From Open Lands to Boarded Mines: Modeling Conflicting Values in the WebAI Gold Rush

Nico Gießmann[1[0009−0007−4712−2338]] and Moreen Heine[1[0000−0002−7431−4251]]

Institute of Human-Centered Interactive Systems, University of Lübeck, Germany
{nico.giessmann,moreen.heine}@uni-luebeck.de

Abstract. The rise of Large Language Models (LLMs) let loose a gold rush for web data, slowly eroding the implicit bargain of an open web. While conceptual modeling of this WebAI ecosystem—especially regarding actor values that conflict internally with each other—presents significant opportunities for resolving these emerging tensions, the modeling community has largely been absent from this discourse. To address this gap, we propose a strategic dependency model between the actors using the i* (iStar) 2.0 modeling language. We briefly discuss measures and actors that could help in overcoming the tensions in the ecosystem, such as the implementation of advanced "no-train" directives and the potential for web archives to serve as "public data trusts" that mediate fair data use. Finally, we advocate for including actors of the ecosystem directly into the modeling process to validate the model and overcome the limitations of modelers' boundary judgments.

Keywords: WebAI Ecosystem · Conceptual Modeling · Enterprise-related Web Data.

1 Introduction

For decades, the web operated under a simple implicit bargain: Access to websites was open for everyone, including bots that crawl content. Many hosting plans charge based on bandwidth usage, thus, bot traffic significantly increases a website's operating costs. A report by Cloudflare shows that around one-third of traffic is caused by bots [6]. Yet, the firms behind bots (e.g., Google) redirected users to the website, allowing website operators to generate revenue from ads or sales.

The rise of AI, specifically LLMs, has disrupted this bargain. Lately, the Economist titled "AI is killing the web. Can anything save it?" [34]. Now, AI can directly provide users with answers, either by incorporating website content into the training data [13] or by using AI-assisted search [15]. Nearly 80% of AI bots are used to collect training data [36]. In both cases, users no longer have the need to visit the original website, substantially reducing opportunities to generate revenue for the website operator. Consequently, an increasing number of websites block AI bots (e.g., GPTBot) from accessing their website via `robots.txt` [11]. Yet, this raises the dilemma of reduced visibility as users increasingly favor

AI-mediated information access [29]. Moreover, AI model developers have been found to disregard `robots.txt` directives [16] and to face practical and procedural challenges in deleting training data (e.g., personally identifiable information according to the GDPR) [24]. This creates an environment of distrust against AI web crawling [20], and leads to more aggressive measures against crawling (e.g., infinite mazes that trap bots [17]).

This environment also affects many small startups that rely on accessing a specific subset of web data: enterprise-related web data. Today, web crawling is not only used for internal purposes [18], but has also become the foundation of startups whose core business model is based on such data [12]. They fit the pattern of data analytics provider startups proposed by [39]. As [7] describes these so-called 'WebAI' startups, they are "leveraging ML [machine learning] to systematically extract, validate, and interpret economic indicators from unstructured web content across multiple dimensions of organizational activity" (p. 3). This enterprise-related web data stems from corporate disclosures, such as job advertisements or press releases, which signal intentional strategy, and from digital trace data, such as employee reviews or professional networks activity, which may reveal actual organizational behavior [10]. Further amplified by AI-assisted web scraping tools [2], building data-driven products that 'mine' enterprise-related web data has become more accessible.

While data was once called "the new oil" [31], that analogy fits a previous era in which large tech tycoons hoarded value in closed silos. We reject this analogy for the WebAI ecosystem because the primary challenge is no longer access or ownership. A gold rush analogy is more suitable for the WebAI ecosystem because enterprise-related web data is, for now, publicly accessible and relatively easy to obtain. This has led to a chaotic frenzy involving small actors and big tech alike. However, this gold rush is analogous to the one in the 19th century in more than just the promise of wealth; it also replicates its destructive footprint. Just as mining washed away hillsides and choked rivers with sediment [23], the modern rush for enterprise-related web data is eroding the bargain the web was built upon. AI model developers and WebAI startups, acting as modern prospectors, leverage AI tools to sift through the web's sediment, often indifferent to the collateral damage.

This positions the current market on fragile stilts. Driven by the prospect of quick profits (and VC funding pressure), these WebAI startups act upon a short-term growth strategy, largely overriding the sustainability of their business model. As more and more enterprise-related web data is being used without proper consideration of consent, compensation, and bandwidth load, an increasing number of enterprises may choose to restrict access to their web data behind paywalls, login screens, or anti-crawling protections. While AI model developers might have the resources to negotiate licensing deals with companies [37], small startups are likely to become collateral damage in this 'gold rush.' The same may be happening to researchers dependent on enterprise-related web data [19].

