Who Is Accountable When the Agent Speaks? Toward Accountability Graphs for Conversational AI Chains

Mehmet Haklidir
Artificial Intelligence Institute
TUBITAK BILGEM
Kocaeli, Turkiye
mehmet.haklidir@tubitak.gov.tr

Abstract

Conversational user interfaces (CUIs) are shifting from single-model dialogue systems to multi-agent architectures where an orchestrating agent delegates tasks to sub-agents, external APIs, and third-party tools. When these agent chains produce harmful or misleading outputs, a governance question follows: *who is accountable?* Current approaches blame either the visible interface or the system developer, but neither captures how responsibility is distributed across a chain. We identify three research threads (structural ecosystem mapping, formal causal responsibility theory, and operational provenance tracking) that have developed in isolation and have not been unified for conversational AI. We propose *Accountability Graphs*, a framework that models agent chains as directed acyclic graphs enriched with provenance metadata and causal responsibility semantics. Motivated by *Moffatt v. Air Canada* (2024), where an airline was held liable for its chatbot's misrepresentation of bereavement fare policy, we extend this case into a multi-agent architecture and show how backward responsibility propagation can assign graded responsibility scores to chain components. We discuss CUI-specific challenges including real-time dialogue, dynamic delegation, modality effects, and the design of user-facing *accountability surfaces*, and propose a research agenda connecting accountability to consent in agentic AI.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**; • **Computing methodologies** → **Intelligent agents**; • **Software and its engineering** → *Software creation and management*.

Keywords

accountability, conversational AI, multi-agent systems, causal responsibility, agentic CUI, governance

ACM Reference Format:
Mehmet Haklidir. 2026. Who Is Accountable When the Agent Speaks? Toward Accountability Graphs for Conversational AI Chains. In *ACM Conversational User Interfaces 2026 (CUI '26), July 21–24, 2026, Bremen, Germany.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3816046.3816264

ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of a national government. As such, the Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only. Request permissions from owner/author(s).
CUI '26, Bremen, Germany
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2741-2/26/07
https://doi.org/10.1145/3816046.3816264

1 Introduction

In 2024, a Canadian tribunal held Air Canada liable after its customer-service chatbot misrepresented the airline's bereavement fare policy, advising a grieving passenger that he could retroactively claim a discounted fare [5]. The airline argued its chatbot was "a separate legal entity"; the tribunal disagreed, ruling that a company is accountable for all information on its website. Yet this accountability question was resolved only because the system was monolithic: one chatbot, one airline, one policy database.

Now consider a near-future agentic version: a user asks the same airline's conversational assistant about bereavement fares. An orchestrating agent delegates to a policy retrieval sub-agent, which queries a knowledge base drawing on both structured fare rules and an unstructured FAQ corpus. A synthesis agent reconciles the results and states that retroactive claims are permitted, but they are not. Responsibility is now distributed across multiple components. Who is accountable?

This question is not hypothetical. CUIs are shifting to *agentic architectures* that act on behalf of users [1]. Frameworks such as LangChain, AutoGen, and CrewAI enable multi-agent orchestration where a primary agent delegates to specialized sub-agents, each of which may call external tools or models [21]. The resulting *conversational agent chain* can span multiple organizations and data sources.

Despite this complexity, CUI research has not developed systematic accountability frameworks for agent chains. Work on *trust* [6, 8], *transparency* [16], and *explainability* [9] addresses related but distinct concerns: trust asks whether users *will* rely on the system, transparency whether they *can see* what it does, and explainability *why* it produced a particular output. None answers *who bears responsibility* when the output causes harm.

This accountability gap is acute in CUI contexts. The conversational interface creates an *illusion of singularity*: users interact with what appears to be a single agent, but the singular persona masks multiple technical and organizational actors jointly producing each utterance. This is both an agency and an identity problem, directly relevant to the CUI 2026 theme. Real-time dialogue compounds decisions within milliseconds, making post-hoc tracing difficult. And *dynamic delegation* means the orchestrator decides at runtime which sub-agents to call, so the accountability structure emerges during conversation rather than being fixed at design time.

