Designing Responsible AI: A pedagogical blueprint for interdisciplinary reflection and applied ethics competencies

Amanda Horzyk
amandahorzyk@outlook.com
School of Informatics, University of Edinburgh
Edinburgh, United Kingdom

Jacqueline Rowe
School of Informatics, University of Edinburgh
Edinburgh, United Kingdom

Benedetta Cantanzariti
School of Informatics, University of Edinburgh
Edinburgh, United Kingdom

Fabio Tollon
School of Social and Political Science, University of Edinburgh
Edinburgh, United Kingdom

Osman Batur Ince
School of Informatics, University of Edinburgh
Edinburgh, United Kingdom

Cyndie Demeocq
School of Informatics, University of Edinburgh
Edinburgh, United Kingdom

Hannah Rohde
School of Informatics, University of Edinburgh
Edinburgh, United Kingdom

James Garforth
School of Informatics, University of Edinburgh
Edinburgh, United Kingdom

Abstract

Evolving industry and policy landscapes require CS graduates to navigate ethical frameworks in practice. We designed a workshop exploring how postgraduates conceptualise and operationalise "responsibility" in AI in Designing Responsible Natural Language Processing, we combine keyword ranking, structured reflection, and a scenario-based co-design task with a disruptive "rug-pull." Run across nine doctoral groups ($n = 33$), the intervention supports collaborative sense-making and translation of ethics principles into design decisions. We offer a blueprint including workshop structure, facilitation guidance, and reusable materials for CS ethics education.

Keywords

Responsible AI; CS Ethics Pedagogy; Applied AI Ethics; Workshop

ACM Reference Format:

Amanda Horzyk, Jacqueline Rowe, Benedetta Cantanzariti, Fabio Tollon, Osman Batur Ince, Cyndie Demeocq, Hannah Rohde, and James Garforth. 2018. Designing Responsible AI: A pedagogical blueprint for interdisciplinary reflection and applied ethics competencies. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Emerging AI-related risks and harms drive demand for ethical AI and more responsible research and development practices. Policy agendas and societal pressures prompt conversations about limited CS curricula developing students' critical ethical, social, and professional skills [7]. The evolving information technology industry requires broader professional competencies to prepare students for transitions from academia to employment [5, 9]. However, developing applied ethics skills remains challenging, as high-level ethics principles often lack operational specificity [2, 7]. Our contribution increases student engagement by emphasising ethics through real-world examples, critical reflection, and instructional frameworks that surface students' understanding of responsible AI [2, 6]. It also supports collaborative interdisciplinary problem-solving under realistic constraints. Our intervention aims to surface how students conceptualise responsibility in AI/NLP, translate abstract ethics principles into applied competencies, and foster situated design reasoning through a scenario-based interdisciplinary exercise.

2 Intervention

Developed within a Centre for Doctoral Training ('CDT') in Designing Responsible Natural Language Processing ('NLP'), the workshop was run across three cohorts of interdisciplinary doctoral students in NLP and machine learning systems, with 9 groups (n = 33). While originally piloted to investigate how CDT students conceptualize "responsibility" in AI/NLP, we observed that the intervention supported peer-to-peer learning, collaborative sense-making, critical AI literacy, and evaluation, suggesting relevance for CS ethics pedagogy. Small group sizes were intended to foster deeper reflection and participation from each participant. Facilitators were given a workshop script, slides, and supplementary materials.[1]

2.1 Workshop

Building on work in CS ethics and translational interdisciplinary approaches, the workshop promotes active, competency-based collaborative learning [2, 3, 6, 8, 9]. The design is guided by principles inspired by [5, p. 42]: system-based design thinking, risk and policy-aware planning, disruption-resilient ethical agility, and collaborative interdisciplinary sense-making. The intervention comprises (i) pre-work through an institutional survey and (ii) a three-part workshop activity. The survey explores how disciplinary backgrounds

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

[1] Workshop Supplementary Materials available: https://detailed-governor-d79.notion.site/Translating-Responsible-AI-Workshop-materials-and-guide-f0d3cf1647cf83c6aa9f01ec1a300976

Table 1: Student Feedback

Abbreviated Questions	Rating out of 5 (Mean ± SD)
The extent to which the workshop was rated as engaging (1 - not engaging, 5 - very engaging).	4.9 ± 0.3
The language used during the workshop activities was accessible.	4.4 ± 0.5
The workshop helped me to develop my understanding of core concepts on responsible AI research and practice.	3.8 ± 1.0
The workshop helped me to critically identify actionable steps to take in my research approach to be more responsible.	3.4 ± 0.5
The scenario-based component was useful to understand how I could translate abstract ethical frameworks in practice.	4.1 ± 0.9