Different disciplines are involved in finding solutions to this problem. These range from regulatory approaches [25] to technical advancements [15], yet the modeling community has largely been absent. This is critical, as models can serve as boundary objects that can bridge knowledge gaps between stakeholders [1][21]. Formal modeling imposes a meta-language that forces normative thinkers to be precise enough for implementers to build upon [21]. Implementation requires trade-offs, often sparking conflict. On one side, individuals seek to protect the privacy of their personally identifiable information embedded in enterprise-related web data, and enterprises strive to shield the intellectual value and controlled narrative of their corporate disclosure; on the other side, WebAI startups and AI model developers seek to harvest enterprise-related web data to fuel their revenue. Normative disciplines usually resist these trade-offs (reflecting a deontological approach to ethics commonly associated with the European context; [18]). Modeling mediates this by making the costs of values visible [4][3], thereby shifting modeling from mere implementation support to an analytic lens for understanding value conflicts in this WebAI ecosystem.

Against this backdrop, understanding how different actors rationalize their behavior becomes essential. Thus, this paper asks: *What actor values characterize the WebAI ecosystem?* The study's focus is descriptive, aiming to clarify the diverse values of actors whose interests intersect around publicly available enterprise-related web data.

To explore actor values that conflict internally with each other in the WebAI ecosystem, we adopt a qualitative, model-based analysis. Our approach aims to structurally explicate how different actors articulate their values and depend on one another around enterprise-related web data. We use the i* (iStar) 2.0 framework as an analytic lens to represent these values and their inter-dependencies at a conceptual level. The resulting model is not intended as an exhaustive or predictive representation, but as a boundary object that supports reflection and discussion across disciplinary perspectives, as well as informing future empirical studies.

2 Actor Values in the WebAI Ecosystem

The i* (iStar) framework is a goal and actor-oriented modeling language designed to support early-phase requirements engineering by focusing on the organizational context and rationales (the 'Whys') behind system requirements [8][41]. We use the strategic dependency (SD) view to visualize these concepts, which maps external relationships where actors depend on one another for goals, softgoals (values), tasks, and resources. We modeled the actors as roles rather than concrete agents to serve as an abstract characterization of their behavior. Actor associations (e.g., 'is-a') are not modeled to reduce complexity. The following presents each actor:

Enterprises seek to present their products to the public, attract customers and investors, and maintain the market value of their work. Their web presence becomes a key resource that researchers, WebAI startups, and AI model developers rely on. However, enterprises also pursue signaling-intensive communication (e.g., marketing) that enhances visibility but can reduce the validity of their public claims. They can depend on internet intermediaries to host and distribute their content, and on AI model developers and WebAI startups to respect their exclusion preferences or provide compensation when their content is reused.

WebAI startups collect and process large-scale enterprise-related web data to deliver commercial insights. They depend on AI model developers for foundational AI capabilities [22] and rely on web platforms' APIs, which can be volatile, restricted, or expensive [9]. Their differentiation rests on unique processing pipelines or crawled datasets.

AI model developers require vast quantities of web data to train foundational AI models. They depend on enterprises and internet intermediaries as the primary sources of this training data. In parallel, they supply analytic capabilities to WebAI startups that build products on top of their foundation models. AI model developers must navigate compliance pressures, alignment concerns, and compute constraints, while managing the reputational and regulatory risks associated with crawling and training on proprietary content. Their incentives are rarely aligned with the enterprises that produce the data they rely on, especially when compensation or consent is disputed. To date, the analytic capabilities needed by WebAI startups can only be achieved by using state-of-the-art LLMs, which are proprietary and operated by AI model developers. Therefore, WebAI startups not only rely on AI model developers for the models, but also on the inference (API costs).

Researchers aim to generate insights grounded in enterprise-related web data. They depend on enterprises for the availability and validity of this data and on regulators for legal clarity that determines what they are allowed to collect or process, even more so than WebAI startups, as they operate in institutional environments that demand higher levels of compliance (i.e., ethics committees). Their pursuit of data validity, stability of access, and sampling integrity makes them particularly sensitive to distortions created by enterprises' strategic communication or by internet intermediaries whose algorithmic decisions shape what researchers can observe. They also rely on WebAI startups and AI model developers insofar as these actors supply tools for crawling and analysis [7].