We address this gap with *Accountability Graphs*, a framework for modeling responsibility attribution in conversational AI agent chains. This paper is a conceptual work-in-progress: we develop the framework and illustrate it through a scenario, but do not yet

provide full formalization or empirical evaluation. Our contributions:

(1) We identify three disconnected research threads (structural ecosystem mapping, causal responsibility theory, and operational provenance tracking) and argue for their unification in the CUI context.
(2) We introduce Accountability Graphs, modeling agent chains as directed acyclic graphs with provenance metadata and causal responsibility semantics, adapted for CUI-specific challenges.
(3) We demonstrate the framework through a CUI scenario and propose a research agenda for empirical investigation.

2 Background and Related Work

2.1 Accountability in AI: Three Disconnected Threads

Accountability in AI has been addressed from three perspectives not yet integrated. *Structural mapping*: Bommasani et al.'s Ecosystem Graphs [2, 3] model asset-level dependencies across the foundation model ecosystem but perform no causal analysis and attribute no responsibility. *Causal responsibility theory*: Halpern and Pearl's [14] structural-model approach to actual causation, extended by Chockler and Halpern's [4] notions of degree of responsibility and blame, provides formal machinery for attribution but has been applied only at the model-internal level, never to pipeline or agent-chain accountability. *Operational provenance*: data provenance initiatives [17], model cards [18], datasheets [10], and accountability metrics catalogues [22] record decisions and lineage but lack formal frameworks for computing responsibility scores from the metadata.

2.2 The Accountability Gap in CUI Research

CUI research has studied trust [8], transparency [16], and explainability [9] in depth, but none addresses Nissenbaum's [19] *many hands problem*, where multiple agents contribute to an outcome and responsibility becomes diffuse. Cooper et al. [7] updated this analysis for algorithmic systems. The many-hands problem is amplified in CUI architectures, where the conversational interface acts as a singular facade behind which agents, tools, and services operate, each contributing causally to the output yet invisible to the user.

Santoni de Sio and Mecacci [20] identify four types of responsibility gaps in AI. Hindriks and Veluwenkamp [15] reframe these as *control gaps*, where no agent has sufficient control to be held responsible, a framing that applies directly to CUI agent chains. Separately, CUI research on conversational breakdowns, repair, and system self-disclosure [6] is relevant because accountability disclosures would arise at precisely such moments of failure; the design of accountability surfaces (Section 4.1) should build on this work.

3 Accountability Graphs for Conversational AI

3.1 Framework Overview

An *Accountability Graph* for a conversational agent chain is defined as a tuple $AG = (N, E, P, H)$ where:

- $N = \{n_1, \ldots, n_k\}$: *decision nodes*, each representing a point where a component makes a choice influencing the conversational output (e.g., intent classification, agent delegation, tool invocation, response generation).
- $E \subseteq N \times N$: directed *causal edges* representing information or control flow. Edges are typed: *dataflow edges* carry information between steps, while *configuration edges* represent design-time parameter settings (e.g., an orchestrator configuring a sub-agent's policy). Configuration edges do not create runtime cycles.
- $P : N \to \mathcal{P}$: a *provenance function* mapping each node to a record containing the responsible component, degree of decision autonomy, information available at decision time, temporal context, and contractual or design obligations.
- $H : N \to [0, 1]$: a *harm propagation function* capturing how a detected harm at a terminal node relates to upstream decisions, drawing on Halpern and Pearl's [14] structural-model approach to actual causation.

Node granularity is a modeling choice: a node may represent an entire agent, a single tool call, or a specific decision within a component. Finer granularity yields more precise attribution but requires more detailed provenance logging.

3.2 CUI-Specific Adaptations

Conversational agent chains raise accountability challenges absent in static ML pipelines. Table 1 summarizes five adaptations.

The most distinctive of these adaptations is the *accountability surface*, a user-facing representation of the internal accountability graph. Unlike traditional accountability analyses aimed at regulators or developers, the accountability surface translates responsibility distributions into natural language explanations for end users. We discuss the design challenges this raises in Section 4.1.

