

The Need for Assessment of the Cascade Effects from Sequencing AI Systems in Policing

Evdoxia Taka
University of Glasgow
Glasgow, UK
evdoxia.taka@glasgow.ac.uk

Temitope Lawal
University of Northumbria
Newcastle, UK
temi.lawal@northumbria.ac.uk

Michele Sevegnani
University of Glasgow
Glasgow, UK
michele.sevegnani@glasgow.ac.uk

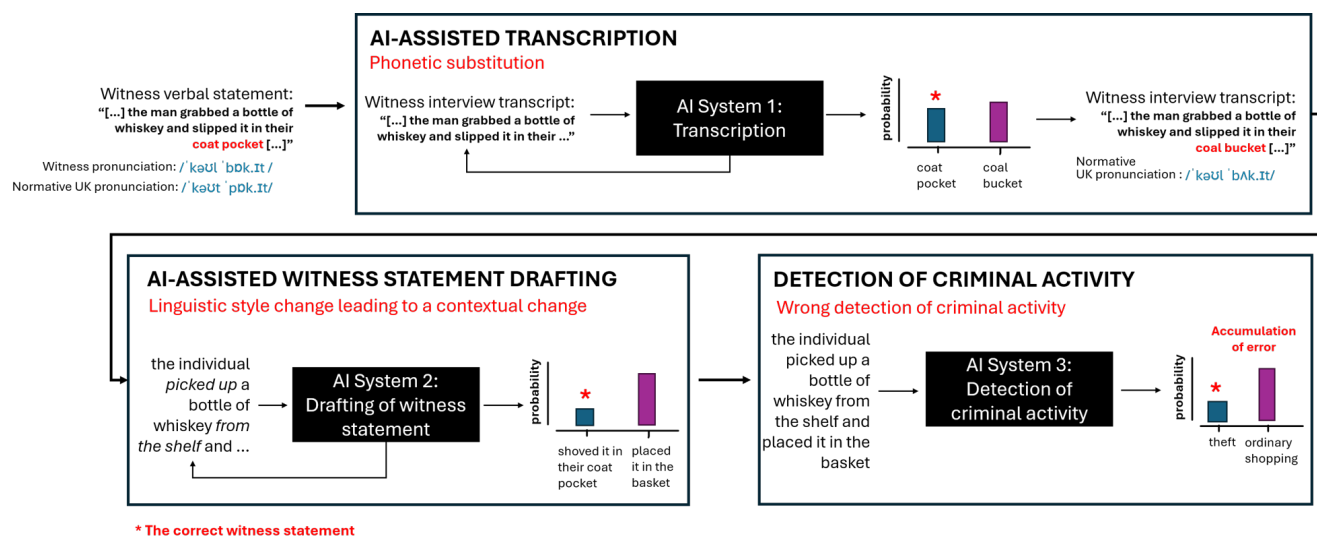


Figure 1: Schematic illustration of error propagation from sequential AI operations in policing when a non-normative speaker is involved.

Abstract

This position paper examines the use of multimodal AI in policing in England and Wales focusing on the risks that arise when the output of one system becomes the input to another. Drawing on 25 stakeholder interviews across the criminal justice system, we find that outputs from AI systems used for transcription, classification, summarisation, or information retrieval are used as inputs to downstream tools. We argue that this sequencing creates a distinct inclusion and governance problem: errors and biases introduced at early stages may be propagated, amplified, or transformed downstream, with particular risks for people whose speech, language, behaviour, accent, disability-related communication patterns, or distress diverge from normative assumptions embedded in training and evaluation data. Even minor inaccuracies, such as transcription errors, may distort evidential narratives and affect legal outcomes. Using examples from AI-assisted transcription and witness statement drafting, we identify the need for chain-level assurance methods that make provenance, uncertainty, and error

propagation visible to practitioners. We conclude by proposing a research agenda for inclusive assessment of chained multimodal AI in high-stakes public-sector settings.

CCS Concepts

• **Computing methodologies** → Artificial intelligence; • **Social and professional topics** → Governmental regulations; • **Security and privacy** → Privacy protections; Social aspects of security and privacy; • **Applied computing** → Law.