Regulators pursue the enforcement of legal frameworks such as data protection, competition rules, and copyright, as well as the evaluation of policies. They depend on researchers to obtain timely, granular, and trustworthy insights that supplement or replace slower traditional survey methods [7]. Conversely, researchers rely on regulators to interpret ambiguous legal obligations and to provide actionable guidance. Regulators also influence WebAI startups and AI model developers through their enforcement powers, shaping the incentives and constraints under which these actors use enterprise-related web data, as regulators are likewise interested in the economic success of these actors. Conversely, commercial actors may seek to influence regulators, for example, through lobbying, to establish favorable legal frameworks.

Internet intermediaries are social media platforms or content-hosting services that provide enterprises with the infrastructure necessary to communicate with stakeholders beyond their self-managed web presence [26]. They are motivated by user retention and monetize data through licensing or API access, which puts them in a structurally ambivalent position: enabling visibility for enterprises while profiting from the downstream extraction of that content. They control access pathways that researchers, WebAI startups, and AI model developers rely on, and their frequent API policy changes create instability in these dependencies. Currently, internet intermediaries such as Cloudflare are adopting a rather critical stance toward AI model developers [36].

Individuals are motivated by the productivity gains, personalized utility, and reduced information overload that AI tools offer [27]. To access these capabilities, they depend on AI model developers. This places individuals in a prosumer dilemma [30]: They are simultaneously consumers of AI models and uncompensated producers of the training data that powers them. While they intentionally signal professional value (e.g., LinkedIn profiles) or provide employer reviews, they are frequently unaware of the digital traces they leave behind. These are harvested by WebAI startups, AI model developers, and researchers to infer enterprise-related insights, often without the individuals' explicit consent or awareness of the systemic insights being derived. Lacking the bargaining power of enterprises or internet intermediaries, they rely heavily on regulators to enforce protections against algorithmic bias, manipulation, and the unauthorized extraction of their personal digital footprint.

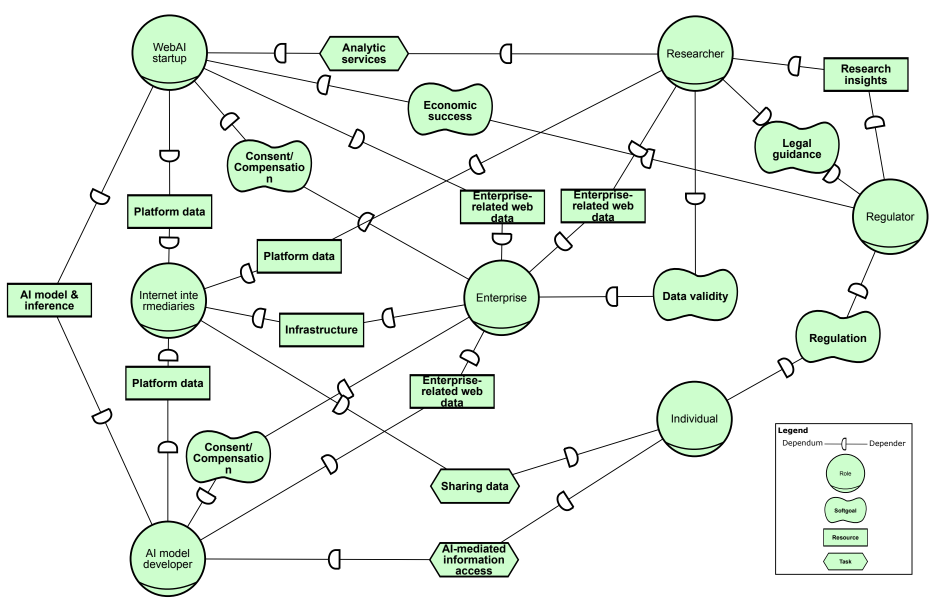

Fig. 1. Strategic Dependency Model of the WebAI Ecosystem.

3 Discussion

The SD model reveals that the WebAI ecosystem is characterized by deep-seated tensions arising from the diverging values of its actors. Enterprises seek visibility and market signaling; while this boosts their communication goals, it often undermines the data validity required by researchers. Similarly, intermediaries pursue monetization strategies that restrict API access, clashing with the operational needs of researchers and WebAI startups, who rely on stable and affordable data flows. While AI model developers and WebAI startups extract significant value from enterprise-related web data, enterprises increasingly demand compensation and consent, creating sharp disputes over the limits of fair use.

