Abstract of "Using Contrasting Cases to Teach Socially Responsible Computing" by Yanyan Ren, Ph.D., Brown University, May 2024.

With the rise of machine learning and growing attention to issues of racial injustice in the USA, there is renewed energetic discussion about how to teach students about ethics and the social impacts of computing. Talks and papers on these projects largely focus on case studies and examples that can be included in assignments. My thesis instead takes a pedagogic perspective. Drawing on papers from various disciplines including math education and cognitive science, we explored various pedagogic approaches to teach socially responsible computing. One recurring challenge in our exploration of pedagogical approaches has been guiding students to move beyond surface-level responses when engaging with case studies. An intriguing strategy we have identified is contrasting cases, where students are presented with multiple distinct instances of a concept, each differing in deep features. We conducted a study to evaluate the effectiveness of employing contrasting cases in teaching socially responsible computing (SRC), aiming to explore how this approach helps students understand and analyze the nuances of SRC concepts.

Using Contrasting Cases to Teach Socially Responsible Computing

by
Yanyan Ren
B. A., Swarthmore College, 2018

A dissertation submitted in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy
in the Department of Computer Science at Brown University

Providence, Rhode Island
May 2024

This dissertation by Yanyan Ren is accepted in its present form by
the Department of Computer Science as satisfying the dissertation requirement
for the degree of Doctor of Philosophy.

Date

Kathi Fisler, Director

Recommended to the Graduate Council

Date

Shriram Krishnamurthi, Reader

Date

Julia Netter, Reader

Approved by the Graduate Council

Date

Thomas A. Lewis
Dean of the Graduate School

Curriculum Vitae

Yanyan Ren was born and raised in Nanjing, China to parents Yuhong Hou and Zhaojie Ren. She attended Nanjing Foreign Language School and later Swarthmore College. During her time at Swarthmore, she worked as a research assistant to Professor K. Ann Renninger. In 2018, after graduating with a Bachelor of Arts in Computer Science, she enrolled in the Ph.D. program of Computer Science at Brown University. She was advised by Professor Kathi Fisler and was part of the Programming Languages Team. While at Brown, she served as a co-organizer of the department's mentorship program for Ph.D. students, and also worked as a graduate advisor for the Socially Responsible Computing program.

Acknowledgements

My math teacher in high school, Mr. Yin, had a catchphrase: ”Don’t be afraid.” He claimed that this strategy would help one solve any math problem, no matter how hard it seems. I always tried to follow this advice, although I must admit, I struggled to apply it effectively, whether in tackling math problems or other challenges in life. There always seemed to be so many things to be afraid of and to worry about. Luckily, I had a lot of support to help me navigate through my fears.

I am grateful to my advisor, Kathi, for teaching me to engage with the uncertainty of research. Through numerous emails, Zoom meetings, and discussions in her office, I frequently voiced my concerns about the potential failures of my research. Each time, Kathi encouraged me to push through despite my fears and doubts about the outcome, often reminding me that ”it’s not research if we already know all the answers”. Kathi, your steady support and guidance have been invaluable throughout my Ph.D. journey. Thank you for everything.

I am grateful to my committee members, Shriram and Julia, for guiding me through my hesitations about venturing into new disciplines. They encouraged me to incorporate concepts from cognitive science and philosophy into my work, areas I initially felt unprepared and unqualified to explore. Their insightful feedback not only enriched my thesis, but also helped alleviate my fear about stepping into interdisciplinary realms, fostering a more interesting approach to my research.

I am grateful to my academic family for helping me get over the fear of struggling alone. Tim, Milda, Rob, Jack, Preston, Francis, KC, Elijah, Siddhartha, Skyler, Ben, and Will, thank you for being there whenever I needed support, both academically and emotionally, and for reminding me that we were all sharing this journey together.

I am grateful to my professors at Swarthmore for encouraging me to pursue my interest in research, and for urging me not to dwell on my fears about completing the program or achieving successful results. Ann, Aimee, Joe, Joshua, Kevin, and Zach, thank you for generously sharing your experiences and guidance when I first started graduate school, and for continuing to offer support and advice throughout my academic journey.

I am grateful to everyone who made the mentorship program possible. Nihal, Marilyn, Jasmine, Lauren, John, and all the participants in the mentorship program, thank you for creating this supportive space together, and for helping me realize I had the strength help others with their fears and struggle.

I am grateful to my teacher friends who showed me that good teachers can come in many forms. Ms. Li, Debbie, Ms. Elaine, Tom, Jeana, and Iris, thank you for helping me overcome my doubts, and for inspiring me to bravely step into the teaching profession.

