an independent scientific institute dedicated to the systematic collection and analysis of experimental behavioral data. The institute will maintain a large pool of long-term semi-professional participants, representing a cross-section of the global population, who will complete a comprehensive set of empirical benchmark tasks for behavioral science over several years. The goal is to produce a publicly available body of knowledge that can support the development and evaluation of formal models of psychological processes and enable algorithmic discovery of novel theories. We argue that such an observatory is necessary to overcome the current crisis in psychological research, characterized by weak theories, unreliable findings, and low generalizability. We also discuss the practical benefits and the logistical challenges of implementing this project. We invite the scientific community to engage in a constructive dialogue about the feasibility and potential impact of such an observatory.

Psychology is facing a crisis of confidence. Despite decades of empirical and theoretical work, hundreds of thousands of researcher hours, and millions spent in funding, the field has progressed slowly[1–4]. Psychological theories tend to be weak and vague[3], empirical findings are often unreliable[5] and non-generalisable[6], and researchers struggle to agree on a core body of knowledge. These problems are interdependent – they stem from a vicious cycle of poor theory development[3,7], publication bias, questionable research practices[8], toxic incentives[9], small and biased sampling[10,11] (Figure 1). As a result, psychological theories, rather than being explicitly rejected or being gradually incorporated into a cumulative theoretical structure, usually rise and fall with the whims of individual researchers. This state of affairs have led some to question whether psychology has delivered on its promises to the public[12] or even to call it "technologically worthless"[1].

What if there was a way to break free from this vicious cycle? What if we borrowed a leaf from astronomy and built an observatory, not for the stars, but for the mind?

Our proposition is radical: To create an independent scientific institute, much like the Space Telescope Science Institute that operates the Hubble Space Telescope in astronomy, tasked with the systematic collection and analysis of experimental behavioral data. This institute will maintain a large pool of long-term semi-professional participants for behavioral research, representing a cross-section of the global population. These participants will each undertake a comprehensive set of empirical benchmarks tasks for experimental behavioral sciences, such as behavioral economics, cognitive psychology, psycholinguistics, social psychology, and others, over many years, providing a continuously updated, fine-grained, and individual-level dataset of various behavioral and cognitive tasks. The goal is to produce an agreed-upon publicly available body of knowledge by accumulating large-scale high-quality data for reliable benchmark findings at the level of individual participants.

This is more than an answer to psychology's replication crisis and a fulfilment of open science reform ideals. It's a way to give theoretical development a solid empirical base, to support the development of integrative formal models, and to enable algorithmic discovery of novel theories. This novel approach presents an opportunity to reclaim psychology's credibility as a scientific discipline and make it "technologically useful", allowing us to build personalized models of learning, decision-making and reasoning, or to identify early markers for cognitive decline.

However, there are logistical challenges that need to be addressed. Here, we present the project and explain why it is necessary, the problems it solves, and the logistical issues we anticipate. We also urge the scientific community to contribute innovative solutions, express concerns, and make suggestions. The completion of this project would require a massive international collaboration and new funding mechanisms.

A crisis of what exactly?

The open science movement has brought excellent suggestions to improve the reliability of results, such as making data and analysis software publicly available[13], preregistering research designs and analysis plans before data collection[14], making publication decisions without seeing the data[15], making replication efforts mainstream[16], and rewarding such practices appropriately[13,17]. In addition, distributing data collection among multiple labs aims to drastically increase the size and representatives of studied samples[18]. Analyzing data via multiple teams[19], sometimes blinded to experimental conditions[20], aims to reduce bias and increase generalization of results. Although controversial, such policies have the potential to solve at least some of the problems mentioned above.

