Who Pilots the Copilots?

Mapping a Generative AI's Actor-Network to Assess Its Educational Impacts

Francesco Balzan[1(✉)], Monique Munarini[2], and Lorenzo Angeli[3]

[1] University of Bologna, Via Zamboni 33, 40126 Bologna, Italy
francesco.balzan3@unibo.it
[2] University of Pisa, Largo B. Pontecorvo, 56127 Pisa, Italy
monique.munarini@phd.unipi.it
[3] University of Trento, Via Sommarive 9, 38123 Povo (TN), Italy
lorenzo.angeli@unitn.it

Abstract. Generative AI (GenAI) is praised as a transformative force for education, with the potential to significantly alter teaching and learning. Despite its promise, debates persist regarding GenAI impacts, with critical voices highlighting the necessity for thorough ethical scrutiny. While traditional ethical evaluations of GenAI tend to focus on the opacity of AI decision-making, we argue that the true challenge for ethical evaluation extends beyond the models themselves, and to the socio-technical networks shaping GenAI development and training. To address this limitation, we present an evaluation method, called Ethical Network Evaluation for AI (ENEA), which combines Latour's Actor-Network Theory—used to map network dynamics by tracing actors' interests and values—with Brusseau's AI Human Impact framework, which identifies ethical indicators for evaluating AI systems. By applying ENEA to GenAI "copilots" in education, we show how making Actor-Networks visible lets us unveil a great variety of dilemmas, guiding ethical auditing and stakeholder discussions.

Keywords: AI for Education · Actor-Network Theory · AI human impact · socio-technical systems · GitHub Copilot · Programming classes

1 Introduction

> Here is the question I wish to raise to designers: where are the visualization tools that allow the contradictory and controversial nature of matters of concern to be represented?
>
> Bruno Latour [12]

The field of Science and Technology Studies (STS) sees AI systems as socio-technical constructs integrating technical components with social elements [7].

A. M. Olney et al. (Eds.): AIED 2024, LNAI 14830, pp. 448–456, 2024.
https://doi.org/10.1007/978-3-031-64299-9_42

In this paper, we argue that the heterogeneous nature of AI systems gives rise to two kinds of intractability for their ethical assessment: technical intractability, stemming from the AIs' opaque technical aspects, and socio-technical intractability, arising from the transparency (i.e., invisibility) of the socio-technical networks behind AI development and training. While much attention has been given to technical intractability, our focus is on illuminating the socio-technical aspect.

For this purpose, we propose ENEA (Ethical Network Evaluation for AI), a method combining Actor-Network Theory (ANT) with James Brusseau's AI Human Impact framework, to systematically map the flow of interests and values of stakeholders involved in the development and training of GenAI chatbots operating as "copilots"[1] in educational settings. We pose three directions to guide our inquiry a) focusing on mapping interests and values, b) integrating descriptive mapping with a normative framework, and c) understanding the educational impacts of GenAI copilots.

In Sect. 2 we briefly present the paper's positioning in the literature (i.e., Actor-Network Theory and AI ethical evaluation); in Sect. 3, we present our analytical model, ENEA, from its structure to a worked example on GitHub Copilot (Sect. 4). To close the contribution, Sect. 5 presents the current limitations of our approach and avenues for future work.

2 Background

Actor-Network Theory (ANT), conceptualised by Michel Callon and Bruno Latour, is a framework adept at unravelling the complexities of socio-technical systems. ANT explores the interplay between technology and social processes [6,14,18] through the definition of "actants", emphasising the interdependent nature of technological and human agency [13]. In ANT, artefacts embody and transmit values and interests in the form of "prescriptions" enacted by the agents deploying such technology. ANT's focus is on evaluating and elucidating the role of "mediators": actants producing unpredictable outputs from inputs, which Latour characterised as being so embedded in our networks to become unquestioned, accepted facets of the status quo [13]. Much like our elementary mental states and cognitive processes [15], mediators and their prescriptions are transparent—they operate below the threshold of attention. ANT aims to unravel the mediators' transparent interactions and prescriptions, making it suitable for addressing the issue of AI's socio-technical intractability. Yet, tracing the actants' values and interests is only a descriptive step. Can we empower ANT with normative capacities to guide the ethical assessment of GenAI in education?