3.3 Illustrative Scenario: An Airline Bereavement Fare Agent

We extend the *Moffatt v. Air Canada* case [5] into a multi-agent CUI architecture. In the original 2024 case, a monolithic chatbot misrepresented bereavement fare policy; here, we imagine the same airline deploying an agentic CUI with distributed responsibilities.

User utterance: *"My grandmother just passed away. I need to fly from Vancouver to Toronto tomorrow. Do you have bereavement fares? Can I get a refund later if I book now at full price?"*

The system processes this through a seven-node agent chain (Figure 1):

(1) **Intent Parser** (n_1): Extracts bereavement context, route (YVR–YYZ), urgency, and flags the interaction as sensitive.
(2) **Orchestrator** (n_2): Routes to the policy retrieval sub-agent based on the bereavement flag.
(3) **Policy Retrieval Agent** (n_3): Formulates and sends a retrieval query to the knowledge base API.
(4) **Knowledge Base API** (n_4): Searches a structured fare rules database and an unstructured FAQ corpus; returns ranked results (third-party service).
(5) **Synthesis Agent** (n_5): Reconciles retrieved fragments. The structured rules require bereavement fares to be requested

Table 1: CUI-specific challenges for accountability attribution and corresponding framework adaptations.

CUI-Specific Challenge	Why It Matters	Framework Adaptation
Real-time dialogue	Decisions are made and compounded within milliseconds during a conversational turn	Temporal proximity weighting: nodes closer in time to the harm event receive proportionally adjusted responsibility consideration
User-facing accountability	Users interact with a single interface, but responsibility is distributed across the chain	*Accountability surface*: a mapping from the internal graph to user-legible responsibility attribution
Conversational context dependency	Earlier dialogue turns shape later decisions; the context window is a shared, evolving resource	Contextual provenance tracking that links node decisions to the conversational state at invocation time
Modality effects	Voice vs. text CUI may distribute responsibility differently (e.g., ASR misrecognition introduces a causal contributor absent in text-based CUI)	Modality-aware node types that distinguish ASR/TTS processing from language understanding
Dynamic delegation	Agents decide at runtime which sub-agents to invoke, meaning the graph structure emerges during conversation	Runtime graph construction: the Accountability Graph is built incrementally as delegation decisions are made, rather than defined statically at design time

before travel; the FAQ contains an outdated passage suggesting retroactive claims within 90 days.

(6) **Empathy Filter** (n_6): Adjusts tone for bereavement context.
(7) **Delivery** (n_7): Presents the response via text or voice.

Harm event: Facing contradictory sources, n_5 selects the FAQ passage (possibly because its retrieval score was higher) and tells the user that retroactive claims are permitted. The user books a full-price ticket, travels, and is denied the refund.

Accountability Graph analysis. The chain has edges $n_1 \to n_2 \to \ldots \to n_7$ plus a configuration edge from n_2 to n_5 (the orchestrator sets n_5's reconciliation parameters before execution, preserving acyclicity within a single turn; multi-turn iterative interactions would require unrolled graphs, left for future work). Causal backtracking identifies n_5 and n_4 as primary contributors and n_2 as secondary.

Worked example: computing H for n_5. Following Halpern and Pearl [14], let S_{n_5} be the binary variable representing n_5's source-selection decision ($S_{n_5} = faq$ in the actual world). Let Y be the binary harm outcome ($Y = 1$). The counterfactual intervention sets $S_{n_5} \leftarrow structured$, holding all other nodes fixed. Under this intervention, $Y = 0$ (the correct policy is returned). Since changing n_5's decision alone suffices to prevent the harm, n_5 satisfies Halpern-Pearl actual causation with degree of responsibility $\rho = 1$ (no other nodes need to be changed in the contingency). We then apply provenance weighting: n_5 operated with full decision autonomy ($w_{\text{autonomy}} = 1.0$), had access to both sources ($w_{\text{info}} = 1.0$), and was temporally proximate to the harm ($w_{\text{temporal}} = 0.9$). The final weighted score is $H(n_5) = \rho \cdot \bar{w} = 1.0 \cdot 0.97 = 0.97$. By contrast, n_4 has $\rho = 1$ (the harm is also counterfactually dependent on n_4's failure to flag the conflict) but lower autonomy weighting ($w_{\text{autonomy}} = 0.5$, as a retrieval service operating under prescribed parameters), yielding $H(n_4) = 0.80$. Node n_2 requires a two-step contingency (change n_2's configuration *and* re-run the pipeline), giving $\rho = 0.5$ and $H(n_2) = 0.43$.

