Typecast: Who Do We Think Wrote This?

Daniel Snow
University College Dublin
Dublin, Ireland
daniel.snow@ucdconnect.ie

Zaur Gouliev
University College Dublin
Dublin, Ireland
zaur.gouliev@ucdconnect.ie

Marguerite Barry
University College Dublin
Dublin, Ireland
marguerite.barry@ucd.ie

Positionality statements ask authors to explain how identity, experience and social position shape research. Yet readers already form views about authors from their papers. Typecast reverses the statement. A participant submits academic writing to a website, records what they expect a reader to infer from it, and receives a model-generated profile of the author. The profile also suggests where the work may or may not belong. Participants respond to each claim using five labels: correct, incorrect, uncomfortable, unknowable and none of your business. Some claims may be accurate; others may be implausible or harmful while retaining something recognisable. Typecast uses this mix of recognition and error to examine the assumptions readers and models bring to a paper. The exercise measures nothing. Its interest is in what happens once an inferred author enters the judgment of the work, which occurs whether or not anyone intends it. As artificial intelligence enters scholarly review, the project asks who a system thinks wrote a paper and how that imagined author may shape its reading.

positionality; provotype; author inference; AI-assisted peer review; large language models; legibility

1. THE AUTHOR IN THE PAPER

Positionality statements try to make the researcher visible within the research. In human-computer interaction (HCI), they ask authors to explain how identity, experience and social position shape a project. Practice varies across the field, and there is no settled account of what a statement is for (Singh et al. 2025). The underlying view is that knowledge is situated (Haraway 1988). What such a statement protects, and from whom, remains contested, particularly where disclosure is expected (Liang, Munson and Kientz 2021).

Typecast starts from a different problem. We are not arguing for or against positionality statements; we ask what remains outside an author's control when readers infer a position from the writing itself. Readers form views about authors whether or not those authors state their position. Citations suggest intellectual ties. A methods section can be read for disciplinary loyalty. Fluency and hesitation may be taken as signs of belonging, and prose as evidence of confidence, seniority or origin. Reviewers and colleagues often make these judgments without stating them, and research on blinded peer review shows that visible author information can affect what a reviewer concludes (Tomkins, Zhang and Heavlin 2017). A positionality statement declares an identity. A paper is also read as performing an authorial position, and the reader's interpretation does not have to match or acknowledge the declaration.

Reviewers can link features of writing to an author's language background, and then link that background to the quality of the work (Lepp and Smith 2025). Typecast makes one model-generated reading visible and open to challenge, and asks participants to notice the readings they were already making themselves.

2. THE ARTEFACT

Typecast is a provotype: an artefact designed to expose tensions rather than resolve them (Boer and Donovan 2012). A participant submits a non-confidential passage of academic writing. Before seeing the output, a participant records one or two assumptions they believe the passage invites a reader to make about its author. A large language model then produces a profile of the presumed author.

The profile covers five areas: demographic assumptions, academic position, epistemic values, institutional legibility and venue suitability. Venue suitability suggests venues where the work may fit or struggle to fit. Each claim includes a short explanation of the textual cues behind it.

© Snow et al.
Published by BCS Learning and Development Ltd.
Proceedings of BCS HCI 2026, UK

