Am I Destined to be Me? Ethically approximating the predictability of human lives with artificial intelligence

Charlie Rahal, Jiani Yan, and Tom Douglas

Nuffield College/LCDS Working Group, 05.03.2025.

LEVERHULME CENTRE FOR DEMOGRAPHIC SCIENCE

- Some technical slides about a grant application in progress.
- Specifically, it's for the ESRC 'Responsive Mode' (£1m).
  - Importantly: an '**AI Highlight Notice**' is currently out for this call.
- The team is listed below:

Team Member	**Initials**	**Short Role Descriptor**	**Academic Website**
Charles Rahal	CR	Project Lead	https://crahal.com
Tom Douglas	TD	Project Co-Lead	https://sites.google.com/view/tomdouglas
Jiani Yan	JY	Project Co-Lead	https://vallerrr.github.io
Philosopher PDRA	PDRA1	Research/Innovation Associate	N/A
Data Science PDRA	PDRA2	Research/Innovation Associate	N/A
Hamza Shams	HS	Professional Enabling Staff	https://www.ndph.ox.ac.uk/team/hamza-shams

- Please do not share anything from this meeting externally.

https://doi.org/10.1038/s43588-025-00776-y

On the unknowable limits to prediction

The classic dichotomization of prediction error into statically defined 'reducible' and 'irreducible' terms is unable to capture the intricacies of progressive research design where specific types of truly reducible error are being rapidly eliminated at differential speeds. While this is generalizably true – including in the domain of large language models (LLMs) – it is none more apparent than in social systems, where questions regarding the 'predictability' of outcomes are gradually reaching the fore[1]. Canonical work that convenes common predictive tasks involving conventional social surveys[2] shows low power, but rapidly emerging computational approaches that utilize non-standard administrative data and information contained within things such as children's autobiographical essays – often combined with generative artificial intelligence and distributed computing[3,4] – begin to allow high accuracy. We propose an enhanced framework that builds upon existing work[5] in a way that we believe helpfully decomposes various types of truly reducible 'epistemic' error, which are in theory eliminable (resulting from a lack of knowledge), and isolates residual, 'aleatoric' error (inherent randomness that can never be modeled) that research endeavor will not be able to capture in the limit of human progress. We further emphasize the impossibility of claiming whether things are 'predictable', especially with regards to open, dynamically evolving systems.

Decomposing prediction error into reducible and irreducible terms is not new. The work of Trevor J. Hastie et al.[6] – a landmark instruction manual in the field – uses the term 'reducible' to taxonomically classify errors into a type that can be eliminated, and the term 'irreducible' to denote errors that cannot be eliminated, but only given current models and information on target variables and feature sets. For example, the insightful mixed-methods work of Ian Lundberg et al.[7] begins its abstract with 'Why are some life outcomes difficult to predict?'. In that case, they are difficult to predict based on information that is currently measurable, and the 'task' at hand. An outcome that is difficult to predict with one feature set may not be 'unpredictable' in an aleatoric sense: life trajectories may be trivial to predict in the future.

Reflexive terminology is essential, lest conclusions be drawn that there is no space for future positive policy interventions based on ever increasingly accurate predictions that prevent negative outcomes. It is impossible to know whether we have eliminated all reducible epistemic error to approach the practical 'ceiling' of accuracy, to make statements about how predictable outcomes will become, or to understand where we are in the process. Predictability of an outcome can only be asserted when all epistemic errors are eliminated: when 'unmeasured' information is measured and constructed perfectly, features are constructed perfectly, and when functional forms are able to be exactly recovered.

