BEYOND 20 QUESTIONS WITH NATURE: INTEGRATIVE EXPERIMENT DESIGN

(ALMAATOUQ, GRIFFITHS, SUCHOW, WHITING, EVANS AND WATTS, 2022)

Karlo Doroc

PhD Student – CBMM – University of Melbourne

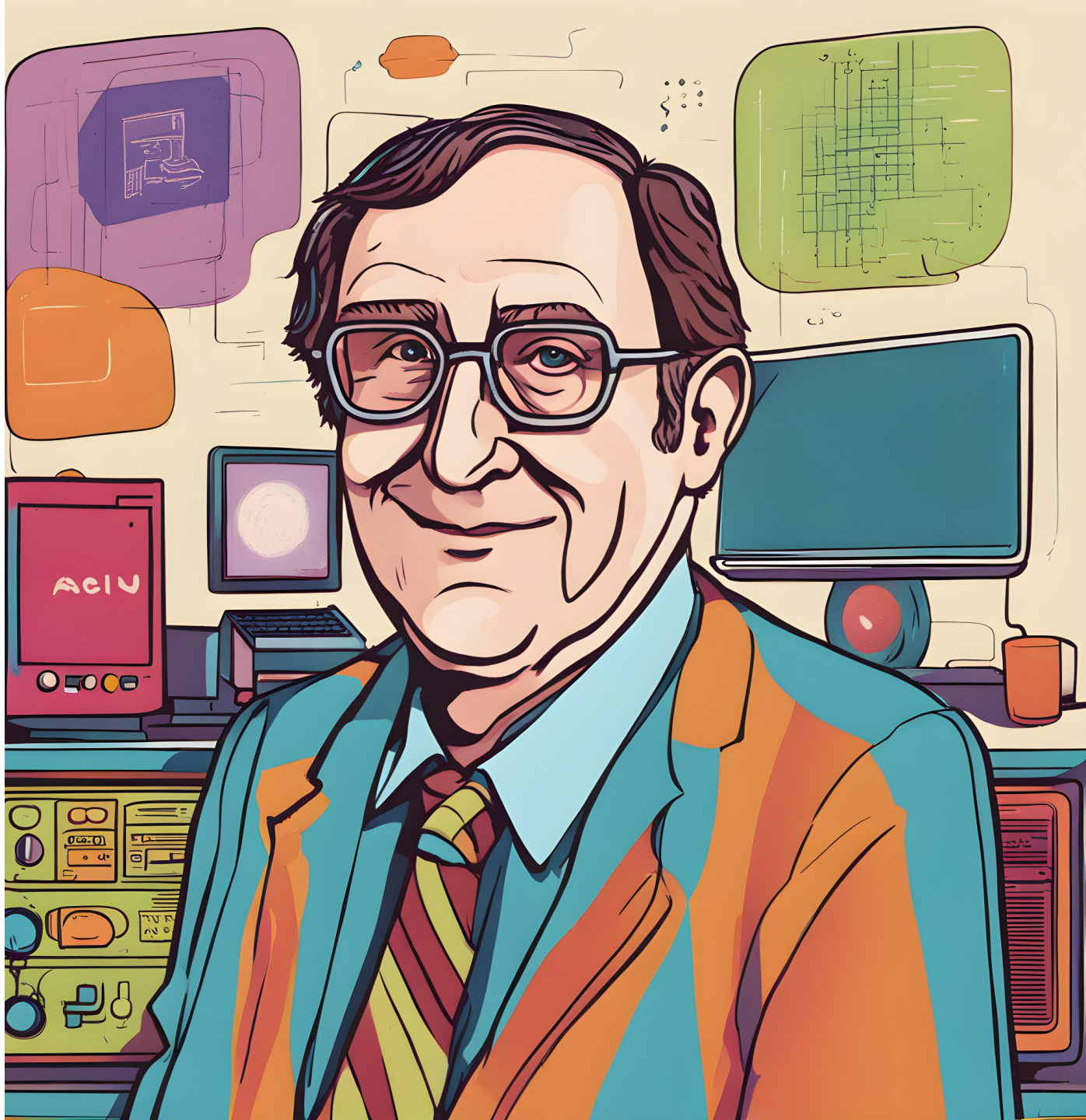
OUR ROADMAP

“Science advances by playing twenty questions with nature. The proper tactic is to frame a general question, hopefully binary, that can be attacked experimentally. Having settled that bits-worth, one can proceed to the next…

Unfortunately, the strategy does not seem to work… We never seem in the experimental literature to put the results of all the experiments together.