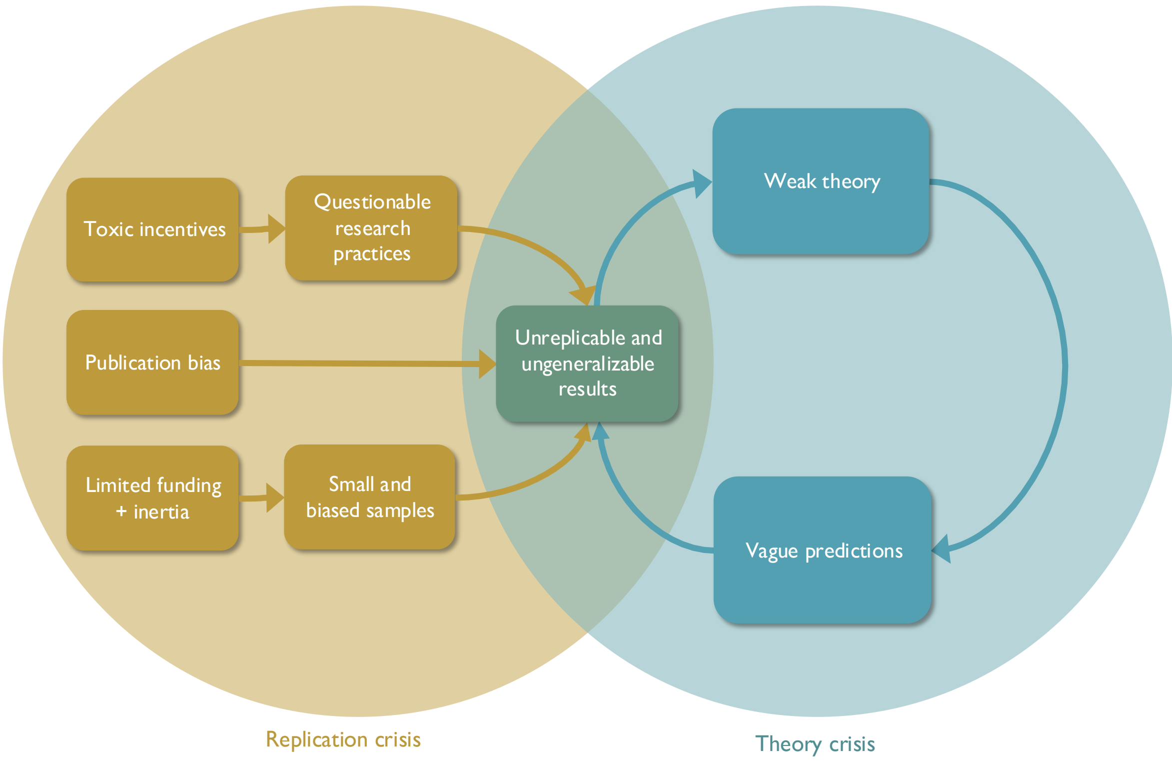

Figure 1. A vicious cycle in which weak theories combined with structural problems lead to unreliable results, which in turn reinforce weak theorizing.

There is a deeper problem, however, because the open science reform has focused on improving the replicability and reliability of empirical results, without addressing the question of what to do with such results. The primary goal of science is not only to collect empirical regularities, but to also explain such regularities in an integrated, parsimonious, precise, intellectually satisfying and practically useful way. Such theories in psychology are few and far between. The weakness of our theories is itself one source of the low replicability of empirical findings[3,21]. The replication crisis is, to a large part, a symptom of a theory crisis in behavioral sciences.

There is growing awareness in the field that overcoming the theory crisis requires us to formalize theories into mathematical or computational models – models whose predictions can be derived independently of the researchers who created them, through either formal calculation or computer simulations. Formal models offer many advantages for strengthening the link between theories, hypotheses, and data, and are becoming increasingly popular in many subfields of the experimental behavioral sciences.

There are, however, several hurdles for the systematic development and evaluation of formal models of psychological processes. Currently, most models are evaluated in an idiosyncratic fashion. Typically, the researchers who have developed a model select a few findings in the field that seem most relevant or interesting to them. As a result, even though multiple models aim to explain the same cognitive capacity (for example, working memory or decision making), each of them only applies to a subset of partly overlapping findings, making model comparison difficult, if not impossible[4,22].

This issue can be overcome by the systematic compilation of agreed-upon "benchmark findings" – key results in a field that the community agrees are valid, have some degree of generality, and are important for theory development. As an example, one of us (K.O.) previously led a collaborative effort to compile such a set of benchmark findings in the domain of working memory[22].