Over time, WebAI startups may rely less on AI model developers for inference because open-weight LLMs are becoming increasingly capable and can be either self-hosted or run by third-party inference providers, which operate in a highly competitive market likely to keep inference prices low.

Currently, the ecosystem operates in a regulatory vacuum. Voluntary directives, such as `robots.txt`, fail to enable the nuanced inclusion or exclusion of content that enterprises require (e.g., allowing product data to be crawled by LLMs while disallowing the use of personally identifiable information). To address this, the implementation of *data use declarations* is proposed, providing enterprises with granular control over the access, processing, and distribution of their web content [40] (e.g., via advanced "no-train" `robots.txt` directives or meta-tags [14]).

However, unrestricted blocking poses a risk to the public interest, as not all content should be subject to deletion or exclusion. So-called digital commons—the collective intellectual and cultural output of humanity—must remain accessible. Certain enterprise-related web data, such as press releases, patents, or news articles, may fall under this definition. To mediate the distinction between proprietary and commons data, a "public data trust" [5, p. 1] is proposed.

While not modeled as focal actors due to their limited involvement in direct value extraction, actors such as *web archives* or *civil society organizations* are uniquely positioned to assume this role. Historically, web archives were viewed primarily as repositories for cultural heritage—a 'museum' of the digital past. In the AI age, this conceptualization shifts dramatically; web archives can evolve into public data trusts that balance preservation with privacy. [33] notes the tension between preservation and the 'right to be forgotten' or communities who traditionally reject archiving. By providing a mechanism for enterprises or individuals to opt-out or negotiate terms (based on the regulators' laws), web archives could perform a 'trusted' crawling,' currently impossible when commercial entities scrape the open web indiscriminately.

The explicit orchestration of the ecosystem through such public data trusts could generate benefits for all actors, including reduced redundant crawling, lowered legal uncertainty, and decreased dependency on the volatile APIs of intermediaries. Nevertheless, key challenges remain regarding funding, control (e.g., public vs. private governance), resistance from incumbent actors, and the design of incentives to encourage participation and prevent circumvention.

Finally, navigating these tensions requires advanced methodological tools. Here, goal modeling, specifically using the strategic rationale view in i*, can help 'mediate' between actors. [35] argue that such mediation is too complex to be left to 'amateurs,' requiring a professional identity distinct from both the subject matter expert and the system builder (i.e., modelers). [4] calls for modelers who ensure that the intent of policy is not lost when translated into implementation. Furthermore, because modeling is inherently domain-dependent [32], it is prone to "boundary judgements" [28][38]. This limitation calls for advancing our approach by engaging ecosystem actors directly in the modeling process. Validating our model with these stakeholders will reveal their internal reasoning and help determine whether our proposed solutions can successfully restore the implicit bargain of the web.

Future research should empirically examine the modeled dependencies and assess how regulatory or institutional contexts shape the identified dependencies. Building on this foundation, the model can be used to explore (assisted by computational conflict detection) how interventions such as data use declarations or public data trusts redistribute value and responsibilities in the WebAI ecosystem.

References

1. Alberto Franco, L.: Rethinking soft or interventions: Models as boundary objects. European Journal of Operational Research **231**(3), 720–733 (2013). https://doi.org/10.1016/j.ejor.2013.06.033
2. Ayuso, E., Dumfeh Brogya, M.S., Kumar Ahlawat, V., Sain, M.: From manual to machine: How ai is redefining web scraping for superior efficiency: A literature review. In: 2024 International Conference on Communication, Control, and Intelligent Systems (CCIS). pp. 1–9. IEEE (2024). https://doi.org/10.1109/ccis63231.2024.10931912
3. Bammer, G.: Disciplining Interdisciplinarity: Integration and Implementation Sciences for Researching Complex Real-World Problems. ANU E Press, Canberra ACT 0200, Australia, 1 edn. (2013)
4. Bammer, G., O'Rourke, M., O'Connell, D., Neuhauser, L., Midgley, G., Klein, J.T., Grigg, N.J., Gadlin, H., Elsum, I.R., Bursztyn, M., Fulton, E.A., Pohl, C., Smithson, M., Vilsmaier, U., Bergmann, M., Jaeger, J., Merkx, F., Vienni Baptista, B., Burgman, M.A., Walker, D.H., Young, J., Bradbury, H., Crawford, L., Haryanto, B., Pachanee, C.a., Polk, M., Richardson, G.P.: Expertise in research integration and implementation for tackling complex problems: when is it needed, where can it be found and how can it be strengthened? Palgrave Communications **6**(1) (2020). https://doi.org/10.1057/s41599-019-0380-0
5. Chan, A., Bradley, H., Rajkumar, N.: Reclaiming the digital commons: A public data trust for training data. In: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. pp. 855–868. ACM, New York, NY, USA (2023). https://doi.org/10.1145/3600211.3604658
6. Cloudflare, Inc.: State of application security 2024. Techreport, Cloudflare, Inc. (Jul 2024)
7. Dahlke, J., Schmidt, S., Lenz, D., Kinne, J., Dehghan, R., Abbasiharofteh, M., Schütz, M., Kriesch, L., Hottenrott, H., Kanilmaz, U.N., Grashof, N., Hajikhani,