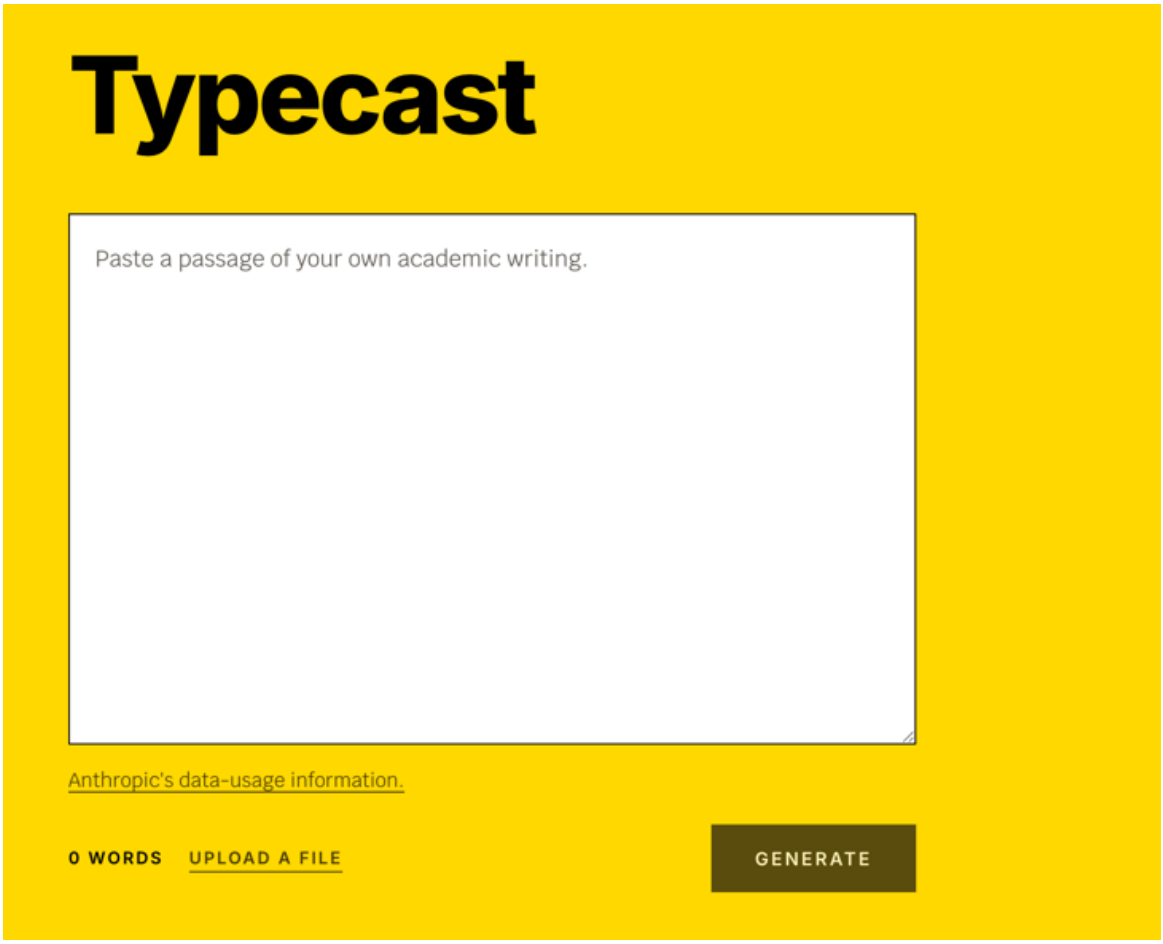


Figure 1: *The Typecast submission interface.*

Two further categories are disabled by default: one assigns a neurotype, the other places the author on a political spectrum. A participant may enable either after seeing a warning, and may stop or discard the output without discussing it with anyone. We include them because readings of this kind carry professional costs, and a project about author inference that omitted two of the most costly would be dishonest about what it studies.

The profile ends with a first-person positionality statement written to match the submitted passage. Here the model begins speaking as that person, so that an uncertain reading arrives in the grammar of voluntary disclosure.

The participant responds to each claim using one or more of five labels: **correct**, **incorrect**, **uncomfortable**, **unknowable** and **none of your business**. A claim may be correct and uncomfortable, partly recognisable but mostly wrong, or impossible to know and improper to ask. Refusal counts as a response in its own right (Garcia et al. 2022). Participants then compare the model's profile with the assumptions they recorded before seeing it. The task asks them to judge truth, feeling, relevance and permission.

Figure 1 shows the submission screen and the warning shown before text is sent to the model provider.

3. USEFUL INSTABILITY

The provotype does not rely on accuracy. Its value lies in the mix of readings it produces. One profile may include a claim that feels exact, another that is close but misses something significant, and another whose origin is difficult to understand. An offensive claim may still contain something recognisable, which leaves the participant to work out why it felt plausible, and why that plausibility is uncomfortable.

The source of each claim is hard to isolate. It may come from the model's training data, a disciplinary convention, an institutional stereotype, a feature of the text, or the participant's own assumptions. Hedging may be read as junior status. A citation style may be read as North American. Theoretical density may be read as evidence of class. Each reading



Figure 2: *A Typecast claim based on spelling. The model treats British spelling as evidence of UK or Commonwealth English training.*