A consensus list of benchmark findings for a research field is only the first step. Formal models aiming to explain the benchmarks need to be evaluated against data that represent them. These data need to meet requirements that current data don't meet. First, we need data that represent all empirical benchmarks in the same sample of participants, tested with comparable experimental protocols. Only then can we reasonably demand that a successful model explains multiple – ideally all – benchmarks with the same set of assumptions and parameter values. Only then can we test model predictions about interactions and correlations between different benchmarks.

Second, the data need to stem from a broad sample not only from the population of people but also from the population of situations in which we observe their behavior. This means that the participant sample consists not only of university students from WEIRD societies[10], but represents people from a broad range of age groups, educational experience, socio-economic status, and cultural backgrounds. It also means that experimental effects are tested for a broad range of stimuli and experimental parameters[23,24]. This is necessary to gauge how far the benchmark effects generalize[6]. Successful models need to accommodate that level of generality. For instance, if a model of human memory predicts an effect that holds only for psychology students tested on memory for English words, we cannot speak of a good match between theory and data.

Third, the data need to be fine-grained, representing behavior on the level of individual responses of individual people. This is necessary to test model predictions about the distribution of responses and response times, as well as trends over time, and about individual differences in behavior. Neglecting individual differences in behavior can result in averaging artefacts. This occurs when the collective behavior of a group presents a distorted image that does not accurately reflect the behavior of any individual[25]. In a recent critique of the memory field, one of us (V.P.) argued that to achieve progress, formal models of memory and cognition need to be able to make parameter-free predictions at the level of individual people across a large set of different tasks[4]. The same applies to models in other subfields of psychology. The only way to achieve "technologically useful" theories is to hold them to this standard; however, we have no dataset that would allow us to achieve this goal.

Agenda for an observatory of the mind

To address the theory crisis in the behavioral sciences and align with open science reform ideals, we propose a radical change in the collection and analysis of empirical benchmark findings. Taking inspiration from the Hubble telescope model in astronomy, we suggest the establishment of an independent scientific institute dedicated to building and maintaining a large pool of long-term semi-professional participants for behavioral research. These participants, representative of the global population, would complete a comprehensive set of empirical benchmark tasks for behavioral science over several years. The aim is to gather a detailed and extensive dataset of various behavioral and cognitive tasks with precise measurements at the individual participant level (Figure 2).

Our approach introduces three key innovations for basic science. First, the same participants will complete all tasks, enabling the evaluation of explanatory models across tasks, separately for each participant. Second, large amounts of high-quality data will be collected for each participant and each task, minimizing measurement and sampling noise. Third, data collection will be continuous, following the initial phase shown in Figure 3. Researchers in the field will have the opportunity to propose data collection with this participant pool, which will be reviewed by an independent committee representing various psychological subfields and scientific societies. The resulting data would be openly available and citable, encouraging data contribution and reuse.

This proposal offers an efficient solution to several pressing issues in behavioral research. It would provide the detailed individual-level data that we need for systematic evaluation of explanatory models, and for the potential algorithmic discovery of novel theories. Additionally, it would help to overcome bias in the empirical record in three ways: It would substantially broaden the inclusiveness of people represented in it. It would buffer it against biases of individual researchers because decisions about studies to run, and about their design, will be made more collectively. Finally, it would protect against publication bias because the data will be published regardless of the outcome. Our proposal would also make two contributions to overcome the replication crisis. One is to provide more robust data through larger samples, and a broader sampling of experimental stimuli and parameters. The other is to enable and encourage the development of stronger theories.