I am grateful to my mentor, Jonathan, for helping me navigate through all kinds of life's problems and unknowns.

I am grateful to my long-term roommates, Qian, Zhenyu, Oreo(🐈‍⬛), and Scarlet(🐈‍⬛), for making sure there's never a dull moment in our home.

Last but not least, I cannot thank my family enough for supporting my decisions throughout this journey. They have been my constant source of reassurance that no matter what challenges or fears I face, we will always find solutions together. Perhaps I'll never stop feeling afraid, but with them by my side, I know I'll always have the courage to face whatever lies ahead.

Contents

[1] a significant part of this section also appeared in my published paper [46]

[2] a significant part of this section also appeared in my published paper [46]

List of Tables

List of Figures

Chapter 1

Introduction

Teaching about ethical and socially responsible computing has been around since 1970s. For example, the ImpactCS framework, introduced in the 1990s [34], proposed a way to situate ethical concerns simultaneously in technical and social contexts. More recently, prominent institutions like Harvard, Stanford, and MIT have all initiated programs in this field: Harvard piloted Embedded EthiCS in 2017 [22]; Stanford introduced Embedded Ethics program in 2020 [60]; and MIT debuted the Case Studies in Social and Ethical Responsibilities of Computing with its first issue in 2021 [27]. At Brown, we also have been running the Socially Responsible Computing (SRC) program since 2018 [14, 43].

A 2022 report from the National Academies on responsible computing [36] calls for embedding responsible computing throughout CS research, education, and training — a perspective we fully support through our SRC program. Our goal is to familiarize students with the common issues related to responsible computing, and cultivate their skills to identify and analyze these issues in software systems, applications and algorithms. We aim to instill in students the mindset that assessing social impacts should be as relevant as—and on par with—run-time analysis. Over the years of developing the SRC program, our journey has evolved from fundamental questions of "what to teach" to more refined inquiries emphasizing "how to teach it effectively".

Our primary research objective revolves around equipping students with the skills to conduct in-depth analysis in SRC. As we delve into the exploration of various pedagogical approaches, they have all led us to a common challenge, which has refined our research question to: *How can we guide students to move beyond superficial responses when engaging with case studies?* One approach that we found particularly promising is contrasting cases. This intervention strategy involves presenting students with multiple small examples of the same concept, but each differing in some deep features.

This approach prompts students to compare these cases and extract the underlying features, thereby fostering a deeper understanding of the concept [20]. Studies have shown that contrasting cases have been effective in teaching students in subjects such as math [61], physics [48], and writing [32]. However, this approach has not been previously utilized in the teaching of SRC. As a result, we aim to focus on exploring this innovative approach.

My dissertation explores the central research question ***In what ways does the use of contrasting cases help students understand and analyze nuances of SRC concepts?*** My dissertation describes our experience designing and calibrating different pedagogical approaches to teach SRC. We first provide a summary of prevalent methods for teaching computing ethics, and emphasize the ways in which our work both builds upon and distinguishes itself from them chapter 2. We then present an overview of our initial designs and findings from previous attempts, detailed in Chapter 3. Following this, Chapter 4 offers a comprehensive description of the design and development stages of our latest study, addressing the significant issues identified in earlier efforts. Finally, Chapter 5 discusses the implementation of the contrasting cases strategy and explores the results derived from this approach.

Chapter 2

Related Work[1]

The SRC program at Brown takes an embedded approach to integrate SRC content throughout the entire CS curriculum, instead of having a standalone ethics course. This ensures that all students, regardless of their chosen courses, will gain exposure to SRC principles. Our primary focus lies in enhancing the learning experience for a diverse spectrum of students. We've collaborated with the SRC team to explore the design of learning objectives, aiming to address the question: ”How can we effectively implement SRC content throughout the CS curriculum to enable students to build up skills?” Figure 2.1 illustrates the three-tier learning objective model that we developed, delineating the scale and progression of skills we aim for students to acquire.

1. **Awareness:** point out the existence of a range of social impacts
2. **Identification:** help students identify potential social impacts related to the technical content they are working with
3. **Analysis:** reflect more deeply on the nuances of those impacts and apply ideas and concepts from humanities and social sciences to think critically about the social context in which technology is embedded

Figure 2.1: Three-Tier Learning Objective

2.1 SRC Concepts

Before delving into effective methods of teaching SRC, it's important to first address what we mean by SRC concepts. The mission of Brown's SRC program is to ”not only familiarize future engineers with the ethical, and political challenges and social impacts of modern digital technology, but to

[1] a significant part of this section also appeared in my published paper [46]

provide them with the tools to reason critically about those challenges" [11]. Hence, we began with a literature review on responsible design in computing, enabling us to comprehend prevalent issues and discussions within the field of ethical computing and socially responsible computing.