You can't play 20 questions with Nature and win”

- Turing Award winner Allan Newell, 1973.



THE "ONE-AT-A-TIME" PARADIGM

- Research question about the relation between independent and dependent variables
- Theory-motivated hypothesis
- Experiment tests the hypothesis
- Analysis is performed, evidence assessed, and findings are generalised
- Future research builds on your findings sequentially

INCOMMENSURABILITY IS BAKED-IN BY DESIGN I

- "each experiment tests at most a small number of theoretically informed hypotheses in isolation by varying at most a small number of parameters."
- The only parameters or variables of interest are those explicitly articulated in the theory - where the theory is silent they are deemed irrelevant
- Critically, if deemed irrelevant these factors are often *not reported*

INCOMMENSURABILITY IS BAKED-IN BY DESIGN II

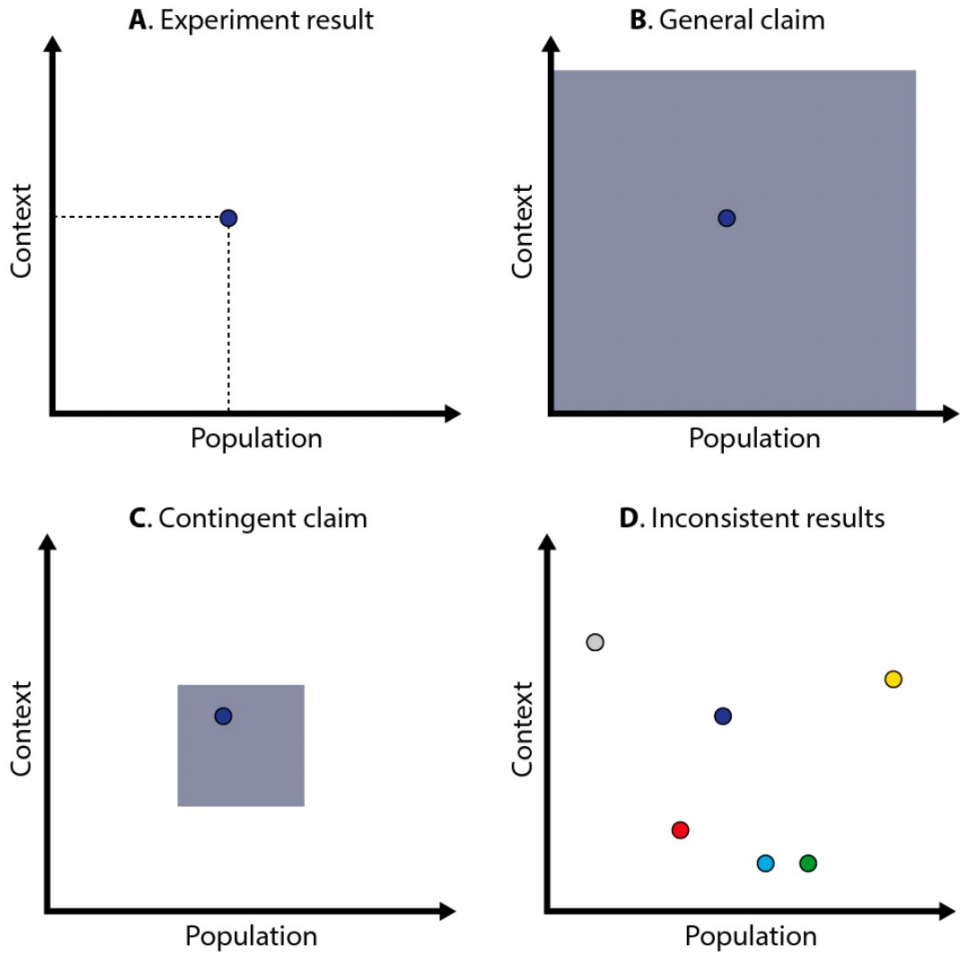
- Different research groups will invariably make different design choices, often in an arbitrary, vague or undocumented fashion, making studies incommensurable
- Experiments are designed with the sole aim of testing a hypothesis, but *not* with the goal of comparing results with other experiments in the same area
- We lack a structured framework with which to articulate how any one experiment differs from others
- Reviews and meta-analysis are post-hoc and take commensurability as given

EXAMPLE: GROUP SYNERGIES

- Question: does the performance of an interacting group exceeding that of an equivalently sized "nominal group" of individuals working independently?