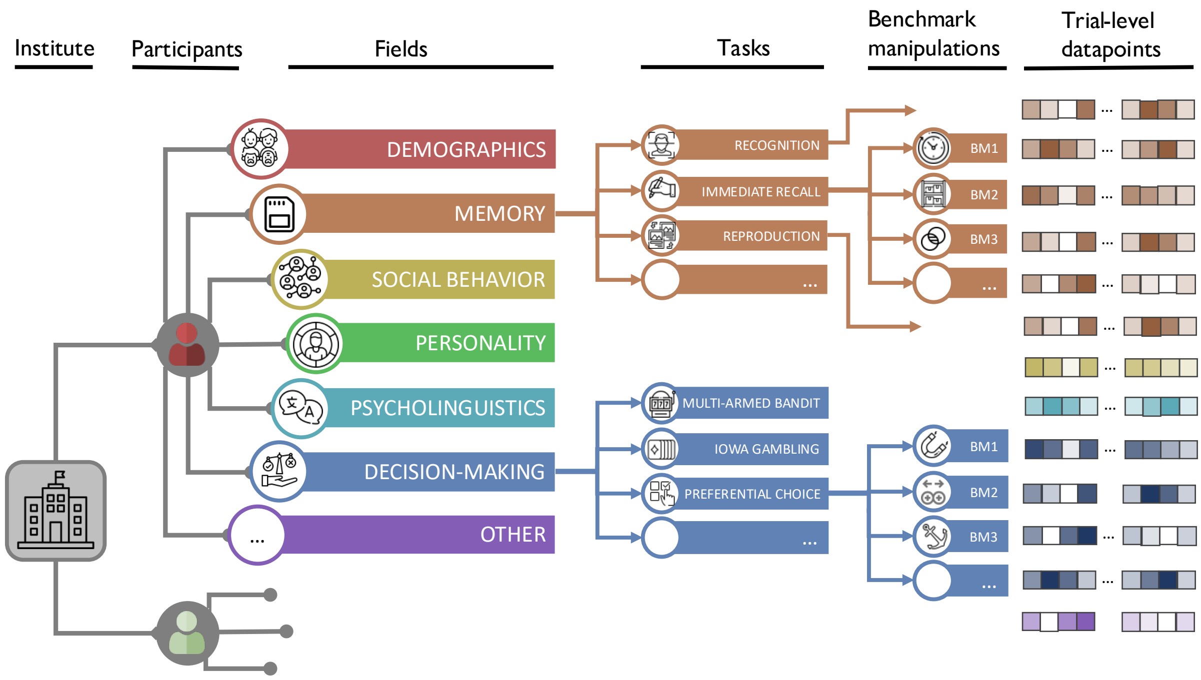

Figure 2. The Observatory of the Mind. The institute maintains a large representative sample of semi-professional participants, each of which completes many experimental tasks in different fields of behavioral science over the course of multiple years. Tasks are selected to be representative of key benchmarks findings in each field, and broad range of experimental manipulations are constructed to elicit such benchmark behavior at the level of individual participants. This results in rich trial-level measurements for each manipulation and participant, allowing comprehensive theory testing and novel theory development. One task for each of two example fields is illustrated, but the complete dataset would comprise multiple tasks per field in a wide array of experimental behavioral sciences.

Initial data collection will focus on established benchmark findings that the community has agreed are important for theory development and evaluation. Although eventually the dataset would include tasks from many different fields, we propose to begin with a field with already established lists of benchmarks, such as human memory or decision making for three reasons: 1) such benchmark findings have already been compiled by a diverse set of researchers[22], allowing us to jump start the process; 2) there exist many competing formal models of memory and decision making, which are currently difficult to evaluate against each other; 3) models of memory would provide some of the most immediate practical benefits, as we explain below. To broaden the scope, concurrently with this process, we invite other subfields of behavioral science to collate their own lists of benchmark findings to propose for collection by the Observatory in a subsequent phase. These two phases will be followed by an open call for proposals, which will continuously update the database with benchmark findings identified in the future. In this way we aim to create a long-term system that responds dynamically to new theoretical and empirical discoveries in the field.

In the scope of this project, given the high value of the limited participant resources we propose to build, data proposals should not aim to test novel hypotheses or to replicate contentious findings. Data proposals should align with the goals of the observatory, which is to collect robust data representing theoretically relevant empirical benchmarks, and connect them to existing data. We aim to treat data proposals similarly to registered reports. Upon approval and data collection, data will be made immediately publicly available. A DOI will be assigned to each independent data contribution that is added when a proposal is accepted. In this way, anyone who uses the data in the future would be required to cite it, creating a strong incentive for researchers to make strong proposals.