Back in 2017, several members of the then-called FAT/ML community (now FAccT) authored a document entitled "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms" [17]. The principles include responsibility, explainability, accuracy, auditability, and fairness, with a note that principles for maintaining privacy and respecting the impact of human experimentation also be included. The document proposes guiding questions for each principle. We found some of these questions (from fairness and some from explainability) align well with the level of our target audience of CS students. However, a significant portion of the questions were aimed at more advanced or professional-level system designers (such as those on designing auditing and accountability systems). Selbst et al.'s [56] five fairness traps are similarly geared at those with a background in machine-learning algorithms.

The OECD has published a collection of privacy principles [37], which are also aimed more at professionals (considering security safeguards, legal issues, and disclosure, among others). While privacy issues are part of the SRC concepts, our emphasis differs. We focus on stakeholders' awareness and control of private information.

Values-based design is practiced in software engineering [67] and HCI [13]. Whittle et al. [68] proposed using the Schwartz basic-value theory (from social science) as a frame for soliciting software requirements. The theory raises concepts such as self-direction, power, achievement, and stimulation (among others) that might guide design decisions. Yu et al. [71] advocated modeling social relationships as part of software-requirements engineering. Social relationships are seen as part of the contextual environment of a system under design, which have long been included in software engineering practice [25]. We also see stakeholders and their relationship as integral components of SRC concepts.

Some software-engineering papers argue that sustainability should be included as a standard in software development [39], and some propose models for thinking about designing systems for sustainability (broadly-construed) [18]. The meaning of the term *sustainability* raises something of a play on words. The common environmental interpretation invokes the resources needed to keep large data centers operational; this in turn connects to questions of keeping data current. From a software-development perspective, this becomes an issue of maintainability: what are the costs to keep a system running smoothly over time. We decided that sustainability and maintainability suggest related questions, and hence should both be included in the SRC concepts.

Indeed, we have a range of additional issues and learning objectives that we aim to cover. The work summarized above represents only a portion of our comprehensive approach. Our long-term goal is to ensure that students develop a robust understanding of the various issues. Furthermore, we aim for students to master the skills necessary to proficiently identify and engage in discussions about SRC concepts across different contexts and in different software systems that they build or interact with.

2.2 Teaching SRC Concepts

Ethics and Social Impacts Frameworks Many have argued for the value of frameworks to aid in eliciting social impacts of systems. To improve AI ethics education, Raji et al. [45] recommended the use of "frameworks of intervention based around existing problems", and also the use of "a frequent analysis of concrete case studies". When introducing a sociotechnical perspective, Selbst et al. [56] argued that "reality is messy, but strong frameworks can help enable process and order, even if they cannot provide definitive solutions".

In an educational context, the ImpactCS framework, which dates from the 1990s [34], situates ethical concerns simultaneously in technical and social contexts. It provides a tabular framework in which rows are the level of social analysis (e.g., individuals, communities, and nations) and columns are ethical issues (e.g., privacy, use of power, and quality of life). They created one table for each of several technologies (e.g., artificial intelligence, medical technology, and computer-aided manufacturing). Table cells are boolean-valued, marked if a social entity might face the corresponding ethical issue within the domain captured by an individual table.

In 2006, a European Commission report [8] proposed an "Ethical Matrix" (as well as an "Ethical Delphi") tool to help professionals, government regulators, and the general public identify ethical issues in bio-technology. As in ImpactCS, the ethical matrix put "interest groups" in the rows and ethical principles (well-being, autonomy, and fairness) in the columns. The matrix provides a "structured approach for reflecting on competing ethical impacts" (page 22).

Drawing on frameworks for identifying security threats, we designed a social threat modeling framework (see section 3.1) for identifying a variety of social threats in the kinds of programs that university students develop in their first two years of study. Similar to ImpactCS and Ethical Matrix, our framework includes a table. However, our unique approach involves populating the table with guiding questions to help students think through concrete issues that might arise in individual cells. Another aspect that sets our framework apart is its centering around architectural components of applications, rather than stakeholders and values. Additionally, we have created a prototype of an automated tutor (see section 3.2) with a similar emphasis on architectural components and how the SRC issues surface in different decision points during system design.

Target Audience In 2011, Wright [69] proposed a framework for ethical assessment of information technologies. He provided a checklist centered around principles from biomedical ethics [7] (respect for autonomy, nonmaleficence, beneficence, and justice) with an additional section on privacy and data protection. Relevant social values and ethical issues were listed per principle (e.g. dignity, safety, accessibility, sustainability), along with corresponding guiding questions. Wright's target user was a "technology developer or policy-maker"; many guiding questions touched on risk assessments, and consultations with outside experts. While these are important in professional practice, they are not relevant for early-stage university students.