This example is simplified: provenance weights are illustrative (in deployment, they could be designer-set, stakeholder-elicited, or learned from incident data), structural equations for all variables would need full specification, and stochastic components would require probabilistic extensions. The example nonetheless illustrates that the framework can produce *derived* responsibility scores from counterfactual reasoning and provenance data, rather than relying on assertion alone.

3.4 Toward Backward Responsibility Propagation

We sketch *backward responsibility propagation*, a procedure for traversing the graph from a harm event to upstream contributors:

(1) **Harm localization**: Identify the terminal node(s) where harm manifests in the conversational output.
(2) **Causal backtracking**: Traverse edges backward, evaluating at each node whether it satisfies Halpern and Pearl's [14] criteria for actual causation with respect to the harm.
(3) **Provenance-weighted scoring**: For each node identified as a cause, incorporate provenance factors (decision autonomy, information availability, temporal proximity, obligation scope) to produce a graded responsibility assessment.
(4) **Counterfactual validation**: For the highest-scoring nodes, verify the attribution through counterfactual reasoning: "Would the harm have occurred if this node's decision had been otherwise?"

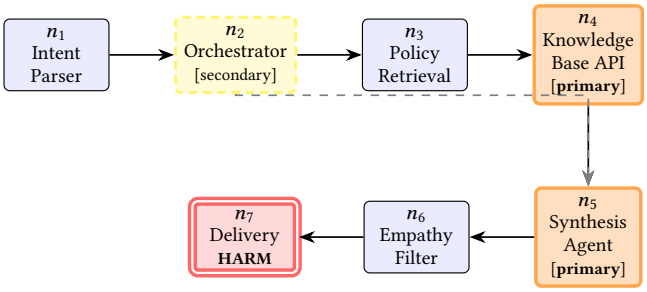


Figure 1: Accountability Graph for the bereavement fare scenario, extending the *Moffatt v. Air Canada* case into a multi-agent architecture. Nodes bearing primary causal responsibility (n_4, Knowledge Base API; n_5, Synthesis Agent) are marked with thick solid borders and the label "[primary]." The node with secondary responsibility (n_2, Orchestrator) has a thick dashed border and the label "[secondary]." The harm manifestation point (n_7) has a double border. The dashed edge represents the orchestrator's configuration of the synthesis agent's reconciliation logic.

Full formalization of this procedure is the subject of ongoing work. We note that the graph traversal itself is $O(|N| + |E|)$; the computational bottleneck is the counterfactual evaluation at each candidate node, which in the worst case requires re-execution of downstream subgraphs. Here we present the conceptual approach and focus on CUI-specific considerations.

4 Discussion

4.1 The Accountability Surface as a CUI Design Problem

The *accountability surface*, a user-facing translation of the internal responsibility graph, poses distinct design challenges. In text CUI, an explanation might read: "Your fare query drew on contradictory internal sources. The synthesis component selected the wrong one. We are correcting this and honoring the fare you were quoted." In voice CUI, explanations must be shorter: "There was an error in how we combined our fare information. We'll honor the quoted fare." The level of detail is a design variable: too little feels like a generic apology; too much risks *cognitive overload*.

Different stakeholders need different renderings. End users need actionable explanations and remediation paths. Auditors and regulators need full provenance trails and formal scores. Customer support staff need intermediate representations for routing complaints. A single graph can support all these surfaces. Importantly, accountability surfaces should not only explain what went wrong but also support *contestability*: providing users with clear pathways to challenge the attribution, request remedy, or escalate to human review.