Various authors have tried to augment ANT with normative capabilities [1,4,11]. Latour himself proposed the concept of "matters of concern" which, in contrast to "matters of facts", allows us to conceive technical systems as embedded within complex socio-technical networks, entangled in social, political, and

[1] Of which a famous example is GitHub Copilot.

ethical relationships. We contend that AI technologies, especially in their application in education, should be understood not as "matters of fact", but as "matters of concern". In the language of this paper, shifting to a "matters of concern" approach mirrors the attention shift from the AI systems' technical intractability to their socio-technical intractability. In other words, conceiving and assessing AI systems as "matters of concern" is the act of following the flow of mediators and their prescriptions, making their transparent networks visible.

We enable this shift of attention by integrating ANT with indicators from James Brusseau's AI human impact framework for ethical assessment. Most ethical assessments of AI emphasise challenges of technical intractability, originating from the models' design (e.g., algorithmic biases [7]) and performance [10] without tackling socio-technical intractability - i.e., treating AI systems as "matters of fact" rather than as "matters of concern". While some proposed strategies aim to extend internal ethical auditing processes to stakeholders involved in the development of the AI system [17], others suggested contextualizing the AI system within the socio-technical network in which it will be deployed [7], we claim that they do not fully address socio-technical intractability. This limitation stems from a narrow conceptualization of actors and agents, which overlooks how these components propagate prescriptions and influence the final product. In essence, the values and interests embedded within these network components remain partially unexplored due to their transparency, highlighting a gap in the current methodologies for ethical assessment. To bridge this gap, our evaluation framework, called ENEA, draws from James Brusseau's AI Human Impact framework [5], as it offers a powerful synthesis of the key ethical issues connected to AI in both research and policy contexts.

3 The ENEA Framework

The AI and Education (AIED) field amply explored the impact of Generative AI (GenAI) on learning. Studies range from investigating its potential to enhance or impede human learning [8] to its influence on meta-learning and critical thinking [2]. In this paper, we aim to discuss some educational implications of GenAI copilots: AI-powered tools designed to support users in various tasks by generating human-like responses, suggestions, or content based on user input. We argue that their mode of side-by-side interaction carries its own ethical dilemmas due to the configuration of GenAI copilots as "actors" that require human-like interaction, which can only be effectively addressed and evaluated if GenAI copilots are conceived as "matters of concern".

The fundamental premise of ENEA lies in the understanding that, in ANT, values and interests can be framed as prescriptions that propagate through the actants' interactions in the network. ENEA thus becomes a tool to reveal the complex interactions between technological advancements, ethical considerations, and educational experiences that establish or replicate power relations and social dynamics. ENEA aims to shed light on how each copilot carries its own prescriptions, and how these prescriptions may affect the educational space.

To do so, we need to address socio-technical intractability. As a first step, we propose a formalisation-visualisation of ANT that represents actants and their inter-relations as vertices and edges in a directed graph. This mapping highlights the heterogeneous nature of the GenAIs' Actor-Networks (the vertices), and visualises the flow of prescriptions (the edges). In line with the broader ANT tradition, we focus on discussing the structure of the Actor-Network rather than specifying precise inclusion/exclusion criteria. While the choice of specific vertices and edges that we present here is deductive, we could also be more inductive, adding actants, prescriptions and connections based on findings from the literature. We argue, though, that the "veracity" of the model would not change, as Actor-Networks are contingent constructions rather than static objects.

Our graphs, "ENEA maps", are structured around two boundaries. The first boundary separates where the GenAI is built (its "upstream") from where it is used (its "downstream"). This boundary reflects the distinction between designers and users, with consequences that are amply explored in interaction design [3]. The second boundary is internal to the upstream, and separates design and development from training data. Each GenAI thus becomes a double boundary object [20], mediating between the upstream and the downstream, but also between designers/developers and data gathered from people. Ultimately, ENEA aims to shed light on the prescriptions that teachers and students using copilots may receive from deep within the Actor-Network. Some of these prescriptions, we argue, may give rise to potential issues in autonomy, dignity, equity, performance and accountability, summarised in Table 1. The result of an ENEA

Table 1. Ethical indicators from AI Human Impact, their applicability to GenAI in education, and how to use ENEA to analyse a Copilot for that indicator.