Let y_{true} denote the actual, observable outcome that is measured after constructing the phenomenon of interest; ε is aleatoric error for (that specific) y_{true}. Next, denote $\mathbf{x}_{\text{true}}$ as the perfectly chosen, constructed and measured feature set from a possibly uncountable, infinite collection of features in the universe, and $f^*()$ as the conceptually true underlying function that maps $\mathbf{x}_{\text{true}}$ onto y_{true}. As opposed to the scenario of perfect prediction described above, y_{observed}, $\mathbf{x}_{\text{observed}}$ and $f(\mathbf{x}|y)$ are defined as the observed target, features, and the model trained on them. Eq. (1) expresses our suggestion for conceptually decomposing predictive error under the assumption of no distributional shift:

$$
\begin{aligned}
y_{\text{true}} &= \underbrace{f^*(\mathbf{x}_{\text{true}})}_{\substack{\text{Predictive}\\ \text{ceiling}}} + \underbrace{\varepsilon}_{\substack{\text{Aleatoric}\\ \text{error}}} \\
&= \underbrace{[f^*(\mathbf{x}_{\text{true}}) - f(\mathbf{x}_{\text{true}}|y_{\text{true}})]}_{\substack{\text{Model approximation gain}\\ \text{(Epistemic)}}} \\
&\quad + \underbrace{[f(\mathbf{x}_{\text{true}}|y_{\text{true}}) - f(\mathbf{x}_{\text{true}}|y_{\text{observed}})]}_{\substack{\text{Measurement gain from } y\\ \text{(Epistemic)}}} \\
&\quad + \underbrace{[f(\mathbf{x}_{\text{true}}|y_{\text{observed}}) - f(\mathbf{x}_{\text{observed}}|y_{\text{observed}})]}_{\substack{\text{Measurement gain from } \mathbf{x}\\ \text{(Epistemic)}}} \\
&\quad + \underbrace{y_{\text{predicted}}}_{\text{Current prediction}} + \underbrace{\varepsilon}_{\substack{\text{Irreducible}\\ \text{(Aleatoric)}}} .
\end{aligned}
\tag{1}
$$

As $\mathbf{x}_{\text{true}}$ and y_{true} are enhanced information sets over and above $\mathbf{x}_{\text{observed}}$ and y_{observed}, they could tautologically reduce what may currently have been termed 'irreducible'. Predictive accuracy could also improve through better construct validity[8]: the extent to which the model's predictions or features accurately capture the theoretical construct or concept the model is intended to measure or represent. This may not necessarily involve the collection of additional information, but simply a re-framing of a variable or new utilization of existing information. Incorporation of previously unmeasured information (from $\mathbf{x}_{\text{observed}}$ towards $\mathbf{x}_{\text{true}}$) or better measurement of existing features (from $\mathbf{x}_{\text{observed}}$ and y_{observed} towards $\mathbf{x}_{\text{true}}$ and y_{true}) will also reduce epistemic error, as will reductions in model approximation error (from f towards f^*). See Supplementary Information for a fully delineated decomposition of Eq. (1). To illustrate how these different forms of error reduction work in practice, we depict our generalized framework in Fig. 1. This demonstrates how

nature computational science

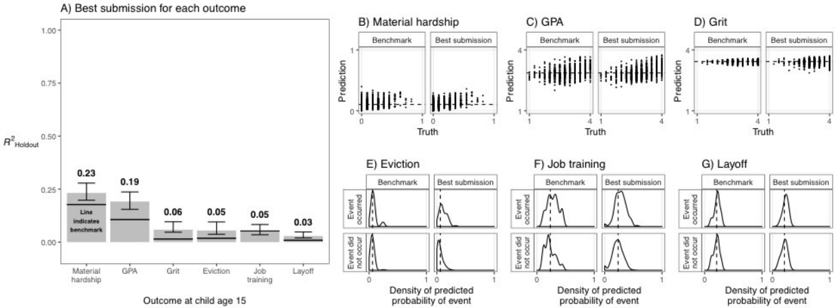
- Published on Tuesday!
- It inspires a lot of our work in this area.
- Careful delineation of prediction error.
- Major application to 'open' social systems.
- Big part of this grant (i.e., in 'Workstream Three').