- Example of factors that are relevant
  - DV: performance on a task
  - IV: a collection of individuals vs teams
  - Non-theory variables: the specific task, task parameters, size of teams, time provided, modality of response, participant incentives, demographics, cognitive abilities, communication skills, interpersonal skills, personality

INCOMMENSURABILITY LEADS TO INCONSISTENT RESULTS



INCOMMENSURABILITY LEADS TO IRRECONCILABLE RESULTS

- The one-at-a-time paradigm cannot distinguish between the following explanations for observed differences in results:
  - due to distinct subdomains being governed by different theories,
  - represent a true disagreement between competing theories that make different claims on the same subdomain,
  - indicate that one or both results are likely to be wrong and therefore require further replication and scrutiny.

- Inconsistent findings essentially become irreconcilable

THE THREE PHASES OF INTEGRATIVE EXPERIMENTAL DESIGN

- Constructing a design space
  - What are the possible variables / dimensions that need to be accounted for?
- Sampling from the design space
  - Intelligently sampling parameters from the design space when designing experiments
- Building theories from data
  - Data-driven theory that maximises out-of-sample prediction

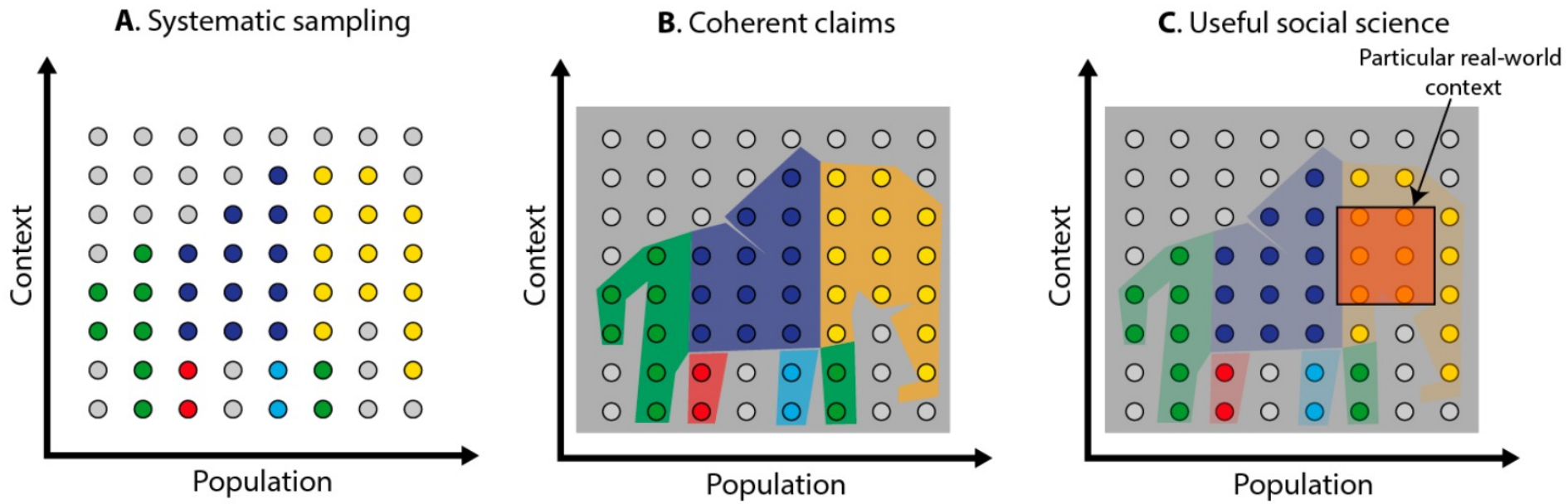
CREATING DESIGN SPACES: EXPLICITLY AND SYSTEMATICALLY FINDING DIMENSIONS

- For a given research question, use the literature and practical experience to decide on each relevant variable / dimension (there could be dozens, even hundreds)
- Construct a quantitative representation of the multi-dimensional design space and use a systematic literature review to assign well-defined coordinates to each experiment
- Iterative process that is refined over time
  - If experiments with the same coordinates yield different results, there may be a dimension missing in the design space
  - If experiments with very different coordinates yield the same results, there may be irrelevant dimensions that can be collapsed
- Potentially a new field unto itself: 'research cartography' will construct design spaces