We have found a variety of frameworks and tools designed for a wide range of target audiences, including undergraduate students in an HCI AI Interaction course [57], master students in a data science program [6], and the general public [64]. However, we have not found any frameworks designed for teaching ethics to general CS students as early as introductory courses.

Class Activity Formats When reading about how others teach ethics, we noticed some patterns in the format of the class activities and homework mentioned. We saw many asking students to do case studies of existing systems, such as a walk-through of privacy options offered by Google [58], small-group discussion on Facebook and Cambridge Analytica's use of data [2], or assigned readings on YouTube's recommender system [59]. Some asked students to do small tweaks and thought experiments on the system they built themselves. For example, Saltz et al. [49] first asked students to implement a logistic regression using a Yelp review dataset, then had students do a thought experiment on how to re-adapt Yelp model for bank creditworthiness score.

2.3 Areas of Improvement

Instructor's Perspective: Enhancing Case Study Instruction While we found these case studies and repositories to be inspiring and helpful since they showcased a diverse range of scenarios and various issues, they primarily function as experience reports detailing someone else's success stories. They often lack guidance on how to effectively teach with these cases and, furthermore, how to generate contemporary and tailored cases that would be current and cater to our students. In our designs, we also made use of case studies, and our focus was on the instructions and pedagogic design around the cases. Our exploration involved using drawing to break down SRC concepts like consent (see section 3.4) and using contrasting cases to encourage students to notice and discuss nuances (see section 3.3 and section 3.5).

Student's Perspective: Lessons from Harvard's Embedded EthiCS Harvard's Embedded EthiCS program, a collaboration between the computer science and philosophy departments, has garnered attention for integrating ethical reasoning into CS curricula [22, 28]. However, student reporters' investigation [30, 66] suggest areas of improvement based on interviews conducted with participants and teaching fellows. Firstly, there's a perception that the program's ethical content sometimes takes a backseat to technical material, leading to a perceived simplicity and a disconnect from the core course content. Additionally, some students expressed feeling that the concepts taught were somewhat watered down and overly obvious, with comments like, "we all know the correct answer here." Finally, some students find the philosophical perspective abstract and unrelated to their coursework. Learning from these challenges, our pedagogic explorations prioritize technical design perspectives, aiming to integrate SRC concepts as integral components of software systems.

2.4 Other Related Work

While this section has provided an overview of relevant literature, it is important to note that additional literature reviews are woven into subsequent sections of this paper. These sections provide deeper insights into various theories and works that underpin our pedagogical designs. Specifically, section 3.1 discusses references related to threat modeling; section 3.2.1 presents teaching model constructions through an automated tutor; section 3.3.1 goes over the theory of contrasting cases, and Section 3.4.1 addresses ontology.

Chapter 3

Prior Prototypes and Designs

Over the last three years, we have explored multiple designs for activities to teach students to think critically about SRC. Each design has been inspired by others' work on teaching people to do a technical learning or analysis task, but not in the context of SRC. For almost every design, we ran a pilot study to assess strengths and weaknesses of the approach and our study setup. Our findings inspired the next design. The changes from one design to the next have sometimes been substantial rather than incremental. Our sequence of prior prototypes motivates how we have come to do a more in-depth study of contrasting cases as a means for teaching nuances in SRC.

3.1 Social Threat Modeling Framework[1]

Drawing on papers on ethics-based design from multiple computing disciplines, as well as frameworks for identifying security threats, we designed a framework for identifying a variety of social threats in the kinds of programs that university students develop in their first two years of study.

3.1.1 Threat Modeling and Adversarial Thinking

Systematic methods to identify vulnerabilities in computing systems have been around since the 1970s. One well-known architecture-based approach, STRIDE [23], was developed at Microsoft in the 1990s to help developers without security training avoid certain pitfalls. With STRIDE, one examines each architectural component (e.g., databases, processors, services) against six threat types (e.g., Spoofing identity, Tampering with data) and identifies concrete vulnerabilities that arise

[1] a significant part of this section also appeared in my published paper [46]

around that component. The component architecture can then be revised as needed (e.g., adding a firewall). The notion of “architecture” in an SRC context is not about databases and processors, but rather the “components” that students see in the software applications they are writing, such as humans, data, and algorithms.