Indicator	Applicability to GenAI in Education	ENEA analysis
Autonomy	Students and teachers should be able to co-define their own rules, without the AI carrying external interference	Does the copilot embed any "rules" (prescriptions) in its design? Who is the source of those rules? Follow the ENEA map to check if any of those rules reach the student-teacher relationship (passing through the Copilot)
Dignity	Students should be able to use AI tools exclusively to learn and enhance their knowledge, without the AI tools serving others' purposes	What other ends might be embedded in the Copilot's design? Who is the origin of those ends? Follow the ENEA map to check whether any of those ends reach the students (passing through the Copilot)
Performance	The output of AI tools should be consistently functional, effective, and efficient	What may alter the quality of the Copilot's outputs? Who affects the quality? Follow the ENEA map to check whether any sources of performance (other than the students' input) reach the Copilot block
Accountability	Students and teachers should recognise when their tools make decisions, how and why these decisions are taken, and know whose responsibility it is for their tools' outputs	Who or what is the origin of the Copilot's text suggestion? Can students recognise this? Follow the ENEA map to see if the students can reach the origin of the text suggestions
Equity	All students in classes using Copilot should be able to successfully leverage the tool	Are there cases in which, within the same class, some students may be able to productively use the Copilot, while others receive fewer benefits? In the ENEA map, check whether the students' block can be subdivided into sub-blocks (of students who would receive different benefits)

analysis, then, is the flagging of further investigation on the deployment of the GenAI copilot according to the relevant indicator.

4 Worked Example: Applying ENEA to GitHub Copilot

In computing education, Yilmaz et al. [22] report that copilot-like GenAI systems like ChatGPT can boost computational thinking, self-efficacy, and motivation among programming students. At the same time, other studies focusing on GitHub Copilot, a GenAI system specifically tuned for code generation, suggest that GenAI's productivity benefits may lead to superficial understanding and reduced problem-solving and creativity [9,21]. Takerngsaksiri et al. [21], in particular, caution that relying on GenAI for coding might foster dependency on autocompletion features rather than developing authentic coding proficiency.

We suggest that, while these studies offer valuable perspectives on the immediate educational effects of GenAI copilots, they lack a comprehensive analysis of the potential impacts emerging from deeper within the copilot's Actor-Network. In this section, we propose a worked example of the ENEA analysis of GitHub Copilot to highlight how prescriptions propagate within its specific Actor-Network, while generating ethical tensions. We will first present the structure of GitHub's Actor-Network, and then conduct a brief ENEA Analysis following the principles outlined in Table 1.

We provide a first approximation of GitHub Copilot's Actor-Network in Fig. 1. In that, the downstream is a schematic representation of the social dynamics that compose the educational setting, while the upstream represents the far more complex situation of Copilot's construction. On the training side, we acknowledge that Copilot's training data is, per GitHub's own documentation[2], largely based on GitHub projects and StackExchange answers; on the design and development, we summarise the network of organisations that design the model and their parent organisations.

Even in this simplified representation, some actants and relationships are noteworthy in their crossing of boundaries: on the training side, different *programming languages* gather different communities on GitHub and StackExchange, and the popularity of programming languages creates impacts on performance and equity, as discussed below; on the design and development side, *Microsoft* holds a crucial role as a common controlling agent for the two companies that develop Copilot, GitHub and OpenAI[3]. Microsoft's position in the design and development Actor-Network means that one single actant can affect Copilot both through its base model and through its control over the company deploying it.

There also are two elements that cross the ENEA structural boundaries. The "maximisation of return on investment" (RoI maximisation) prescription

[2] See https://github.com/features/copilot (FAQs section, General question #4. (Accessed 2024/01/26).

[3] See https://openai.com/our-structure and the profiles of GitHub's leadership at https://github.com/about/leadership (Accessed 2024/01/26).