In emotionally charged contexts such as bereavement, telling a user that "the synthesis component made an error" risks feeling impersonal or deflective. Accountability surfaces in sensitive interactions should draw on conversational repair and empathic response design [6], treating tone and timing as first-class parameters. How users receive accountability information when frustrated by a failure remains an open empirical question.

4.2 Regulatory Alignment

As a preliminary interpretive alignment (not a formal legal analysis), we note that Accountability Graphs map to several EU AI Act obligations: Article 9 (risk management) aligns with node-level provenance; Article 12 (logging) corresponds to runtime graph construction; Article 50 (disclosure to natural persons) extends from identity disclosure to responsibility disclosure via the accountability surface. The Act's provider-deployer distinction (Articles 16, 26) maps to the framework's ability to assign different responsibility weights to nodes maintained by different organizations.

4.3 Connecting Accountability and Consent

This work connects to *consent* in agentic AI. Recent research examines how consent breaks when CUI transitions from responding to acting [13], how a "consent vacuum" emerges in agent-to-agent interactions [12], and how consent chains degrade in embodied multi-agent systems where actions have physical, irreversible consequences [11]. Consent addresses the *before* (was permission obtained?), accountability the *after* (who bears responsibility?). Together they form a governance loop, and CUI is a natural domain to study it: the conversational interface serves simultaneously as the consent mechanism and the accountability surface.

4.4 Limitations, Risks, and Future Work

No empirical validation. We are developing a prototype instrumenting LangChain agent chains via callback hooks to construct Accountability Graphs at runtime. Next steps: a think-aloud study with CUI designers, followed by a controlled user study comparing accountability surface explanations against standard error messages, measuring comprehension, trust calibration, perceived fairness, and emotional appropriateness.

Power dynamics and accountability-washing. Graded responsibility distributions are not inherently benign. Organizations could use them to deflect blame onto third-party providers or users rather than accepting systemic responsibility. One possible governance mechanism is a minimum accountability floor for the entity controlling the orchestrator, regardless of downstream scores.

User agency. The framework does not include a user node. Yet users contribute to outcomes through ambiguous queries or ignored warnings. Whether to represent user contributions raises ethical questions: attributing responsibility to a grieving user who phrased a question imprecisely would be inappropriate, but ignoring user input may misrepresent the causal chain. We leave user-responsibility boundaries as an open question.

Formal responsibility vs. legal liability. Accountability Graphs are diagnostic, not normative: they identify where failures occurred, not what legal consequences should follow.

Traceability and opaque components. Many agent chains include proprietary "black box" components where causal backtracking into internal reasoning is impossible. For such nodes, the framework assigns responsibility based on input-output behavior and contractual

obligations only. A related challenge is *provenance logging in streaming CUI*: capturing metadata without disrupting token-by-token generation requires asynchronous logging hooks.

Cultural and jurisdictional variation. Accountability norms vary across cultures and legal systems. Provenance weights may need to reflect local regulatory expectations, and accountability surfaces should account for differing user expectations about institutional transparency. The current framework does not model these variations.

5 Conclusion

We have proposed *Accountability Graphs*, a framework that models conversational agent chains as provenance-enriched directed acyclic graphs with causal responsibility semantics. By extending the *Moffatt v. Air Canada* case into a multi-agent architecture, we illustrated how backward responsibility propagation can produce graded responsibility scores across a seven-node agent chain through counterfactual reasoning and provenance weighting. The framework unifies structural ecosystem mapping, causal responsibility theory, and operational provenance tracking, and adapts them to CUI-specific challenges including real-time dialogue, dynamic delegation, and the user-facing *accountability surface*. We connect this work to the governance concern of consent in agentic AI, arguing that consent and accountability together form a governance loop for agentic CUI systems.

Acknowledgments

The author used Claude (Anthropic) for brainstorming, structural feedback, and editing assistance on author-written text. All intellectual contributions, framework design, and analytical arguments are the author's own. The author takes full responsibility for the content, in accordance with the ACM Policy on Authorship.

References

[1] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arber, et al. 2022. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258* (2022).