The cybersecurity community talks about a “security mindset” and “adversarial thinking” as important parts of identifying potential threats to systems [16, 51, 21]. These approaches involve adopting the perspective of a hacker who is trying to attack a system. While adversarial thinking might help a student identify some kinds of social threats, many adverse social impacts arise from second-order impacts that do not result from malicious intent. Young and Krishnamurthi [70] assessed adversarial thinking on socially-responsible computing assignments in an introductory CS courses. Their rubric categorizes whether students run experiments, speculate on system design, or take others’ perspectives when trying to identify social threats on various systems. Our work differs in trying to propose a framework of guiding issues and questions, rather than techniques for actually identifying threats.

We thought the checklist-style approach to analysis from STRIDE could work well for SRC. More importantly, we thought the term ”threat modeling” might better appeal to CS students. Anecdotally, we have heard students stating that they do not consider SRC to be relevant to their computing careers. The word “ethics” often suggests “philosophy”, which is not what job-seeking students associate with CS. In addition, early-stage students can perceive ethics as being about values which in turn are either right or wrong; early-stage students tend to believe that there are authorities that hold the right answer, and won’t rebel against authority and learn to incorporate personal experience until later years in college [41]. Thus, a framing of values and ethics might not be the most effective for early-stage CS students. In contrast, “threat modeling” is a term from computing security. While early-stage undergraduates won’t already know the term, it sounds more technical and the connection to security can give it credibility. Furthermore, “threats”—like “responsible computing”—has broader connotations than “ethics”, which may also help engage CS students and faculty.

3.1.2 Framework Design

Our social threat modeling framework (shown in fig. 3.1) walks through the most common social threats categories that are of the complexity suitable for students in their first two years of college-level CS.

Instructions: Use the following questions to determine the socio-technical context of the program or system, then complete the table *separately for each identified stakeholder*.

- What problem is the program (being) designed to solve? Is the program replacing an existing (manual or computational) process? Focus on the problem at the level of individuals, populations, or organizations.
- List the individuals and organizations who will interact with or be impacted by the program (the *stakeholders*). For each one, state their relationship to the program. (e.g., job applicants will submit data to and be evaluated by the hiring software, hiring managers will use output from the program to decide who to interview)
- What relationships (if any) exist among stakeholders that should be preserved or changed as a result of this new system? These could be positive relationships or conflicts of interest.

	Data	**Agency**	**Algorithm**
Fairness and Inclusivity	In what ways might data used for decision-making within the system be biased or unfair to the stakeholder? Are the data representative of the population? Are the data trustworthy? Can the data structures within the system represent everyone in the target group for this class of stakeholder?	What should the stakeholder be able to do or control in the system? Do all potential members of this stakeholder class have similar rights and privileges? Who has the power or privilege to manipulate the data or actions that are relevant to this class of stakeholder?	What outcome would make the algorithm fair or unfair for different individual stakeholders? Does the algorithm produce different outputs for two stakeholders in this class who should be considered similar? Can the algorithm be gamed to the advantage of some stakeholders over others?
Privacy and Reputation	Are data that pertain to this stakeholder class protected from leakage and tampering? What can the system infer about an individual stakeholder based on the data available to the system? What tradeoffs does the system make between privacy and its services on behalf of the stakeholder?	Did the stakeholder willingly or knowingly provide their data? Did they know how it would be used? Can an individual stakeholder in this class have rights to view, modify, or remove data about them (whether provided, computed, or inferred)? Does an individual stakeholder have the ability to adjust outcomes of the system that could affect their reputation? What impact would such actions have on the intended purpose of the system?	Is the algorithm transparent, meaning that the reason for a specific outcome on a specific individual can be inspected by an appropriate person?
Sustainability and Maintainability	Do critical data about individual stakeholders need to be kept current? How much energy does it take to collect, store, and maintain the data?	How does the system impact an individual stakeholder's ability to thrive, exercise their rights, and develop freely? How does the system impact the relationships between individuals and their community, including trust, communication, and participation?	Does the system have the ability to maintain and evolve the system over time? Would there be any technical debt when scaling? How much energy does the algorithm consume as a function of data size?

Figure 3.1: Social Threat Modeling Framework

Axes of the Framework Our framework, shown in fig. 3.1, arose from a process of reading and benchmarking. The rows are key themes emerged from a literature review on responsible design across multiple subfields of CS.