A., Liu, L., Riccaboni, M., Balland, P.A., Wörter, M., Rammer, C.: The webai paradigm of innovation research: Extracting insight from organizational web data through ai. ZEW Discussion Papers 25-019, Mannheim (2025), https://hdl.handle.net/10419/319890
8. Dalpiaz, F., Franch, X., Horkoff, J.: iStar 2.0 Language Guide (06 2016). https://doi.org/10.48550/arXiv.1605.07767, retrieved December 1, 2025
9. Freelon, D.: Computational research in the post-api age. Political Communication **35**(4), 665–668 (2018). https://doi.org/10.1080/10584609.2018.1477506
10. Gießmann, N.: Enterprise architecture traces on the web: An ontology-driven integrative review. In: Companion Proceedings of the 18th IFIP Working Conference on the Practice of Enterprise Modeling (POEM 2025). Geneva, Switzerland (December 2025), https://ceur-ws.org/Vol-4171/paper_7.pdf
11. Gillham, J.: Block AI bots from crawling websites using robots.txt (aug 2024), https://originality.ai/ai-bot-blocking
12. Hartmann, P.M., Zaki, M., Feldmann, N., Neely, A.: Capturing value from big data – a taxonomy of data-driven business models used by start-up firms. International Journal of Operations & Production Management **36**(10), 1382–1406 (2016). https://doi.org/10.1108/ijopm-02-2014-0098
13. Huang, S., Siddarth, D.: Generative ai and the digital commons (3 2023). https://doi.org/10.48550/arXiv.2303.11074, retrieved December 16, 2023
14. Ippolito, D., Yu, Y.W.: DONOTTRAIN: A metadata standard for indicating consent for machine learning. In: Proceedings of the 40th International Conference on Machine Learning. PMLR, Honolulu, Hawaii, USA (2023), https://blog.genlaw.org/CameraReady/42.pdf
15. Jiménez, J., Arkko, J.: AI, robots.txt. In: Submissions to the IAB AI-CONTROL workshop. p. 5. Washington DC, USA (Sep 2024)
16. Kim, T., Bock, K., Luo, C., Liswood, A., Poroslay, C., Wenger, E.: Scrapers selectively respect robots.txt directives: Evidence from a large-scale empirical study. In: Proceedings of the 2025 ACM Internet Measurement Conference. pp. 541–557. ACM, New York, NY, USA (2025). https://doi.org/10.1145/3730567.3764471
17. Koebler, J.: Developer creates infinite maze that traps AI training bots (jan 2025), https://www.404media.co/developer-creates-infinite-maze-to-trap-ai-crawlers-in
18. Krotov, V., Johnson, L., Silva, L.: Legality and ethics of web scraping. Communications of the Association for Information Systems **47**, 539–563 (2020). https://doi.org/10.17705/1cais.04724
19. Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., Van Alstyne, M.: Computational social science. Science **323**(5915), 721–723 (2009). https://doi.org/10.1126/science.1167742
20. Liu, E., Luo, E., Shan, S., Voelker, G.M., Zhao, B.Y., Savage, S.: Somesite i used to crawl: Awareness, agency and efficacy in protecting content creators from ai crawlers. In: Proceedings of the 2025 ACM Internet Measurement Conference. pp. 78–99. ACM, New York, NY, USA (2025). https://doi.org/10.1145/3730567.3732913
21. Luna-Reyes, L.F., Black, L.J., Ran, W., Andersen, D.L., Jarman, H., Richardson, G.P., Andersen, D.F.: Modeling and simulation as boundary objects to facilitate interdisciplinary research. Systems Research and Behavioral Science **36**(4), 494–513 (2018). https://doi.org/10.1002/sres.2564
22. Maruping, L.M., Yang, Y.: From open to balkanized data flows in digital platforms: Toward a framework of the paradox of data openness in the age of algorithms. Strategic Management Review (2025), forthcoming