Keywords

Multimodal AI; Inclusive AI; Policing; Cascade effects; Error propagation; AI governance

ACM Reference Format:

Evdoxia Taka, Temitope Lawal, and Michele Sevegnani. 2026. The Need for Assessment of the Cascade Effects from Sequencing AI Systems in Policing. In *Companion of the INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI Companion '26)*, October 05–09, 2026, Napoli, Italy. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3776591.3833843>

1 The problem and our position

Policing and the criminal justice domain more broadly, is a rapidly evolving landscape for the experimentation with and deployment of artificial intelligence (AI) systems. Established technologies, including facial recognition, risk assessment tools, and predictive policing



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
ICMI Companion '26, Napoli, Italy
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2319-3/2026/10
<https://doi.org/10.1145/3776591.3833843>

systems, are now being complemented by applications that support operational activity, case management, criminal investigations, and interactions with the public [8, 10]. The role of multimodal AI is central to this shift because it mediates interactions between members of the public, police officers, legal professionals, and institutional databases.

Adoption of such systems is often driven by legitimate operational pressures, in particular the potential to save officer time and process large volumes of material [3, 8, 10, 11, 15]. These tools also create less visible inclusion risks [3, 10, 15]: the people affected by these systems are not only professional users, but also victims, witnesses, suspects, callers, and communities whose speech, language, behaviour, accent, disability-related communication patterns, or distress may diverge from the assumptions encoded in AI models. Existing research makes differential performance a concrete, testable concern.

Koenecke et al. [5] reported materially higher error rates for Black compared to White speakers across five commercial automatic speech recognition (ASR) systems. Tatman [13] identified differences by gender and dialect, including lower caption accuracy for female and Scottish speakers. Li and Wu [7] documented performance disparities for Mandarin stuttered speech. Most directly, Ezema et al. [2] found that lower ASR accuracy for Black tutors affected downstream automated feedback. While these studies arise outside policing in England and Wales, they do show that upstream performance differences and their downstream effects are empirically plausible and should be tested.

In this context, we conducted a study to map the current and potential future AI landscape in policing in England and Wales [6]. It involved 25 semi-structured interviews, conducted between May 2025 and January 2026, with professionals across the criminal justice system, spanning policing, the Bar, government and oversight bodies, the research and policy community, and commercial technology providers collaborating with law enforcement. Respondents were recruited through purposive sampling, supplemented by snowball sampling to broaden the range of professional perspectives (i.e. participants were invited to refer additional qualified experts). We interviewed 10 police officers (across 7 police forces), 4 legal practitioners, 2 industry practitioners, 8 regulators/representatives of oversight bodies, and 1 academic. The analysis of transcripts was conducted according to a hybrid coding framework that combined deductive codes from the research questions with inductive codes developed from the data [4].

From these interviews, we found that many of the AI systems at various stages of deployment — experimental, pilot, or live — across policing and the broader criminal justice process are multimodal. Central to them are transformer-based architectures and large language models (LLMs), which frequently interact with other modalities such as audio, speech, image, and video data. These systems are routinely deployed in sequence with other unimodal or multimodal systems because one component must typically convert material from one form into another before a subsequent component can process it. In such pipelines, errors and embedded biases can interact, producing effects on accuracy, accountability, and legal compliance that are not visible in single-system evaluations.

We use the terms ‘sequencing’ and ‘chaining’ interchangeably to describe an arrangement in which the output of one system becomes

the input of another, and we use ‘cascade’ or ‘compounding’ effects to describe the consequences that follow from it.

In this setting, the relevant users and affected parties include both system operators and people represented within system inputs and outputs. A non-emergency caller, a witness giving an account, a suspect appearing in video, or a victim whose statement is summarised may all be affected by multi-modal and -stage processing without being the formal system user. This is central to the inclusion problem: exclusion or misrecognition may arise not only at the interface used by officers, but also in the automated interpretation of speech, behaviour, affect, language, and contextual cues.