Table 2: Overview of the workshop activities

Part I: Conceptual Warm-up (~10 minutes)	
Objective	Icebreaker introduces the workshop theme and explores which characteristics of "responsible" AI/NLP students consider important. Participants receive five keywords (cards printed), drawn from words most associated with responsible AI/NLP from the pre-work survey.
Learning Outcomes	Collaborative sense-making and communication in interdisciplinary teams; definition-alignment and agility when navigating changing technologies, languages, and frameworks.
Instruction	Students in 5 minutes collaboratively rank five responsible-AI-related terms in order of perceived importance. Blank cards allow students to introduce new terms. Cards are then flipped to reveal definitions drawn from academic papers and policy frameworks, with another 5-minute discussion to reconsider their ranking.
Rationale	Boundary objects help align perspectives on complex concepts [10].
Part II: Structured Reflection & Dialogue (~30 minutes)	
Objective	Students reflect on three questions: how they define "responsible" AI/NLP, what sources shape their understanding of the topic, and what experiences they have with responsible AI in practice.
Learning Outcomes	Professional communication and reflection on how assumptions, disciplinary backgrounds, and experiences shape judgement.
Instruction	Students first undertake 2 minutes of silent reflection per question before contributing to semi-structured group discussions lasting approximately 6–8 minutes each.
Rationale	Mitigates typical focus-group power imbalances by giving quieter participants time to formulate ideas before sharing [4, p. 455–456].
Part III: Scenario-Based Co-Design with Rug-Pull (~30 minutes)	
Objective	Activity moves from abstract ethical principles to practical implementation by asking students to design a responsible AI system under realistic constraints for a given scenario.
Learning Outcomes	Real-world, situated problem solving, agile interdisciplinary teamwork, ethical design thinking, and professional communication.
Instruction	Students role-played an AI development team designing a product according to a specified project scenario and brief. They are given 12 months and £10,000 to design the project strategy using a systems-design pipeline [1] defining objectives, methods, success metrics; practical steps, risks and harms and responsible AI measures. After 20 minutes, a "rug-pull" reduces the budget to £2,000 and timeline to three months, forcing teams to re-prioritise, cut or adapt aspirational responsible AI measures under new conditions.
Rationale	Scenarios are carefully selected to be relatable, realistic, and adaptable to students' backgrounds; Full scenario constraints and project goals and an AI project development pipeline PDF (see supplement).

and organisational cultures shape interpretations of responsible AI/NLP and prior ethics exposure. Students then participate in a 90-minute workshop involving three tasks: a conceptual warm-up using keyword ranking and definition alignment; structured silent reflection followed by group discussion; and a scenario-based co-design activity with a disruptive "rug-pull" element (**Table 2**).[2]

3 Student Feedback

Facilitators reported strong engagement across groups. In a survey of 10 students during the third iteration of the workshop, participants rated the activity highly for engagement, accessibility, and usefulness in developing a critical understanding of responsible AI research and practice (see **Table 1**). The scenario-based activity was most frequently identified as the most useful component (n=5), followed by definition alignment (n=2), and peer discussion and reflection (n=1). The "rug-pull" element was also highlighted as particularly valuable, prompting discussion on how responsibility priorities shift under realistic project pressures.

Suggested improvements. Some students suggested allowing more time and opportunities for cross-group discussion of project strategies. Others requested additional facilitator guidance on interpreting responsibility in practice, and one participant suggested simplifying the scenario task.

4 Conclusion

We present a workshop that frames responsible AI as a collaborative design challenge shaped by institutional and resource constraints. The intervention supports interdisciplinary reflection, encourages translation of ethical principles into design decisions, and helps students reason about socio-technical trade-offs across the AI development lifecycle. Future work will report and qualitatively analyse students' conceptualisations and practices of responsible AI/NLP.

Acknowledgments

This research is supported by UKRI AI Centre for Doctoral Training in Responsible and Trustworthy in-the-world Natural Language Processing grant (EP/Y030656/1) from UK Research and Innovation and the University of Edinburgh's School of Informatics.

References

[1] Jesmond J Allen and James J Chudley. 2012. *Smashing UX design: Foundations for designing online user experiences*. Vol. 34. John Wiley & Sons.
[2] Richard J Botting. 2005. Teaching and learning ethics in computer science: Walking the walk. *ACM SIGCSE Bulletin* 37, 1 (2005), 342–346.
[3] Alexi Brooks. 2024. Agile Ethics: A Low Stakes, Skills-based Framework for Teaching CS Ethics. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*. 492–498.
[4] Alan Bryman. 2016. *Social research methods*. Oxford university press.
[5] Tony Clear, Åsa Cajander, Alison Clear, Roger McDermott, Mats Daniels, Monica Divitini, Matthew Forshaw, Niklas Humble, Maria Kasinidou, Styliani Kleanthous, et al. 2025. AI integration in the IT professional workplace: a scoping review and interview study with implications for education and professional competencies. *2024 Working Group Reports on Innovation and Technology in Computer Science Education* (2025), 34–67.
[6] Amir Hedayati-Mehdiabadi. 2022. How do computer science students make decisions in ethical situations? Implications for teaching computing ethics based on a grounded theory study. *ACM Transactions on Computing Education (TOCE)* 22, 3 (2022), 1–24.
[7] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature machine intelligence* 1, 9 (2019), 389–399.
[8] James W Malazita and Korryn Resetar. 2019. Infrastructures of abstraction: how computer science education produces anti-political subjects. *Digital Creativity* 30, 4 (2019), 300–312.
[9] Inioluwa Deborah Raji, Morgan Klaus Scheuerman, and Razvan Amironesei. 2021. You can't sit with us: exclusionary pedagogy in AI ethics education. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 515–525.
[10] Susan Leigh Star and James R Griesemer. 1989. Institutional ecology,translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social studies of science* 19, 3 (1989), 387–420.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

[2]Study approved by the University of Edinburgh School of Informatics Ethics Committee Application number rt #8794, granted 25 February 2025; and rt #8794 follow up workshops, granted 30 July 2025