23. National Park Service: Environmental consequences of the gold rush (Jul 2024), https://www.nps.gov/klgo/learn/historyculture/environmental-impacts.htm
24. noyb: ChatGPT verbreitet falsche infos über personen – und OpenAI kann nichts tun. https://noyb.eu/de/chatgpt-provides-false-information-about-people-and-openai-cant-correct-it (Apr 2024), retrieved November 29, 2025
25. Pasqual, F., Sun, H.: Consent and compensation: Resolving generative AI's copyright crisis. Virginia Law Review Online **110**, 207–247 (Aug 2024), https://heinonline.org/HOL/P?h=hein.journals/inbrf110&i=207
26. Perset, K.: The Economic and Social Role of Internet Intermediaries. OECD Digital Economy Papers 171, OECD Publishing, Paris (2010). https://doi.org/10.1787/5kmh79zzs8vb-en
27. Przegalinska, A., Triantoro, T., Kovbasiuk, A., Ciechanowski, L., Freeman, R.B., Sowa, K.: Collaborative ai in the workplace: Enhancing organizational performance through resource-based and task-technology fit perspectives. International Journal of Information Management **81**, 102853 (2025). https://doi.org/10.1016/j.ijinfomgt.2024.102853
28. Reynolds, M., Wilding, H.: Boundary critique: an approach for framing methodological design. In: de Savigny, D., Blanchet, K., Adam, T. (eds.) Applied systems thinking for health systems research: A methodological handbook, chap. 3, pp. 38–56. Open University Press, Maidenhead (2017)
29. del Rio-Chanona, R.M., Laurentsyeva, N., Wachs, J.: Large language models reduce public knowledge sharing on online q&a platforms. PNAS Nexus **3**(9) (2024). https://doi.org/10.1093/pnasnexus/pgae400
30. Ritzer, G., Jurgenson, N.: Production, consumption, prosumption. Journal of Consumer Culture **10**(1), 13–36 (2010). https://doi.org/10.1177/1469540509354673
31. Rotella, P.: Is data the new oil? Forbes (Apr 2012), https://www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/, retrieved November 27, 2025
32. Roussos, J.: Modelling in normative ethics. Ethical Theory and Moral Practice **25**(5), 865–889 (2022). https://doi.org/10.1007/s10677-022-10326-4
33. Schafer, V., Winters, J.: The values of web archives. International Journal of Digital Humanities **2**(1-3), 129–144 (2021). https://doi.org/10.1007/s42803-021-00037-0
34. The Economist: Ai is killing the web. can anything save it? The Economist (Jul 2025), https://www.economist.com/business/2025/07/14/ai-is-killing-the-web-can-anything-save-it
35. Tolk, A., Oren, T.: The Profession of Modeling and Simulation. Wiley (2017). https://doi.org/10.1002/9781119288091
36. Tomé, J.: The crawl-to-click gap: Cloudflare data on AI bots, training, and referrals. https://blog.cloudflare.com/crawlers-click-ai-bots-training/ (Aug 2025), retrieved November 27, 2025
37. Tong, A., Wang, E., Coulter, M.: Exclusive: Reddit in AI content licensing deal with Google. Reuters (Feb 2024), https://www.reuters.com/technology/reddit-ai-content-licensing-deal-with-google-sources-say-2024-02-22/, accessed: 2025-12-16
38. Ulrich, W., Reynolds, M.: Critical Systems Heuristics, pp. 243–292. Systems Approaches to Managing Change: A Practical Guide, Springer, London (2010). https://doi.org/10.1007/978-1-84882-809-4_6
39. Weber, M., Beutter, M., Weking, J., Böhm, M., Krcmar, H.: Ai startup business models. Business & Information Systems Engineering **64**(1), 91–109 (2021). https://doi.org/10.1007/s12599-021-00732-w

40. Yang, C., Liao, H.: Using the robots.txt and robots meta tags to implement online copyright and a related amendment. Library Hi Tech **28**(1), 94–106 (2010). https://doi.org/10.1108/07378831011026715
41. Yu, E.: Towards modelling and reasoning support for early-phase requirements engineering. In: Proceedings of ISRE '97: 3rd IEEE International Symposium on Requirements Engineering. pp. 226–235. IEEE Comput. Soc. Press (1997). https://doi.org/10.1109/isre.1997.566873