The interviews highlighted two specific inclusion-relevant examples that show how system design can redistribute both access and risk. A respondent discussing AI-assisted post-call analysis reported difficulties with road and proper names and noted that equivalent Welsh capability was required but difficult to obtain (*Interview 20*). Other respondents observed that AI-assisted statement drafting helped dyslexic officers with composition, spelling, and grammar (*Interviews 2, 9*).

Compounding errors arising from the sequencing of AI systems could heighten the risk to affected parties. Chaining of AI systems was discussed in 20 of these interviews. Respondents highlighted several specific risks associated with chaining.

First, inaccuracies introduced at an early stage in a processing pipeline are not confined to that stage; rather, they are propagated forward and may be amplified at each subsequent step. As one respondent noted in relation to AI transcription, “*if we don’t illuminate [accuracy concerns] and we just assume the transcription is accurate, then potentially everything else could become biased that follows*” (*Interview 20*). Second, where systems are trained on the outputs of other systems without a sufficiently broad base of original data, deficiencies in the first tool may be inherited and magnified downstream. One respondent called this “*doubling your risk*” (*Interview 24*). Risk, in this sense, is not merely transferred between systems but actively transformed through their interaction.

These compounding effects may become increasingly significant as information progresses through the evidential chain within the criminal justice system [1], from policing activities (such as community policing, intelligence gathering, crime investigation, and case management), through prosecution (where charging decisions depend on the admissibility and reliability of evidence), to trial and sentencing, and finally correctional processes (including prison and probation supervision). An error introduced at an early stage may therefore have far-reaching consequences across the whole of this process.

Although sequential AI processing is already occurring, awareness of the concept among respondents was uneven (Fig. 2). Some reported it already embedded, although still limited, in operational tools (*Interviews 4, 13, 18, 19, 20*). A larger group treated it as a near-term prospect (*Interviews 5, 6, 8, 9, 11, 12, 16, 17, 21, 22, 23, 24*), and others had not previously encountered the term (*Interviews 1, 2, 7*).

We read this unevenness as indicative of the limited extent to which current governance frameworks address systemic interactions across multiple AI models and modalities. As these technologies become more autonomous and agentic — able to execute sequences of tasks without human validation at each stage — chained

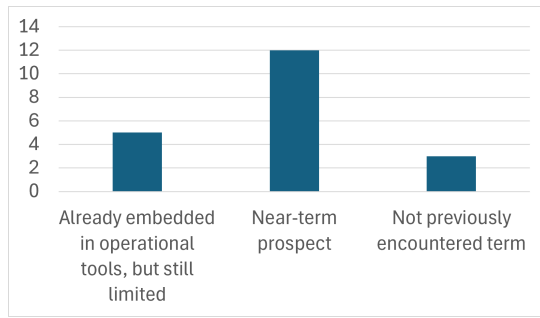


Figure 2: Respondents' awareness of AI "chaining". Most respondents acknowledged this as a near-term prospect.

systems are likely to become more prevalent. This development raises a critical inclusion and governance challenge. Specifically, effective frameworks for the assurance and oversight of chained AI systems must be established proactively to protect the inclusivity of AI in this sensitive domain. Retrospective governance, introduced after widespread deployment, risks significant loss of oversight and control. Accordingly, there is an urgent need to develop governance mechanisms in advance, while there remains an opportunity to shape the design, constraints, and accountability structures governing these systems.

2 Examples of multimodal and sequential AI processing in policing

The need for such anticipatory governance is most easily grasped from the deployment of multimodal processing systems. Examples identified in our study include the following:

- **Witness statement drafting:** a system transcribes recorded witness interviews, and the resulting transcript is subsequently used as input for an AI-assisted tool that generates formal witness statements.
- **Audio redaction in video recordings:** the system identifies instances of Personally Identifiable Information (PII) through analysis of the transcript. Based on these detections, an officer may select specific segments of the audio track for redaction. The system also allows for manual specification of time ranges to enable more precise redaction.
- **Object detection in video recordings:** an officer provides a search query in the form of a textual prompt, which the system uses to locate and retrieve instances of specified objects within the video footage.
- **AI-assisted handling of non-emergency calls:** the system transcribes incoming calls and applies analytical models to identify indicators of vulnerability by conducting assessments of threat, harm, risk, and vulnerability.
- **Non-emergency call triage:** the system transcribes incoming calls and analyses key terms and patterns to determine whether the caller should be connected to a call handler or redirected to alternative forms of assistance.