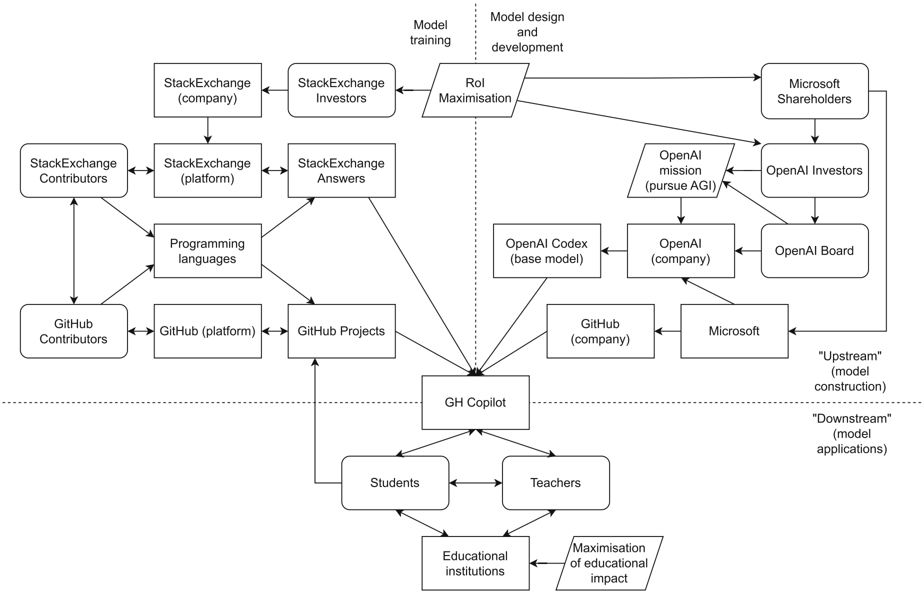

Fig. 1. The general ENEA map for Copilot. Rounded rectangles represent human actants, rectangles represent non-human actants, parallelograms represent prescriptions. Arrows show the propagation of prescriptions.

connects to actants on both sides of Copilot's upstream, and creates an alignment between the model's training set and its design and development, blurring the upstream boundary. Students also enable boundary-crossing when they publish their code on GitHub, which may eventually become part of the model's data set. In this way, students, who would normally be passive "users" of an education technology [19] acquire a direct avenue to become part of their own educational tool's upstream.

As for the analysis of GitHub Copilot, in Table 1, we defined *Autonomy* as "giving rules to oneself" [5]. We should then check whether the self-determination of students and teachers is preserved in the educational context. Since there is a path that connects the RoI maximisation and OpenAI Mission to the students and teachers that pass through GH Copilot, we can claim that Copilot requires further investigation for Autonomy. We can apply a similar logic for *Dignity*: RoI Maximisation and the pursuit of AGI are not just rules, but also ends, and they trickle down to the students passing through Copilot, meaning that Copilot should also be thoroughly assessed for Dignity. As for *Performance*, the popularity of programming language is a substantial source of performance[4], which reaches the Copilot block, other than the students' input. GitHub Copilot should also be assessed for Performance. In *Accountability*, there is no clear path for the

[4] See https://docs.github.com/en/copilot/using-github-copilot/getting-started-with-github-copilot (Accessed 2024/01/26).

students and teachers to recognise the origin of Copilot's suggestions, and the Actor-Network shows no clear way for the students to recognise who is responsible for the tools' output, meaning that Copilot deserves further investigation in Accountability, too.

Equity deserves its own deeper discussion. In Equity, the ultimate goal is respecting inclusion and diversity to integrate individuals within a community, and avoid marginalisation [16]. GitHub's documentation acknowledges Copilot's better performance with English prompts[5] and certain programming languages (as discussed above), which may disadvantage students using underrepresented (natural or programming) languages. The student block in the ENEA map could thus be sub-divided on the basis of natural or programming language. This gives rise to Equity concerns and is an issue of intersectionality, since this dual linguistic challenge impacts students twice based on their language and programming preferences, affecting a broad range of learners beyond legally protected groups.

5 Limitations and Conclusions

In this contribution, we presented ENEA, an evaluation method to orient ethical assessment when considering the deployment of GenAI copilots in education. ENEA aims at challenging the characterisation of AI systems as inherently intractable, a narrative supported by the major AI companies[6] to create a discourse that sees AI systems as superhumanly autonomous and sophisticated. Our work on ENEA challenges the public perception of AI-generated content as an objective, distilled essence of the collective human knowledge which, we argue, conceals the links between the AI tools' functioning and venal interests that exist in its Actor-Network.

Current limitations of our work are linked to the tradition ENEA comes from: the visualisations we propose, along with many of the considerations that it lets us draw, are contingent and easily subject to individual biases. We argue, however, that this is not necessarily a shortcoming, but rather should be seen as a feature of ENEA, and an immediately-acknowledged step towards disclosure of inevitably-present personal stances in AI assessment.

ENEA can aid in pinpointing problematic actants or prescriptions in an Actor-Network, and help plan interventions. With due adaptations, we see ENEA as potentially useful beyond educational settings. We hope that ENEA can become a tool for the scientific and educational communities to highlight points of attention, focus collective action, and ultimately build GenAI systems that respect all the involved humans and non-humans.