- From the machine-learning community, issues of fairness and bias, especially at the hand of algorithms
- From the security community, issues of privacy and controlling access to systems or data
- From the software-engineering community, identification of stakeholders and issues of long-term maintenance, including (environmental) sustainability

We initially set out to create a STRIDE-like 1-dimensional framework with a simple list of concrete issues and guiding questions. We quickly realized, however, that STRIDE's simple structure rested on the assumption that developers knew the system architecture. Early-stage CS students are still learning to identify communicating components in their programs and how components directly or indirectly interact with social contexts. We thus chose to highlight (as table columns) three components of modern software applications that aligned with the threats we saw in the literature:

- **Data:** information that the system uses, collects, or manages about entities (whether human or organizational)
- **Algorithm:** how the system generates or processes the data
- **Agency:** what rights entities have regarding the handling of data or the processing by the algorithm

While users are also part of the architecture of a socio-technical system, we put users in a third dimension about *stakeholders*. This supports the variety in both types of stakeholders (individual and institutional) and their relationships, goals, and values to certain classes of threats. This is consistent with other social impact frameworks that treat stakeholders as their own dimension.

While these are not conventional architectural components, they are core pieces of early-stage software projects in our department ("agency" is a generalization over concepts like "access control").[2] We refer to these components as architecture even though they are higher-level and cross-cut more conventional notions of an architecture based on physical or software-as-service components.

[2] We do not list the user-interface among these components because our department doesn't cover interfaces in the early-level courses.

While users are also part of the architecture of a socio-technical system, we put users in a third dimension about *stakeholders* (see the paragraph below titled Bringing in Stakeholders). This supports the variety in both types of stakeholders (individual and institutional) and their relationships, goals, and values to certain classes of threats. This is consistent with other social impact frameworks that treat stakeholders as their own dimension.

We refined our theme (row) descriptions by working through several concrete case studies, including a discussion of threats in AirBnB [18], information about problems with Yelp [42], and problems in the design of YouTube [15]. The issues raised in these papers helped us see that threat topics often had both *individual and collective angles*. For example, one might seek to protect one's reputation (with others) in addition to one's privacy. An individual decision might be fair or unfair, but interactions across decisions might affect an entire subgroup, leading to exclusivity. We thus refined the *fairness* and *privacy* themes to include the more collective concepts of *inclusivity* and *reputation*, respectively. *Inclusivity* also covers software-development guidelines such as the "falsehoods programmers believe" series [35, 47] about representations in data.

The *sustainability* theme raises something of a play on words. The common environmental interpretation invokes the resources needed to keep large data centers operational; this in turn connects to questions of keeping data current. From a software-development perspective, this becomes an issue of *maintainability*: what are the costs to keep a system running smoothly over time. We decided that sustainability and maintainability suggest related questions, and hence we grouped them together as a theme in our framework.

Bringing in Stakeholders The framework includes general instructions on *identifying stakeholders* (above the table in fig. 3.1). The instructions prompt thinking about the various users of a system, as well as non-users who might be impacted by it. The latter clause creates something of a chicken-and-egg problem: the framework should help students identify threats, but also asks them to think about potentially impacted users (which will be associated with threats). The data we collected (see section 3.1.3) indicates (a) that students do raise some threats based on being asked to identify stakeholders, and (b) there are systematic categories of stakeholders that we could suggest to help students perform better at this task.

Guiding Questions Inside each table cell are the guiding questions to help students identify specific threats. We developed the questions with two goals in mind. First, we wanted to expand on the meanings of the terms in the themes. Second, we wanted to reflect on specific threats that we had read about in the literature or seen raised in case studies. Several of the questions were influenced by security threat modeling, such as protecting data from leakage and tampering and stakeholder rights and privileges. Some arose from well-acknowledged issues such as whether data are representative and system transparency. Others arose from consultations with faculty colleagues outside of computer science, such as rights to thrive and develop freely.

3.1.3 Evaluations

We tested our framework in two offerings of a data-structures and algorithms course. The first offering had 150 students: most were sophomores or juniors, a handful were seniors, and roughly a quarter were first-years. The second offering had 112 students, nearly all first-years with a handful of sophomores. Both offerings were conducted online (due to, and in the midst of, COVID-19), taught by the same instructor. Students came from three different first-semester courses that had discussed responsible computing to some, yet differing, extent. None of the first semester courses had presented a framework or other systematic presentation of threat categories, or even framed responsible computing from the perspective of identifying threats.

Baseline without the Framework

During the first two days of the course, students completed a background survey about prior computing experience, self-assessment of skills, and interests in computing. One section gathered baseline data on how students think about socially-responsible computing. One question asked about students' criteria for fairness:

> *Consider a restaurant review website (like Yelp, TripAdvisor, or Dazhong Dianping). What would it mean for the system to be "fair"? (yes, this is open-ended – we want to see how you think about this as we get started)*

Common responses concerned checking the legitimacy of reviews (data), preventing restaurants from paying to influence which results get displayed (agency), allowing restaurants to respond to but not change posted reviews (agency), and collecting reviews at different times from different populations of customers (policy regarding data). Many of these issues are external to the application itself: this is good for having students think in terms of social context, but also suggestive that students need guidance to think of how technical decisions contribute to these problems. Very few students mentioned impacts on the surrounding community or how to keep the reviews current (a form of sustainability), both questions that the framework questions would prompt them to consider.