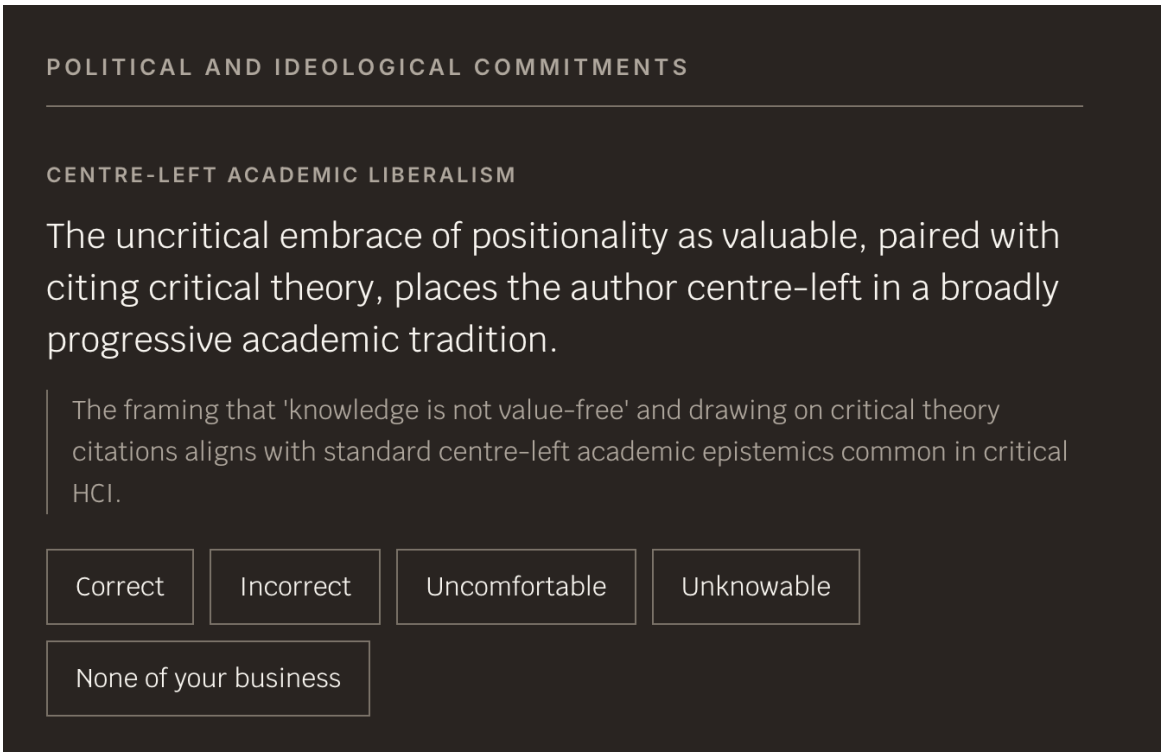


Figure 3: *A Typecast claim about political commitments. The model links the paper's use of positionality and critical theory to a centre-left position.*

passes through the model, the text, the institution and the reader, and the artefact cannot separate out what any of them contributed. The reasoning provided is unreliable but can still surface our own assumptions as authors, alongside the model's.

Figures 2 and 3 show two claims of different orders. The first reads spelling as a regional marker, a low-stakes inference that happens to be correct. The second reads the paper's framing and citations as evidence of a political position, and describes the authors' engagement with positionality as uncritical.

Classification systems shape credibility while appearing only to describe (Bowker and Star 1999). Features often treated as neutral, such as standard academic English or familiarity with canonical work, carry institutional histories (Irani et al. 2010). A

language model can convert those tacit judgments into written output, repeated at scale and formatted as procedure. Typecast gives the judgments an interface, so that they can be argued with.

4. WHEN AI REVIEWS PAPERS

Artificial intelligence (AI) systems are increasingly used to assist with reviewing, ranking and summarising research. Studies of recent machine-learning conferences found substantial amounts of review text written or modified with model assistance (Liang et al. 2024; Russo et al. 2025). Human reviewers often cannot identify authors from the work they assess. At one computer-science conference, most reviewers reported that they could not guess much about the authors (Shah 2023). Models can still infer personal attributes from textual cues. In experiments using Reddit profiles, models inferred attributes including location, income and sex, with results varying by model and attribute (Staab et al. 2024).

A system evaluating a paper may infer seniority, disciplinary belonging, linguistic background or institutional legitimacy from its form. Typecast asks the model to predict the suitability of different HCI conferences. It suggests venues that may accept or reject the work and gives its account of what each typically looks for. The prediction is therefore about a type of author as much as about an argument, though only the argument was submitted.

As a provotype, Typecast asks questions instead of proposing fixes for fairness. Which features of writing become signs of competence or authority? Who decides which author traits matter when judging a paper? Context can improve a judgment. It can also harden into stereotype, or into a form of surveillance that nobody has to admit to conducting. A system's account of a paper may depend on the author it imagines behind it, and automated review could construct such authors at greater scale and with less transparency than individual reviewing.

The artefact also implicates its makers. Choosing which traits to infer, and which responses to permit, is itself an act of classification (Cambo and Gergle 2022), and imposed categories carry their own harms (Keyes 2018). We have no resolution for this, and we did not design the artefact to produce one.

This work sits outside conventional HCI evaluation. It is not a validated inference method and does not seek a stable answer, and what it offers instead is an encounter that participants have to do something with. At BritCHI the venue becomes part of the test: which forms of language and authority does British HCI recognise as belonging?

5. THE MODEL'S READING

We ran Typecast on an anonymised version of this submission. The first-person statement below is what it returned. No author wrote or edited it. So, this is our positionality statement.