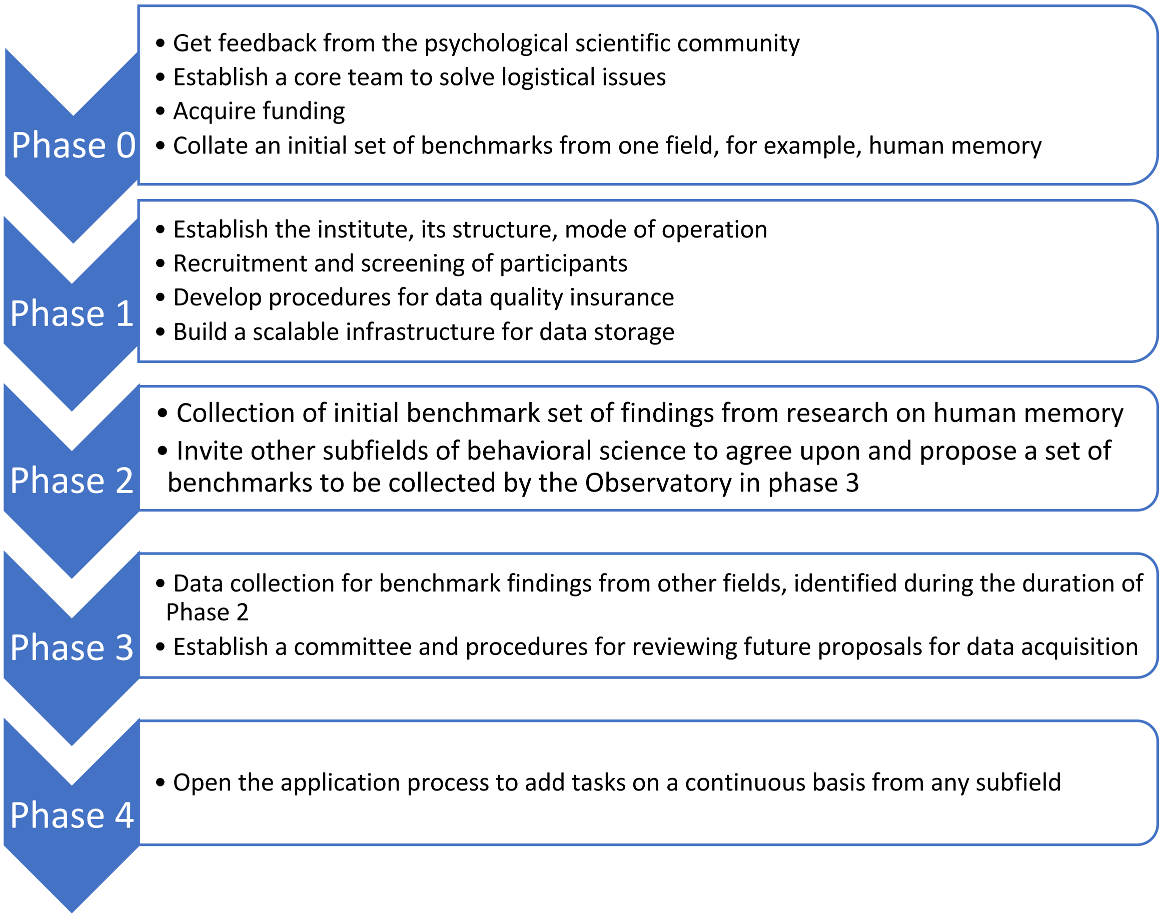

Figure 3. Proposed phases of the Observatory of the Mind

The goal of the Observatory is not to replace traditional laboratory science, but to complement it. In the standard model of doing behavioral science, individual scientists derive predictions from theories, and test those predictions in ad hoc participant samples. This approach is ideally suited for discovering novel empirical effects, due to its generally low cost, flexibility, and fast-paced nature, leading to the establishment and verfication of novel benchmarks for theories. The Observatory aims to consolidate the existing benchmarks, and to integrate future discoveries on an ongoing basis, providing solid ground for theory testing.