Identifying Stakeholders and Threats

A week and a half into the course, students were asked to read or watch danah boyd's [sic] "Be Careful What You Code For" presentation [15] about environmental and social consequences of consumer technologies and broad questions people need to be asking about them. Students in the second offering were asked to choose one of several papers (on topics such as privacy, health insurance matching, predictive policing, etc) and comment on which of boyd's [sic] general questions the developers failed to account for adequately. We provided the table axes (components and themes), but students were neither required to use nor limited to using them in answering.

Observations Across the paper options, roughly 75% of students identified users who (perhaps unknowingly) provided data (e.g., browsing history), and users who access data provided by apps (e.g.,ads). Roughly half named the platform providers (e.g., Google) who stood to profit from gathered data. Just under half identified individuals or groups who would be directly displaced by new tech or harmed by gathered data (e.g., subjects of search results). More subtle stakeholders (identified by just a few students) included those who could build new products from the technology, competitors who might lose business, supplies of underlying components (e.g., hardware or energy companies), third-parties who had promoted similar technologies (e.g., thought leaders), and the earth (in the case of bitcoin). These observations are consistent with those of Prioleau et al. [44], who found that "students often associated technologies with the vulnerable populations that are often victimized in those domains, as seen on the news and media."

During the process of summarizing categories, we detected some subtleties that we hadn't previously observed when looking at student work (on other assignments). We had originally identified "data producer" and "data consumer" categories, but students' responses to the ProPublica article on recidivism [5] led us to add additional explanation to distinguish the kinds of data involved: primary (e.g., judges, who use the system to render decisions) and secondary (e.g., defendants, whose personal information are also provided to the system). The concept of data providers also highlights how the same category of stakeholder can occupy rather different roles in different systems: data are provided knowingly in some systems and unknowingly in others.

We observed cases in which students seemed to identify each of threats and stakeholders from the other. Writing about cryptocurrency, for example, one student wrote about "regular people who are set to feel the impacts of climate change and strain on energy systems." The same student later adopted the perspective of a different stakeholder: "For people who mine cryptocurrency, these oversights add an enormous energy bill to their costs, but in many cases these costs are significantly outweighed by the profits from mining cryptocurrency."

Identified Threats To get a sense of which threats students considered without seeing the cell questions, we coded the threats that students listed in terms of the cell labels (a component-topic pair). Regardless of which article they read, students spoke about fairness with respect to data and algorithms. Fairly few raised agency-based concerns within fairness, even though such responses were (a) plausible for some articles and (b) well-represented in the survey-question data (so the general idea was within scope for them).

Regarding privacy, students talked about collection and leakage of private data, as well as consent (agency component). They also discussed the role of algorithm transparency in helping users maintain and understand privacy settings. No student raised comments related to the question about how a system trades off privacy and services, or about the potential for a system to infer private information (both questions in our framework). In the sustainability topic, students did not reflect much on sustaining relationships between users and their communities. Students did comment on environmental sustainability, which is not surprising given the article they read about that as part of the background survey.

These observations suggest that framework questions target issues that most students do not naturally identify on their own. While students raised some issues that were not covered by our framework, they were all impacts on broader economic systems that are beyond the scope of content in an early-stage CS course.

3.1.4 Lessons Learned

The data analysis has demonstrated the potential of the framework in providing systematic scaffolding questions to break down the analysis process. Yet, while the table and guiding questions provide a valuable way to break down the SRC analysis process, the framework is primarily a reference tool rather than a comprehensive solution. We learned that merely providing the framework to students does not ensure a deeper or more comprehensive analysis.

Another limitation of the framework became apparent during the refinement process. As we refined our themes (row) descriptions by working through several concrete case studies and mapping the related issues to the framework table, we came to the realization that certain issues were highly context-dependent, making it challenging to neatly categorize them within the framework. For example, in the Yelp case, the mere existence of the online platform intensifies competition among the local restaurants, resulting in unfairness and damage to the community's well-being. While one could argue for fitting and expanding on these issues within the framework's fairness-agency and sustainability-agency cells, it's also evident that the current guiding questions are too generic to assist individuals who might not naturally consider these issues on their own.

These findings underscore the need for ongoing development and implementation of suitable pedagogical approaches to help students utilize the tool effectively.