[2] Rishi Bommasani, Thomas I. Liao, Dilara Soylu, and Percy Liang. 2024. Ecosystem Graphs: Documenting the Foundation Model Supply Chain. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, 1 (2024), 196–209. doi:10.1609/aies.v7i1.31638

[3] Rishi Bommasani, Dilara Soylu, Thomas I. Liao, Kathleen A. Creel, and Percy Liang. 2023. Ecosystem Graphs: The Social Footprint of Foundation Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. ACM. doi:10.1145/3593013.3594060

[4] Hana Chockler and Joseph Y. Halpern. 2004. Responsibility and Blame: A Structural-Model Approach. *Journal of Artificial Intelligence Research* 22 (2004), 93–115. doi:10.1613/jair.1391

[5] Civil Resolution Tribunal of British Columbia. 2024. Moffatt v. Air Canada. https://www.canlii.org/en/bc/bccrt/doc/2024/2024bccrt149/2024bccrt149.html. 2024 BCCRT 149.

[6] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garber, Emer Maney, Brendan O'Brien, Ross O'Connor, Brendan Spillane, and Benjamin R. Cowan. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, 1–12. doi:10.1145/3290605.3300705

[7] A. Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. 2022. Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM, 864–876. doi:10.1145/3531146.3533150

[8] Benjamin R. Cowan, Leigh Clark, Heloisa Candello, and Janice Tsai. 2023. Exploring Trust in Conversational AI: A Multidisciplinary Perspective. In *Proceedings of the 5th International Conference on Conversational User Interfaces (CUI '23)*. ACM. doi:10.1145/3571884.3603757

[9] Upol Ehsan and Mark O. Riedl. 2021. Explainability Pitfalls: Beyond Dark Patterns in Explainable AI. In *Human-Centered AI Workshop at NeurIPS 2021*.

[10] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (2021), 86–92. doi:10.1145/3458723

[11] Mehmet Haklidir. 2026. Consent Chain Degradation in Embodied Multi-Agent Systems: From Digital Delegation to Physical Irreversibility. In *WoRoBet 2026 Workshop at IEEE ICRA*.

[12] Mehmet Haklidir. 2026. The Consent Vacuum: Governing Agent-to-Agent Interaction in Multi-Agent AI Systems. In *IAIL 2026 Workshop at HHAI*. Under review.

[13] Mehmet Haklidir. 2026. From Chat to Agent: Why Consent Breaks When AI Acts on Your Behalf. In *CHI 2026 Workshop on Consent in Agentic AI*.

[14] Joseph Y. Halpern and Judea Pearl. 2005. Causes and Explanations: A Structural-Model Approach. Part II: Explanations. *The British Journal for the Philosophy of Science* 56, 4 (2005), 889–911. doi:10.1093/bjps/axi148

[15] Frank Hindriks and Herman Veluwenkamp. 2023. From Responsibility Gaps to Control Gaps. *Synthese* 201, 1 (2023), 21. doi:10.1007/s11229-022-04009-3

[16] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-User Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, 1–14. doi:10.1145/3290605.3300641

[17] Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeid, Damien Chan, Luca Soldaini, et al. 2024. A Large-Scale Audit of Dataset Licensing and Attribution in AI. *Nature Machine Intelligence* 6 (2024), 975–987. doi:10.1038/s42256-024-00878-8

[18] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, 220–229. doi:10.1145/3287560.3287596

[19] Helen Nissenbaum. 1996. Accountability in a Computerized Society. *Science and Engineering Ethics* 2, 1 (1996), 25–42. doi:10.1007/BF02639315

[20] Filippo Santoni de Sio and Giulio Mecacci. 2021. Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them. *Philosophy & Technology* 34 (2021), 1057–1084. doi:10.1007/s13347-021-00450-x

[21] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. In *Proceedings of the ICLR 2024 Workshop on LLM Agents*.

[22] Bo Xia et al. 2023. Towards a Responsible AI Metrics Catalogue: A Collection of Metrics for AI Accountability. *arXiv preprint arXiv:2311.13158* (2023).