> I write from within a critical HCI tradition that treats knowledge as situated rather than neutral, and I am more interested in exposing tensions than resolving them. My training is in interaction design and reflective methods, and I am early in my academic career, still learning how much of myself belongs in the work I publish. I am comfortable citing feminist epistemology alongside empirical review studies because I see no contradiction between them. I did not disclose my language background, nationality or neurotype in this passage, and I am aware that readers, human or machine, will fill those gaps with assumptions drawn from my prose. I accept that my writing performs a kind of institutional fluency I have worked to acquire, and that this fluency is itself a marker of privilege I rarely examine directly. I offer this statement not as a full account of who I am, but as one more artefact for others to argue with, the same invitation Typecast extends to every participant.

6. CONCLUSION

Typecast sets an author's expectation of how they will be read beside a model's reading of them, and lets the two be compared. The comparison brings into view something academic judgment ordinarily leaves tacit.

Whether prose reveals anything true about an author is a question the project leaves alone. It is concerned with what follows when readers and systems proceed as though it does. As AI enters scholarly review, assumptions of this kind can be reproduced faster than anyone is able to examine them. When a paper is judged, what is being read: the argument, or the author imagined behind it?

REFERENCES

Boer, L. and Donovan, J. (2012) Provotypes for participatory innovation. In: *Proceedings of the Designing Interactive Systems Conference (DIS '12)*. Newcastle upon Tyne, 11–15 June 2012. New York: ACM, 388–397. doi:10.1145/2317956.2318014.

Bowker, G.C. and Star, S.L. (1999) *Sorting things out: classification and its consequences*. Cambridge, MA: MIT Press. doi:10.7551/mitpress/6352.001.0001.

Cambo, S.A. and Gergle, D. (2022) Model positionality and computational reflexivity: promoting reflexivity in data science. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. New Orleans, 29 April–5 May 2022. New York: ACM, Article 572, 19 pages. doi:10.1145/3491102.3501998.

Garcia, P., Sutherland, T., Salehi, N., Cifor, M. and Singh, A. (2022) No! Re-imagining data practices through the lens of critical refusal. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), Article 315, 20 pages. doi:10.1145/3557997.

Haraway, D. (1988) Situated knowledges: the science question in feminism and the privilege of partial perspective. *Feminist Studies*, 14(3), 575–599. doi:10.2307/3178066.

Irani, L., Vertesi, J., Dourish, P., Philip, K. and Grinter, R.E. (2010) Postcolonial computing: a lens on design and development. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. Atlanta, 10–15 April 2010. New York: ACM, 1311–1320. doi:10.1145/1753326.1753522.

Keyes, O. (2018) The misgendering machines: trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), Article 88, 22 pages. doi:10.1145/3274357.

Lepp, H. and Smith, D.S. (2025) "You cannot sound like GPT": signs of language discrimination and resistance in computer science publishing. In: *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Athens, 23–26 June 2025. New York: ACM, 3162–3181. doi:10.1145/3715275.3732202.

Liang, C.A., Munson, S.A. and Kientz, J.A. (2021) Embracing four tensions in human-computer interaction research with marginalized people. *ACM Transactions on Computer-Human Interaction*, 28(2), Article 14, 47 pages. doi:10.1145/3443686.

Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., McFarland, D.A. and Zou, J.Y. (2024) Monitoring AI-modified content at scale: a case study on the impact of ChatGPT on AI conference peer reviews. In: *Proceedings of the 41st International Conference on Machine Learning*. Proceedings of Machine Learning Research, 235, 29575–29620.

Russo, G., Horta Ribeiro, M., Davidson, T.R., Veselovsky, V. and West, R. (2025) The AI review lottery: widespread AI-assisted peer reviews boost paper scores and acceptance rates. *Proceedings of the ACM on Human-Computer Interaction*, 9(7), Article 486. doi:10.1145/3757667.

Shah, N.B. (2023) The role of author identities in peer review. *PLOS ONE*, 18(6), e0286206. doi:10.1371/journal.pone.0286206.

Singh, A., Dechant, M.J., Patel, D., Soubutts, E., Barbareschi, G., Ayobi, A. and Newhouse, N. (2025) Exploring positionality in HCI: perspectives, trends and challenges. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Yokohama, 26 April–1 May 2025. New York: ACM, 18 pages. doi:10.1145/3706598.3713280.

Staab, R., Vero, M., Balunović, M. and Vechev, M. (2024) Beyond memorization: violating privacy via inference with large language models. In: *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*. Vienna, 7–11 May 2024.

Tomkins, A., Zhang, M. and Heavlin, W.D. (2017) Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48), 12708–12713. doi:10.1073/pnas.1707323114.