These examples involve the sequential operation of individual multimodal or unimodal AI models. This sequential processing

takes two broad forms. The first is chaining within a single product or vendor pipeline, in which steps such as transcription, summarisation, and generation are linked internally making the tool's composite nature often invisible to its users (*Interview 19*). The second is chaining across separately procured systems, for instance, where the output of an AI statement-drafting tool becomes the disclosure material handled by a separate AI redaction tool, a configuration one respondent described as "*an AI for the AI*" (*Interview 4*). Respondents reported that both forms already occur in practice.

One example is a post-call analysis tool, being trialled by two police forces in England and Wales, that produces two outputs: one designed "*for a human to read*" and another "*optimised for further technological processing*" (*Interview 20*). The latter output is then passed into a further stage of the system which: "*identifies elements within that call that are capable of being searched for on the police database [...] automatically searches all of the other police databases that we have and brings all the information back for the human to review*", before proceeding to draft an assessment of the threat, harm, risk and vulnerability (*Interview 20*).

This sequence — from audio to transcription to categorisation to automated database search, and finally to risk assessment — is an operational chain in which several components, many of them AI-enabled, interact in series. Each stage contributes to a composite output that will inform decisions about individuals, and responsibility, interpretability, and the source of any error become difficult to isolate at any single point. A particular concern is whether the automated nature of this sequence is disclosed in the resulting assessment. If the final output—in this case, a threat, harm, risk, and vulnerability assessment—is passed to a decision-maker without any indication that it was generated through a chain of automated processes, that decision-maker has no basis on which to apply the appropriate degree of scrutiny or to understand the provenance of the information on which they are being asked to act.

A further example involved integrating LLM processing with national policing databases. One respondent described a system combining "*classification, the ability to transcribe data and then using a large language model to pick out specific elements of that data that might indicate higher threat, harm or risk*", whose outputs feed into data pipelines connected to the Police National Computer and the Police National Database (*Interview 19*). Respondents also envisaged the emergence of more explicitly agentic forensic tools capable of invoking several analytical functions in sequence and passing outputs between them without human intervention at intermediate points (*Interview 16*).

3 Propagation of error: a use case

Fig. 1 is a schematic illustration of error propagation arising from sequential integration of multiple AI systems in the drafting of a witness statement and the subsequent inference of whether criminal activity has occurred in the case of a non-normative witness speaker. The scenario follows an example reported by a respondent, in which an LLM altered the context of an account concerning the alleged theft of a bottle of whiskey. A statement that a person "*grabbed a bottle of whiskey and shoved it in their coat pocket*" was rendered as "*the bottle of whiskey was picked up from the shelf and placed in the basket*" (*Interview 22*). This distinction is legally significant: the first

description may indicate theft, while the second describes ordinary shopping.

Although the respondent did not make a relevant statement, we assume that, in this scenario, the mispronunciation of a non-normative witness speaker initiates a cascade of downstream errors. This example illustrates why inclusive evaluation cannot be confined to aggregate accuracy. A transcription or generation system may perform acceptably on benchmark data while failing for speakers with particular accents, speech impairments, atypical prosody, emotional distress, code-switching, or non-standard phrasing. When such outputs are normalised by a downstream LLM, the resulting statement may become more fluent but less faithful.