"A social fact is any way of acting, whether fixed or not, capable of exerting over the individual an external constraint"

Émile Durkheim (1895), '*The Rules of Sociological Method*'.

If this notion were true in its strongest form, one would expect high predictability across social systems with minimal stochasticity in individual lives.

But, I've seen this before!



But, I've seen this before!

The FFC is a wonderful study! But:

- There are problems with the dataset.
    - This is in terms of both dimensionality and missingness.
- There are problems with the scope ('target variables').
- There are problems with external validity.
- There are problems with 'time to event'.
- There are problems with interpretation of it.
- There are problems with the undeveloped ethical aspects of it.

It re-invigorated the field. Lundberg et al. (2024) also immensely readable.

What are we proposing? Not limited to:

- Computational re-examination with largest datasets.
- Rigorous analysis of multiple dimensions and facets.
- New frameworks for normative ethics and regulation.
- Further development and integration of new AI tools.
- Development of standalone libraries.
- Decomposition of epistemic errors based on our earlier work.
- Interactive tools for engagement.
- International comparison with focus on UKRI-funded data.

What are we proposing? Three workstreams:

Stream	Output	Output Description	Panels	Census	Biobanks	Time Use	Beneficiaries
One	1	Ethics Paper	✓	✓	✓	✓	All Stakeholders
	2	Regulatory Report	✓	✓	✓	✓	Regulators
Two	3	Pre-registered Report	✓	✓	✓	✗	All Stakeholders
	4	Methods Paper	✓	✓	✓	✗	Academics
	5	Replication Work	✓	✓	✓	✗	Academics
Three	6	Autowrangler	✓	✓	✓	✓	Data Analysts
	7	Predictability Index	✓	✓	✓	✓	Public Engagement
	8	Applied Paper	✓	✓	✓	✓	All Stakeholders
Four	9	Decomposition Paper	✓	✓	✓	✗	Academics

Regulatory Framework and Ethical Risks I

- Develop regulatory frameworks via reflexive equilibria to support later analyses.
- **Risk 1: Undermining Moral Responsibility**
  - Accurate prediction may support incompatibilist determinism.
  - Assess whether it supports or appears to support incompatibilism.
  - Explore communication strategies to mitigate these risks.
- **Risk 2: Excessive Responsibility**
  - Risk of holding agents responsible for predicted future behaviour.
  - Analyse justificatory conditions for pre-emptive treatment.

Regulatory Framework and Ethical Risks II

- **Risk 3: Privacy Infringement**
  - Predictions may approximate knowledge of future behaviour.
  - Risk of infringing privacy rights through predictive knowledge.
  - Examine conditions for ethically permissible predictions.
  - Consider different moral weights of epistemic errors.
- Examine impacts on scientific practice across all risks.
- Identify safeguards and limits for deploying predictive systems.

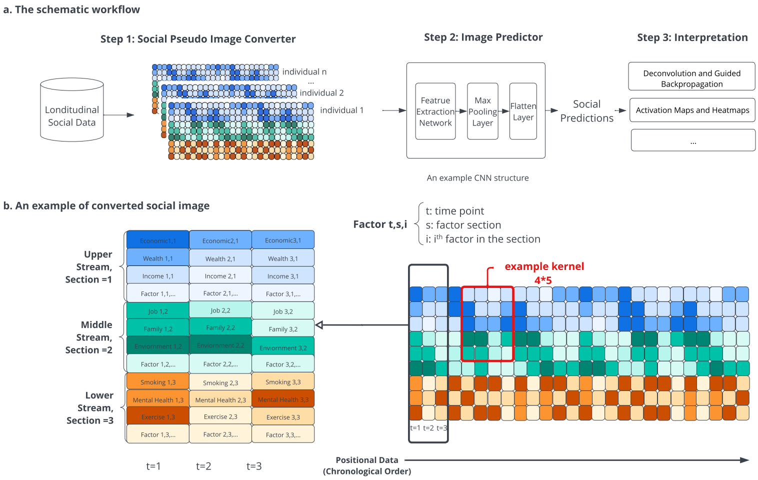
Social Data as Text, Images, and Ensembles