Finally, the proposed continuous data gathering format lends itself well to organizing prediction challenges for the community. As in other fields such as artificial vision[26], protein folding[27] or brain responses to images[28], this has the potential to reinvigorate interest in developing rigorous formal models that can be comapred in their ability to predict held out data. Parts of the initial bechmark data gathered by the Observatory could be embargoed for such a challenge, and due to the ongoing nature of the data collection, annual challenges can be organized to predict novel benchmarks.

Practical benefits

Aside from the clear benefits for basic research, such a project would have substantial practical benefits for society. Long-term tracking of individual participants across many cognitive tasks could provide an opportunity to identify early markers of cognitive decline, semantic dementia, and Alzheimer's, allowing for early detection and treatment. Developing learning models that can successfully make predictions about individual people's memory performance could revolutionize education by creating custom tailored cognitive tutors. Even a 1-5% increase in learning outcomes when employed on a population scale could have a massive impact on the level of skill and knowledge in a society. Such personalized cognitive tutors would be particularly beneficial for disadvantaged populations who cannot afford expensive one-on-one tutoring.

Recent advances in AI and machine learning have also led to the possibility to perform computer-assisted scientific discovery in other fields. Machine learning algorithms have been able to reconstruct Kepler's third law of planetary motion, Einstein's relativistic time-dilation law, and Langmuir's theory of adsorption purely from data without any human input[29]. Even though we practice and support more traditional theory-building, we cannot deny the potential power of such techniques. Applying automated models for scientific discovery of psychological theories is currently impossible, but the availability of a large-scale dataset at the level individual participants will open many possibilities, accelerating progress and allowing us to capture the complexities of human cognition and behavior.

Challenges

There are precedents for a project of such a scale in psychology. One example among several crowdsourced multi-lab initiatives[18] is The Psychological Science Accelerator[30] (PSA), a distributed network of labs whose goal is to coordinate data collection for large-scale studies of human behaviour. Projects are selected by a democratically elected scientific committee. This institution shares similarities with our proposal in the way it operates, but differs in one crucial aspect - studies are not run with the same sample of participants, and hence, data across different studies is not linked through the same participants. Rather studies selected usually have the form of traditional single-lab studies, such that their design is usually limited to a specific paradigm of interest, aiming to test specific hypotheses, rather than gathering an exhaustive dataset of benchmark findings across a field.

The PSA has admirable goals and much overlap with our proposal, but it serves a different, if related goal. Just like NASA operates several different "Great Telescopes" in different regions of the electromagnetic spectrum, the Observatory we propose and the PSA can operate independently to serve their goals. Examining the challenges that the PSA has faced[18] makes it clear that the success of tour project will depend on two things – broad community support and funding.

Large-scale investment is needed to establish and ensure the longevity of the Observatory of the Mind. By far the largest cost would be the payment for participants. To ensure high data quality and sustained participation over years, participants need to be compensated at a level commensurate with median salaries in the originating country. For example, if the project involved 1000 part-time participants who each contribute two hours of data per day, at the median salary level in the US, the annual cost would be around 14,000,000 USD. Additional funds would be required for scientific staff to run the organization, to collect the data, and to do research on maximizing data quality. Continuous research and development of engineering solutions for data collection and quality assurance is also necessary. Tools and computational resources for efficient data storage and access need to be developed and maintained.

These funding aims are not unrealistic, given the potential benefits of solving basic science and societal problems. In 2016, approximately $2.1 billion in federal funding was directed towards psychological research in the US alone[31]; however, these funds are distributed among individual labs in the form of small grants, and current funding mechanisms in behavioral science are inadequate to support such a large-scale project.

Given the fundamental problems that the Observatory will address, the cost is justified. This will become the largest cross-task behavioral dataset in existence. Other solutions for scaling up, such as calls for labs to dramatically increase their sample sizes[32] or the use of proprietary industry data[33], might not be feasible for most individual researchers. Our proposal will give researchers equal opportunities to use carefully curated large-scale data, even if they are not well funded or well connected to acquire it.