SAMPLING DESIGN SPACES: SYSTEMATICALLY AND EFFICIENTLY 'QUERYING' THE SPACE

- Parameters for experiments should be sampled from the design space in an unbiased way that is independent of the experimental design or currently favoured hypotheses
- Where the number of dimensions in the design space (*D*) is small (e.g. <8), random sampling or even exhaustive sampling of the design space can be done e.g. via 'high-throughput' research methods or a consortium of labs
- When *D* is high efficient sampling procedures for high-dimensional spaces can be used, like Thompson sampling or active learning
- Unbiased sampling can leverage a relatively small number of sampled points in the design space to make predictions about every point in the space

SYSTEMATIC SAMPLING HELPS TIE THEORIES AND RESULTS TOGETHER



EXAMPLE OF SAMPLING A DESIGN SPACE USING ACTIVE LEARNING

- Randomly sample a few points in the design space and run the experiments
- Train a 'surrogate model' on the experiment's outcomes and generate predictions about the value of sampling unexplored points in the design space
- Follow a sampling strategy to select a new batch of experiments, ranked by value
- Iterate

BUILDING INTEGRATIVE THEORIES: MORE COMPLEXITY AND MORE BOUNDARIES

- Less emphasis on proposing new theories and more emphasis on identifying theory boundaries
- Simple theories will need to give way to more complex ones which can explain non-linear interactions and moderating effects
- New theories should be able to explain existing experimental results, predict out-of-sample experimental results, and be able to integrate new information to improve the model
- As one example, the active learning 'surrogate models' could store formal representations which both:
  - Explain previously sampled experimental results
  - Can be queried for predictions of unsampled points in the design space, generating hypotheses

INTEGRATIVE DESIGN LEAVES MORE SPACE FOR METATHEORY

- While metatheories could be used to explain differences in experimental results, very few of them currently exist
- Integrative experimental design can facilitate the creation of metatheories
- By empirically identifying distinct regions of the design space where particular theories hold, metatheories can be used to precisely state the conditions and parameter values for which different theoretically informed results should apply

INTEGRATIVE DESIGN IS PRAGMATIC

- We all know that no single intervention, no matter how evidence-based, benefits all individuals in all circumstances
- Over-generalisation from lab experiments can and often does lead to the deployment of suboptimal or even dangerous real-world interventions
- By emphasising the role of both the context and the population and more precisely identifying theoretical boundaries, it will be easier to find which interventions work when

THREE STUDIES THAT SHOWCASE ASPECTS OF INTEGRATIVE EXPERIMENTAL DESIGN

- Awad et al.'s 'Moral Machine': crowd-sourced online experiment sampling a nine-dimensional space to generate over 9 million moral dilemmas re self-driving cars (e.g. Trolley Problem)
  - The sample space led to findings that would not be deducible based on traditional experimental designs or prior research
  - Machine learning used to create a model with over 100 meaningful predictors
- Peterson et al.'s Choice Prediction Competition: automated selection of more than 100 pairs of gambles from a 12-dimensional space. Tested 'human' theories of risky choice and then fit a ML model to see which aspects of human theories worked and when
  - E.g. found prospect theory's predictive power was not uniform across the design space
- Baribault et al.'s subliminal priming metastudy: sampled 5,000 'microexperiments' from a 16-dimensional design space to reveal that a hypothesised priming effect was much more limited in generalisability than previously thought

TAKEAWAYS

- Construct the 'design space' ex-ante, explicitly, and systematically
- Report 'hidden' variables and assumptions transparently
- Communicate context and degree of generalisability not only in the methods or discussion, but in abstracts and introductions
  - "while a methods section might note that the participants were recruited from the subject pool at a particular university, it is not uncommon for research articles to report findings as if they apply to all of humanity"
- Sample the design space in an unbiased way
- Adjust expectations of what theory is: more complexity, more focus on out-of-sample prediction

ISN'T THE CRITICISM OF THE ONE-AT-A-TIME PARADIGM A STRAW MAN?