- Advance CNNs and LLMs for prediction in social contexts.
- Pick six intersecting variables from:
  - Panels
  - Censuses
  - Biobanks

and:

  - Socio-economics
  - Lifecourse
  - Demographics
- Pre-register a ‘domain expert’ (as per Salganik et al., 2020).
- International/multi-dataset context allows tests of external validity.
- Response to Salganik et al. (2024): ‘breaking free from the X matrix’.
- Deep model fusion based models which learn from multiple datasets.

A simple stacked variable representation



Extension to Recurrence Matrices and GAF

- For each person $i \in \{1, \ldots, N\}$ and each variable $j \in 1, \ldots, d\}$, define the individual recurrence plot (with threshold function θ) as:

$$R^{(i,j)}_{t,t'} = \theta\left(\left|x^{(j)}_{i,t} - x^{(j)}_{i,t'}\right|\right), \quad t, t' \in \{1, \ldots, T\}. \tag{1}$$

- Multidimensional Recurrence Plots (MRP) for person i are defined as:

$$MRP^{(i)}_{t,t'} = \prod_{j=1}^{d} R^{(i,j)}_{t,t'}. \tag{2}$$

- This aggregates the recurrence information across all d variables.
- We will also explore representations as Gramian Angular Fields.

In terms of social data as text:

- For prediction tasks, the key is to design a transformation:

$$g : \mathbb{R}^{n \times d} \rightarrow \mathcal{L}, \tag{3}$$

that converts tabular data $X \in \mathbb{R}^{n \times d}$ into a textual representation $\ell = g(X)$.

- Three ideas for how to enhance this.

1. Predictive Narrative Encoding. A high correlation between x_i and x_j (or theroetical relationship) that is indicative of an outcome might be encoded as:

   '*The simultaneous increase in x_i and x_j robustly signals a higher likelihood of y*'

   1.1 A further extension of this would be using an alternate (or same?) LLM to serialize the text.

2. Learn g by optimizing a differentiable surrogate for loss. Train g_θ jointly with a predictor h_ϕ by minimizing:

$$\min_{\theta,\phi} \mathbb{E}\left[L\left(h_\phi\left(g_\theta(X)\right), y\right)\right]. \tag{4}$$

Automatically discover textual constructs that best encapsulate the predictive signals in X.

Deep Model Fusion for Social Prediction

3. Assume an image-like representation:

$$\phi_{\text{img}}(X_i) \in \mathbb{R}^{H \times W \times C}, \tag{5}$$

a text embedding:

$$\phi_{\text{text}}(s_i) \in \mathbb{R}^{d_t}. \tag{6}$$

and incorporate a traditional superlearner on tabular data X for person i:

$$g_{SL}(X_i) \in \mathbb{R}^{d_{SL}}, \tag{7}$$

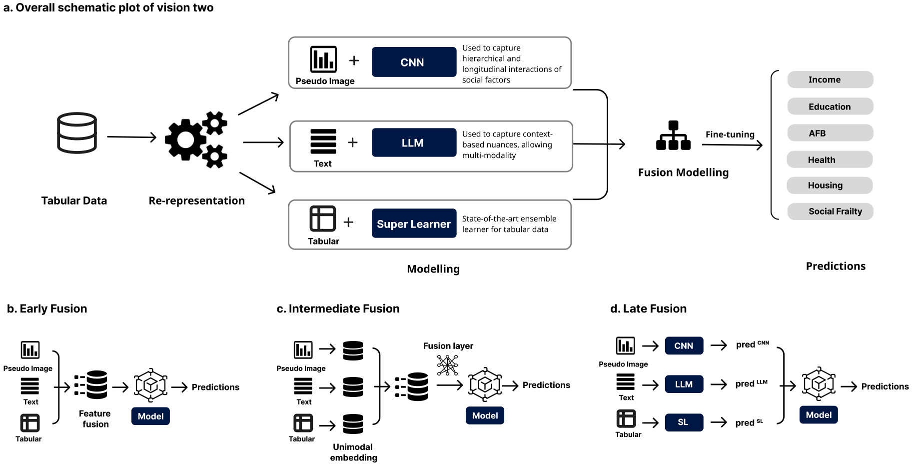
where g_{SL} is an ensemble of classical models.