---

[5] See for example https://docs.github.com/en/copilot/github-copilot-in-the-cli/about-github-copilot-in-the-cli (Accessed 2024/01/26).

[6] Effectively summarised in the TESCREAL acronym: Transhumanism, Extropianism, Singularitarianism, Cosmism, Rationalism, Effective Altruism, and Longtermism.

Acknowledgments. L.A. thanks Fabio Gasparini for the many insightful conversations and comments. F.B. was supported by Future AI Research (FAIR) PE01, SPOKE 8 on PERVASIVE AI funded by the National Recovery and Resilience Plan (NRRP).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Akrich, M.: The description of technical objects (1992)
2. Barana, A., Marchisio, M., Roman, F.: Fostering problem solving and critical thinking in mathematics through generative artificial intelligence. In: 20th International Conference on Cognition and Exploratory Learning in the Digital Age. PRT (2023)
3. Bardzell, J., Bardzell, S.: The user reconfigured: on subjectivities of information. In: Proceedings of the Fifth Decennial Aarhus Conference on Critical Alternatives. CA 2015, pp. 133–144. Aarhus University Press, Aarhus N, August 2015
4. Bowker, G.C., Star, S.L.: Sorting Things Out: Classification and Its Consequences. MIT Press, Cambridge (2000)
5. Brusseau, J.: AI human impact: toward a model for ethical investing in AI-intensive companies. J. Sustain. Financ. Invest. **13**(2), 1030–1057 (2023)
6. Callon, M., Latour, B.: Don't throw the baby out with the bath school! A reply to collins and yearley. In: Science as Practice and Culture, pp. 343–368. University of Chicago Press (1992). https://doi.org/10.7208/9780226668208-013
7. Dignum, V.: Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30371-6
8. Ernst, N.A., Bavota, G.: AI-Driven development is here: should you worry? IEEE Softw. **39**(2), 106–110 (2022)
9. Finnie-Ansley, J., Denny, P., Becker, B.A., Luxton-Reilly, A., Prather, J.: The robots are coming: exploring the implications of OpenAI codex on introductory programming. In: Proceedings of 24th Australasian Computing Education Conference. ACE 2022, pp. 10–19. Association for Computing Machinery, February 2022
10. Hickman, S.E., Baxter, G.C., Gilbert, F.J.: Adoption of artificial intelligence in breast imaging: evaluation, ethical constraints and limitations. Br. J. Cancer **125**(1), 15–22 (2021)
11. Introna, L.D.: Ethics and the speaking of things. Theory Culture Soc. **26**(4), 25–46 (2009)
12. Latour, B.: A Cautious Prometheus? A Few Steps Toward a Philosophy of Design, p. 2. Universal Publishers (2008). https://sciencespo.hal.science/hal-00972919
13. Latour, B.: Reassembling the Social: An Introduction to Actor-Network-Theory. OUP Oxford, September 2007
14. Law, J.: After ant: complexity, naming and topology **47**, 1–14 (1999)
15. Metzinger, T.: Phenomenal transparency and cognitive self-reference **2**(4), 353–39 (2003). https://doi.org/10.1023/B:PHEN.0000007366.42918.eb
16. Minow, M.: Equality vs equity **1**, 167–193 (2021). https://doi.org/10.1162/ajle_a_00019
17. Raji, I.D., et al.: Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: 2020 Conference on Fairness, Accountability, and Transparency. ACM, January 2020

18. Rydin, Y.: Actor-network theory and planning theory: a response to booelens. Plan. Theory **9**(3), 265–268 (2010)
19. Selwyn, N.: On the Limits of Artificial Intelligence (AI) in Education **10**, 3 (2024). https://doi.org/10.23865/ntpk.v10.6062
20. Star, S.L., Griesemer, J.R.: Institutional ecology, 'translations' and boundary objects: amateurs and professionals in Berkeley's museum of vertebrate zoology, 1907–39. Soc. Stud. Sci. **19**(3), 387–420 (1989)
21. Takerngsaksiri, W., Warusavitarne, C., Yaacoub, C., Hou, M.H.K., Tantithamthavorn, C.: Students' perspective on AI code completion: benefits and challenges, October 2023
22. Yilmaz, R., Karaoglan Yilmaz, F.G.: The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation. Comput. Educ. **4** (2023)