This example shows how cumulative effects by the deployment of AI in subsequent processing stages of an atypical speaker's statement — such as phonetic substitutions by an automatic speech recognition system followed by a stylistic and contextual shift introduced by a LLM — can distort an evidential narrative and impact case progression by upgrading or downgrading the severity of a case relative to the original account [1]. When AI mis-transcribes the verbal statement of an atypical speaker, which in turn can contribute to AI smoothing text into fluent, well-formed prose, this “chain” of alterations can quietly change the meaning, and in evidential terms that change is not neutral. In the absence of adequate human oversight and verification, these distortions may result in the misinterpretation of key facts, potentially leading to a failure to detect criminal behaviour or, conversely, to an erroneous conclusion that it has occurred.

4 Future challenges and directions

Error propagation in cascaded speech and language systems is recognised in existing research [2, 14], and prior work has documented performance disparities in ASR [5, 7, 13]. Our narrower claim is that chain-level assurances remain underdeveloped in the AI governance literature, and to our knowledge, its conceptualisation as a distinct governance challenge in law enforcement is a novel contribution of this project.

In practice, responsibility for interpreting the outputs of interconnected AI systems is often shifted onto end users who lack specialised AI knowledge [8] or are not appropriately equipped [3, 15]. Consequently, a key priority for governance frameworks is to make interactions among multimodal AI components, and their implications, transparent and explainable to those without a technical background. We contend that this objective can be advanced through the systematic evaluation and accessible explanation of system robustness and the risk of error propagation. Robustness assessments can draw on probabilistic methods to quantify uncertainty and to model how variation in the inputs of an upstream component affects the variability of downstream outputs. A priority for research is to establish how such methods can be applied within sequential AI operations in order to assess the cascading effects on accuracy and bias.

We identified five requirements to achieve this: 1) a provenance representation should record each component, model, and prompt, intermediate input and output and associated uncertainty, and human intervention; 2) evaluation should test each component and the chain end-to-end, including perturbation tests that trace how

upstream variation changes downstream outcomes; 3) performance should be disaggregated across communication characteristics relevant to the use case, with affected communities involved in defining test conditions and harms; 4) metrics should distinguish lexical error from semantic and legally consequential error; and 5) assessments should identify points for abstention, correction, and human review.

In our prior work, we developed two instruments that could support such assessment before deployment. First, we constructed a map of the AI tool ecosystem within criminal justice across England and Wales and introduced a taxonomy to classify these tools by their information inference mode and modes of input and output [12]. This classification can help identify where chaining occurs. Second, we developed a responsible AI checklist for policing applications [9] to help practitioners involved in the deployment of AI tools recognise instances of chaining and assess the associated risks. A key gap we identify is the lack of an explainable framework capable of rendering system interactions transparent and comprehensible to end-users, explaining the possible mechanisms through which errors may propagate across interconnected systems, and providing robust estimates of error-propagation risk within a chain.

Through this workshop, we aim to share our findings on the challenges of sequencing AI systems in criminal justice, to discuss emerging research directions on cascade effects in multimodal AI, and to build collaboration across the field. In this context, we seek to advance discussion on the following research questions:

- What should a comprehensive framework for assessing cascade effects from the sequencing of AI systems in policing include?
- Which methods and metrics are best suited to quantifying the risk of error accumulation across such systems?
- How can these risks be effectively communicated to system operators, particularly those without technical expertise?
- What guidance should be provided to practitioners to support responsible and informed use of these technologies, and how should affected communities participate in defining test conditions, harms, and acceptable thresholds?

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/Y009800/1], through funding from Responsible Ai UK (KP0003).

Safe and Responsible Innovation Statement

The study received ethical approval from the Ethics Committee of Northumbria University. Participation was based on informed consent, all quotations are anonymised, and respondents were drawn from a range of professional roles. We acknowledge the limitation that affected communities (victims, witnesses, people with disabilities) were absent from the study,

References

- [1] Muffy Calder, Marion Oswald, Elizabeth McClory-Tiarks, Michele Sevegnani, and Evdokia Taka. 2026. Responsible AI in criminal justice: LLMs in policing and risks to case progression. arXiv:2603.18116 [cs.CY]
- [2] Kelechi Ezema, Chelsea Chandler, Rosy Southwell, Niranjan Cholendiran, and Sidney D'Mello. 2025. “It feels like we’re not meeting the criteria”: Examining and Mitigating the Cascading Effects of Bias in Automatic Speech Recognition in Spoken Language Interfaces.. In *Proceedings of the 2025 CHI Conference on Human*

- Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 109, 13 pages. doi:10.1145/3706598.3714059
- [3] Shai Farber. 2026. Machines of justice: A systematic review of AI applications in policing and criminal justice. *The Police Journal: Theory, Practice and Principles* (April 2026). doi:10.1177/0032258X261439572
 - [4] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods* 5, 1 (2006), 80–92. doi:10.1177/160940690600500107
 - [5] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America* 117, 14 (2020), 7684–7689. doi:10.1073/pnas.1915768117
 - [6] Temitope Lawal, Kyriakos Kotsoglou, Evdoxia Taka, Marion Oswald, Michele Sevegnani, and Muffy Calder. 2026. The Use of Probabilistic AI in Law Enforcement and the Criminal Justice System of England and Wales. *Northumbria University* (2026), 1–88. doi:10.25398/rd.northumbria.32769762
 - [7] Qisheng Li and Shaomei Wu. 2024. Towards Fair and Inclusive Speech Recognition for Stuttering: Community-led Chinese Stuttered Speech Dataset Creation and Benchmarking. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI EA '24). Association for Computing Machinery, New York, NY, USA, Article 331, 9 pages. doi:10.1145/3613905.3650950
 - [8] Rick Muir and Felicity O'Connell. 2025. *Policing and Artificial Intelligence*. Technical Report. The Police Foundation, London, UK. <https://www.police-foundation.org.uk/wp-content/uploads/2010/10/policing-and-ai.pdf.pdf>
 - [9] Marion Oswald, Muffy Calder, Claire Paterson-Young, and Jennifer Dunkwu. 2025. Responsible AI Checklist for Policing. https://research.northumbria.ac.uk/probablefutures/?page_id=636
 - [10] Benjamin Perrin, Geoffrey Liew, Isabelle Sweeney, and Peter A. Allard. 2026. *AI & Policing: Research Report on the Governance & Use of Artificial Intelligence by Police in Canada*. Technical Report CanLIIDocs 293. UBC AI & Criminal Justice Initiative. <https://canlii.ca/t/7nrqc>
 - [11] Monika Simmler, Simone Brunner, Giulia Canova, and Kuno Schedler. 2023. Smart criminal justice: exploring the use of algorithms in the Swiss criminal justice system. *Artif Intell Law* 31 (2023), 213–237. doi:10.1007/s10506-022-09310-1
 - [12] Evdoxia Taka, Temitope Lawal, Muffy Calder, Michele Sevegnani, Kyriakos Kotsoglou, Elizabeth McClory-Tiarks, and Marion Oswald. 2025. Mapping the Probabilistic AI Ecosystem in Criminal Justice in England and Wales. arXiv:2512.04116 [cs.CY]
 - [13] Rachael Tatman. 2017. Gender and Dialect Bias in YouTube's Automatic Captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, Dirk Hovy, Shannon Spruit, Margaret Mitchell, Emily M. Bender, Michael Strube, and Hanna Wallach (Eds.). Association for Computational Linguistics, Valencia, Spain, 53–59. doi:10.18653/v1/W17-1606
 - [14] Zhen Wan, Chao-Han Huck Yang, Yahan Yu, Jinchuan Tian, Sheng Li, Ke Hu, Zhehuai Chen, Shinji Watanabe, Fei Cheng, Chenhui Chu, and Sadao Kurohashi. 2025. SpeechIQ: Speech-Agentive Intelligence Quotient Across Cognitive Levels in Voice Understanding by Large Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 30381–30398. doi:10.18653/v1/2025.acl-long.1466
 - [15] Miri Zilka, Holli Sargeant, and Adrian Weller. 2022. Transparency, Governance and Regulation of Algorithmic Tools Deployed in the Criminal Justice System: a UK Case Study. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) (AI/ES '22). Association for Computing Machinery, New York, NY, USA, 880–889. doi:10.1145/3514094.3534200