- Define a fusion module f to obtain the joint representation:

$$F_i = f\Big(\text{CNN}(\phi_{\text{img}}(X_i)),\ \phi_{\text{text}}(s_i),\ g_{SL}(X_i)\Big) \in \mathbb{R}^{d_f}. \tag{8}$$

Fusion Modelling Strategies

- **Early Fusion**
  - Concatenate raw features from all modalities
  - Jointly train a single model
- **Intermediate Fusion**
  - Encode each modality separately
  - Merge embeddings via a fusion layer
- **Late Fusion**
  - Train separate models per modality
  - Combine predictions in a meta-learner


a. Overall schematic plot of vision two
Tabular Data
Re-representation
Pseudo Image
CNN
Used to capture hierarchical and longitudinal interactions of social factors
Text
LLM
Used to capture context-based nuances, allowing multi-modality
Tabular
Super Learner
State-of-the-art ensemble learner for tabular data
Modelling
Fusion Modelling
Fine-tuning
Income
Education
AFB
Health
Housing
Social Frailty
Predictions
b. Early Fusion
Pseudo Image
Text
Tabular
Feature fusion
Model
Predictions
c. Intermediate Fusion
Pseudo Image
Text
Tabular
Unimodal embedding
Fusion layer
Model
Predictions
d. Late Fusion
Pseudo Image
Text
Tabular
CNN
LLM
SL
pred CNN
pred LLM
pred SL
Model
Predictions

Indexing Social Predictability

- New algorithms for pre-processing of social data at scale.
  - Different representations amenable to new algorithms as developed in Workstream Two.
- Apply a range of algorithms to a vast set of normalized data.
- Health and genetic dimension added from UKB+CKB.
  - Allows consideration of GxE.
- Deep consideration of 'Time to Event'.

Universal Multi-Task Predictive Modeling

- Each variable X_i in a high-dimensional dataset $\mathbf{X} = \{X_1, X_2, \ldots, X_p\}$ is both a predictor/outcome.
- Learn a collection of functions

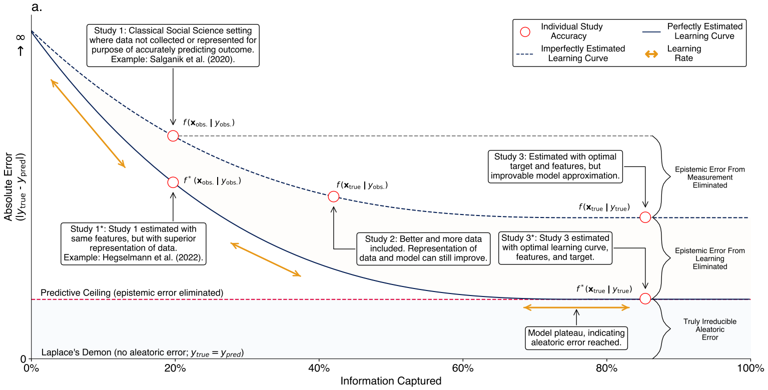
  $$f_i : \mathcal{X}_{-i} \to \mathbb{R}, \quad i = 1, \ldots, p,$$

  where $\mathcal{X}_{-i}$ denotes all variables except X_i, so that the prediction error

  $$\mathbb{E}\left[L\left(f_i(\mathbf{X}_{-i}), X_i\right)\right]$$

  is minimized for an appropriate loss function L.
- **Multi-Task Learning**/**Masked Feature Modeling** possible strategies.
- Another new idea: integrate a **DAG** as a regularizer a scale.