We envision several possible funding routes. First, new international funding mechanisms through existing funding agencies could be negotiated, similarly to how large collaborative research projects in physics, astronomy and biology have been funded. Second, private foundations interested in the societal benefits of this program could become involved. Third, regulated access to the collected data could be sold to companies aiming to develop ethical AI products that align with the goals, mission, and values of our organization. While we would prefer this enterprise to be funded by non-commercial means, commercial partners whose goals align with our missions and ethics could be an asset. Ultimately, a diverse set of public and private funding mechanisms might be necessary for the longevity of the project.In addition to funding, several logistical challenges such as benchmark selection, data collection, data privacy, and dealing with dropouts, will need to be solved. Table 1 provides a non-exhaustive list of challenges and some possible solutions. To democratize the process, to harness existing expert knowledge in the field, and to build innovation from the start, we invite the community to contribute innovative solutions, express concerns, make suggestions and note their interest in being involved in the project.

Our proposal is an ambitious one. But we believe that an ambitious change is precisely what is needed to overcome the current crisis in psychological research. We urge the research community to engage in open and constructive dialogue about the viability and potential impact of such an observatory. Together, we can shape the future of psychological research, bringing it into a new era of technological relevance and scientific robustness. If we are to make psychology truly technologically useful and scientifically credible, then we need to think big. An observatory for the mind might be just what we need.

Table 1. Logistical challenges for implementing an Observatory of the Mind and some possible solutions

Challenge	Possible solutions
Funding - $14 million per year for participant payments (assuming a 1000 participants) - Data storage and infrastructure - Management - Project scientists - Software engineers	- New international governmental funding mechanisms - Private foundations - Commercial businesses
Privacy, data storage and data access - Large-scale publicly open individual data could make it difficult to maintain the privacy of participants due to the possibility of triangulation[34]	- Assessment of privacy concerns - Cloud-based access to data[35,36] - Restrict and monitor access to data, while still making it available

Participant recruitment and dropouts - How to ensure representativeness of sample? How many participants? - It would be expensive and time-consuming to replace participants who quit the study after a prolonged period. How to minimize dropout or replace participants efficiently?	- Well-paid compensation for participants - Stress the scientific value of the project - Identify participants who perform similarly on a set of diagnostic tasks. This would obviate the need for replacement participants to complete all tasks of dropouts
Data quality - Minimize practice, fatigue and transfer effects from one task to another - Minimize measurement noise	- Careful design and selection of studies and materials to avoid task-order effects - Task design aimed at maximizing and sustaining motivation - In-house additional research on psychological measurement
Data collection - Online or in-person?	- Online data collection can be easier to implement, but more difficult to control - In-person data collection could use existing distributed resources, such that individual labs are compensated by the project to offer their resources for testing
Benchmark selection - How will benchmarks or new data be selected for testing?	- Initial benchmarks in Phase 2 can be compiled by subgroups in different fields following the example of Oberauer and colleagues[22] - A diverse committee reviews proposals for benchmark inclusions - Two-tiered review system in Phase 4: 1) applications for using participants' time with applicants funding, 2) applications for using participants' time with the institute's funding. More strict selection criteria for tier 2.
Order effects - How to ensure that the behavior of semi-professional participants is representative of the population? How to guard against task order effects or the development of non-standard strategies?	- Sample naive participants for some key tasks at some intervals and compare - Half sample does the tasks in the same order, half are randomized