- One-at-a-time has been successful in some domains, like auction theory and mechanism design, with reconcilable results and clear theoretical predictions

- One-at-a-time will still be useful, particularly for:
  - Testing for the existence of a phenomenon (but not the conditions under which it exists)
  - Preceding the integrative framework when exploring new topics or identifying the variables that make up the design space

- One-at-a-time is not bad, it just can't do everything that is currently being asked of it

DON'T META-ANALYSES AND REVIEWS ALREADY SYNTHESISE RESULTS?

- They are post hoc
  - It can take years to wait for evidence to accumulate under the one-at-a-time approach before any attempt can be made to put them together
  - Harder to account for publication bias
- They implicitly assume existing studies are commensurate with one another, which is rarely the case
- While meta-analyses are interested in the aggregating across conditions to estimate the average effect size, the integrative framework is interested in the variation across conditions
- The integrative framework can be thought of as an ex-ante 'planned meta-analysis' that builds in commensurability by design

WHAT ABOUT 'CONCEPTUAL' OR 'MEASUREMENT' INCOMMENSURABILITY?

- Could use adversarial collaborations from scholars who have previously published from competing theoretical perspectives
  - Help reduce bias
  - Can iron out conceptual differences
- Could hold consensus conventions where existing scholars commit a priori to lists of dimensions, their operationalization, and the validity and reliability of chosen instruments
- Ultimately it will be hard and we will learn from experience – but the framework is designed to be flexible and iterable
- These problems also apply to the one-at-a-time paradigm

ISN'T THIS BLINDLY EMPIRICAL OR EVEN ANTI-THEORETICAL?

- Their vision for theory: accurately explains observed experimental results and makes accurate predictions about unseen experiments

- Typically, theories are thought to have two subgoals: providing an accurate representation of how the world works AND resonating with human interpretation
  - These often conflict in practice
  - We get inaccurate theories that are intuitive, or accurate ones that are hard to explain

- They unashamedly favour accurate predictions at the cost of a subjective sense of understanding

WHAT IF PREDICTIONS ARE INACCURATE WHEN ADOPTING THIS FRAMEWORK?

- Failures of predictions are not the fault of an integrative framework
- The design space may be mis-specified, with a crucial dimension missing
- The domain may simply have a fundamental limit to prediction, where accurate generalisation is not possible. Better to know this than not and direct resources elsewhere

SOUNDS WONDERFUL, BUT HOW COULD THIS EVER WORK IN PRACTICE?

- Recent innovations mean integrative design is becoming feasible for some. Innovations include:
  - virtual lab environments (allows systematic and automatic sampling of parameter space),
  - participant sourcing platforms (e.g. Prolific),
  - mass collaboration mechanisms (e.g. multi-lab consortiums), and
  - machine learning methods (more suited with bigger datasets, bigger design spaces)

- While more expensive for an individual group, the integrative framework should end up cheaper for science as a whole due to its more efficient approach
  - Admittedly, this will require institutional change among ethics boards, grant committees and funding arrangements

HOW COULD SMALLER LABS POSSIBLY PARTICIPATE?

- Up-front costs are high to design and run integrative meta-experiments, but the methods and infrastructure are inherently shareable

- Marginal costs are low, possible lower than one-at-a-time experiments

THE PROPOSAL IS INCOMPATIBLE WITH ACADEMIC INCENTIVE STRUCTURES

- Change is hard and it will be needed
- Promising progress in some fields:
  - in physics some of the biggest breakthroughs have come from teams with thousands of collaborating researchers
  - 'speaker lists' that prominently feature early career researchers give junior academics a chance to appear as a the 'face' of a collaboration
  - Collaboration leaders write letters of recommendation for team members
  - A 14-role research role taxonomy is gaining increasing use
- Researchers who build the infrastructure used by a collaborative effort could receive 'builder status', earning them co-author rights on subsequent publications that use that infrastructure

DISCUSSION

- What is the role of theory as machine learning methods grow in utility and popularity?
- What are the 'hidden' variables in our own research? Do we report transparently?
- Is integrative experiment design feasible for us? Is it feasible for science as a whole?