Decomposing Epistemic Error



1. **Learning Curve Analysis:** Incremental Information Refinement:

$$L(I) = \mathbb{E}\left[\left|y_{\text{true}} - f_I(\mathbf{x})\right|\right], \tag{9}$$

$$\Delta_\epsilon \approx L\left(I_{\text{observed}}\right) - L\left(I_{\text{enhanced}}\right). \tag{10}$$

2. Approximate Δ_ϵ from perturbations /w **first-order Taylor expansion**:

$$\Delta_\epsilon \approx \nabla_{\mathbf{x}} f(\mathbf{x}, y) \cdot \delta\mathbf{x} + \nabla_y f(\mathbf{x}, y) \cdot \delta y, \tag{11}$$

   where:

   - $\Delta_\epsilon = f(\mathbf{x} + \delta\mathbf{x},\, y + \delta y) - f(\mathbf{x}, y)$ is the change in the predictive error due to small changes in the inputs.
   - $\nabla_x f(x, y)$ and $\nabla_y f(x, y)$ are gradients (Jacobians) of f w.r.t. $\mathbf{x}$ and y.
   - δx (δy): small perturbation/measurement error in $\mathbf{x}$ (y).

3. Epistemic error can be partially captured by the uncertainty in the model parameters **via bootstrap**:

$$\mathrm{Var}\left(f(x_{\text{observed}})\right) \approx \frac{1}{M}\sum_{i=1}^{M}\left(f^{(i)}(x_{\text{observed}}) - \bar{f}(x_{\text{observed}})\right)^2, \qquad (12)$$

where $\bar{f}(x_{\text{observed}}) = \frac{1}{M}\sum_{i=1}^{M} f^{(i)}(x_{\text{observed}})$ is the mean prediction.

- This might be capturing Aleatoric Error, though, and not actually identifying epistemic error (further thinking needed).

4. **Latent Variable Modeling** can be used to “invert” the measurement error process/estimate error reduction through improved measurement:

$$x_{\text{observed}} = x_{\text{true}} + \delta x, \tag{13}$$

$$y_{\text{observed}} = y_{\text{true}} + \delta y, \tag{14}$$

Estimate the true feature set through EM algorithm or variational inference:

$$\Delta_{\text{measurement}} \approx \left\| f\left(x_{\text{observed}}\right) - f\left(\hat{x}_{\text{true}}\right) \right\|. \tag{15}$$

- This is one way to consider the effect of construct validity (of X).
- However, it assumes knowledge of the DGP.

5. **Controlled Experiments/Simulations** with Synthetic Data To assess the impact of measurement errors, we simulate synthetic data using a known true model:

$$y_{\text{true}} = f^*(x_{\text{true}}) + \epsilon, \tag{16}$$

where ϵ represents inherent aleatoric noise. We introduce measurement errors as:

$$x_{\text{obs}} = x_{\text{true}} + \delta x, \quad y_{\text{obs}} = y_{\text{true}} + \delta y, \tag{17}$$

where δx and δy are perturbations capturing imperfections in data acquisition (measure or construct).

6. Also: concentration inequalities (e.g., Hoeffding's or McDiarmid's inequality) to establish probabilistic bounds.

7. Also: differential geometric methods like the Fisher information matrix.

Thanks for your attention!

Things we'd specifically like to discuss:.

1. What can we do that goes even further beyond:
    1.1 BOLT?
    1.2 TabLLM?
2. Are our estimates of UKB/CKB RAP costs too conservative?
3. Is the choice of our six socio-economic/socio-demographic variables sufficiently interesting?