References

1. Meehl, P. E. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *J. Consult. Clin. Psychol.* **46**, 806 (1978).
2. Newell, A. You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. in *Visual information processing* (ed. Chase, W. G.) (Academic Press, 1973).
3. Oberauer, K. & Lewandowsky, S. Addressing the theory crisis in psychology. *Psychon. Bull. Rev.* **26**, 1596–1618 (2019).
4. Popov, V. If God Handed Us the Ground-Truth Theory of Memory, How Would We Recognize It? *PsyArXiv* (2023) doi:10.31234/osf.io/ay5cm.
5. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
6. Yarkoni, T. The generalizability crisis. *Behav. Brain Sci.* **45**, e1 (2022).
7. Gigerenzer, G. Personal Reflections on Theory and Psychology. *Theory Psychol.* **20**, 733–743 (2010).
8. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366 (2011).
9. Lilienfeld, S. O. Psychology's Replication Crisis and the Grant Culture: Righting the Ship. *Perspect. Psychol. Sci.* **12**, 660–664 (2017).
10. Henrich, J., Heine, S. & Norenzayan, A. The weirdest people in the world? *Behav. Brain Sci.* **33**, 61–83 (2010).
11. Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
12. Schlinger, H. D. Why Psychology Hasn't Kept Its Promises. *J. Mind Behav.* **25**, 123–144 (2004).
13. Nosek, B. A. *et al.* Promoting an open research culture. *Science* **348**, 1422–1425 (2015).
14. Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. *Proc. Natl. Acad. Sci.* **115**, 2600–2606 (2018).
15. Chambers, C. D. & Tzavella, L. The past, present and future of Registered Reports. *Nat. Hum. Behav.* **6**, 29–42 (2022).
16. Zwaan, R. A., Etz, A., Lucas, R. E. & Donnellan, M. B. Making replication mainstream. *Behav. Brain Sci.* **41**, e120 (2018).

17. Moher, D. *et al.* Assessing scientists for hiring, promotion, and tenure. *PLOS Biol.* **16**, e2004089 (2018).

18. Coles, N. A., Hamlin, J. K., Sullivan, L. L., Parker, T. H. & Altschul, D. Build up big-team science. *Nature* **601**, 505–507 (2022).

19. Wagenmakers, E.-J., Sarafoglou, A. & Aczel, B. One statistical analysis must not rule them all.

20. Starns, J. J. *et al.* Assessing Theoretical Conclusions With Blinded Inference to Investigate a Potential Inference Crisis. *Adv. Methods Pract. Psychol. Sci.* **2**, 335–349 (2019).

21. Guest, O. & Martin, A. E. How Computational Modeling Can Force Theory Building in Psychological Science. *Perspect. Psychol. Sci.* 1745691620970585 (2021) doi:10.1177/1745691620970585.

22. Oberauer, K. *et al.* Benchmarks for models of short-term and working memory. *Psychol. Bull.* **144**, 885–958 (2018).

23. Baribault, B. *et al.* Metastudies for robust tests of theory. *Proc. Natl. Acad. Sci.* **115**, 2607–2612 (2018).

24. Almaatouq, A. *et al.* Beyond Playing 20 Questions with Nature: Integrative Experiment Design in the Social and Behavioral Sciences. *Behav. Brain Sci.* 1–55 (2022) doi:10.1017/S0140525X22002874.

25. Estes, W. K. The problem of inference from curves based on group data. *Psychol. Bull.* **53**, 134–140 (1956).

26. Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. Preprint at https://doi.org/10.48550/arXiv.1409.0575 (2015).

27. Method of the Year 2021: Protein structure prediction. *Nat. Methods* **19**, 1–1 (2022).

28. Allen, E. J. *et al.* A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. Neurosci.* **25**, 116–126 (2022).

29. Cornelio, C. *et al.* Combining data and theory for derivable scientific discovery with AI-Descartes. *Nat. Commun.* **14**, 1777 (2023).

30. Moshontz, H. *et al.* The Psychological Science Accelerator: Advancing Psychology Through a Distributed Collaborative Network. *Adv. Methods Pract. Psychol. Sci.* **1**, 501–515 (2018).

31. How much federal funding is directed to research in psychology? *https://www.apa.org* https://www.apa.org/monitor/2017/04/datapoint.

32. Griffiths, T. L. Manifesto for a new (computational) cognitive revolution. *Cognition* **135**, 21–23 (2015).

33. Steyvers, M. & Schafer, R. J. Inferring latent learning factors in large-scale cognitive training data. *Nat. Hum. Behav.* **4**, 1145–1155 (2020).

34. Wicherts, J. M. *et al.* How to protect privacy in open data. *Nat. Hum. Behav.* **6**, 1603–1605 (2022).
35. de Kok, J. W. T. M. *et al.* A guide to sharing open healthcare data under the General Data Protection Regulation. *Sci. Data* **10**, 404 (2023).
36. Dennis, S. *et al.* Privacy versus open science. *Behav. Res. Methods* **51**, 1839–1848 (2019).