Generative AI Supply Shocks and Online Reputation: Evidence from Trustpilot

A popular media narrative known as the “dead internet theory” suggests that automated content is increasingly displacing human contributions online. In the context of online reviews, this raises two empirical questions: whether generative AI is being systematically used to produce reviews and, if so, to what ends. Both are difficult to answer because AI-generated content is hard to identify reliably at scale. We study 13 million Trustpilot reviews, exploiting discrete shocks to the supply of generative AI arising from new large language model releases and reductions in model prices. Our difference-in-differences design compares changes in unverified reviews with contemporaneous changes in verified reviews around each shock, allowing us to identify patterns consistent with AI use. We find that following these shocks, unverified ratings become more negative relative to verified ratings: the share of one-star reviews rises, the share of five-star reviews falls, and average ratings decline. New model releases account for most of this shift, which is concentrated among firms in the lowest and highest review-volume quartiles. We also find evidence of within-day same-company spikes in review activity. Although reviewers’ motives are unobserved, the shift toward negative ratings appears more consistent with the strategic use of AI to damage competitors’ reputations than with either changes in how consumers use generative AI when writing reviews or changes in the platform’s review filters. We discuss the implications for platforms, firms, and consumers, as well as for our understanding of the strategic use of generative AI in markets.