3.2 Automated Tutor and Design Principles

To explore other ways to present the guiding questions, we drew inspiration from a graduate seminar on tutoring systems tailored for programming language concepts. This prompted us to consider the feasibility of developing an automated tutor that could aid students in navigating SRC concepts. We began to explore what foundational elements such a tutor would require and whether these needs could be met within the context of SRC.

A tutoring system, as described by VanLehn [62], involves "giving students an electronic form, natural language dialogue, simulated instrument panel, or other user interfacethat allows them to enter the steps required for solving the problem". One example of this concept is implemented by VanLehn et al. [63] to teach algebraic model constructions. They pinpointed four critical skills of model construction. Essentially, these skills involve the ability to deconstruct the problem, identify the relevant components, then match the problem to the mathematical relationships and equations [63].

3.2.1 Prototype Design

We wondered about the potential of designing a tutor with similar underlying logic to teach SRC. Unlike algebra, where there are formal mathematical relationships such as the distance-rate-time schema, we don't have an established set of ground truths with regard to SRC. However, principles relevant to design decisions in SRC can provide a solid basis for a tutoring system. For instance, principles concerning the preservation of privacy in software systems, the manifestation of transparency throughout system components, and the evaluation of mathematical formulas to measure algorithmic fairness. These underlying principles, much like mathematical rules, guide the making of design decisions in SRC.

To clarify these principles, consider the following examples:

- Principle 1: Privacy issues can surface from processes that infer data, highlighting the need for careful data management.

- Principle 2: When an algorithm functions as an opaque "black box," it becomes difficult to identify and correct errors, stressing the importance of transparency.

- Principle 3: Sole reliance on mathematical formulas can neglect social subtleties and lead to unintended consequences, a dilemma referred to as ”the Formalism Trap,” as identified by Selbst et al. [56].

Our end goal with the tutor is to help students master the skills of identifying the issues (the second level skill in the three-tier learning objective as described in fig. 2.1) and applying the design principles. Specifically, we want to ensure that when presented with a scenario, students are capable of discerning the relevant themes of issues. Additionally, for each identified theme, students should be able to pinpoint the pertinent parts of the system and adeptly apply the corresponding design principles.

On a high level, we envision the tutor providing students with a scenario to consider, asking students to identify the tools, algorithms, or decisions that they see embedded in a solution to the scenario, and then asking students to assess potential threats associated with those decisions in that context. Through this sequence of questions, students gain a deeper understanding of how the SRC issues surface in different decision points during system design. Our long-term vision was for different courses to utilize the tutor with different scenarios, all showcasing the same set of design principles. In the end, students ideally would have a better grasp of the design principles by repeatedly encountering and applying them to different systems.

We developed a prototype of the tutor using YouTube as an example. Below is the scenario that would be presented to students when they start the tutor:

> **YouTube Recommendation System**: Imagine that you've been asked to design an application to recommend videos to users based on initial search terms. A user would enter a search query, and the system would return a sequence of related videos based on the query. The system would show the videos in order until the user stops scrolling down or closes the web page.

After reading the scenario, students would be guided to answer a series of questions in both free-response and multiple-choice format. The tutor would immediately process some of the responses and automatically follow up with additional questions based on the response. Below are some example questions:

1. What potential social threats might lie within this scenario? (free response)

2. Which of the following issue(s) did your answer touch on? (multiple choice)

   □ Bias

   □ Privacy

□ Transparency

□ Accessibility

□ Sustainability

3. I see that you didn't check "privacy". Do you think any of the following situations might arise? (multiple choice)

   □ The user may not be willingly or knowingly provide their data.

   □ The user doesn't understand how their data will be used.

   □ The system may collect personal data in order to provide more personalized content.

4. I see that you checked "transparency". Who get impacted by this issue? Please select from below (multiple choice) and write a few sentences to explain (free response)

   □ Viewers

   □ Video uploader

   □ YouTube

   □ Ad sponsors

   □ News media

   □ Other

3.2.2 Feedback and Lessons Learned

We presented the prototype to our lab, and received valuable feedback. One prominent theme of the feedback was on students' diverse perspectives. Students with different experiences will see the same issues in different ways. For example, in the YouTube case mentioned above, a student who owns a YouTube channel and spends a lot of time creating content might have a distinct understanding of how the recommendation system works, compared to a student who primarily watches shows on TV. Students will also have very different vocabulary when interpreting the same issue. For example, what falls under the umbrella of the term "accessibility"? Consequently, this variance in perspective and vocabulary presents a significant challenge when attempting to provide automated feedback that respects and accommodates the diverse experiences of students.

Another theme of the feedback centered around the generality of the design principles. To ensure the tutor's effectiveness in different scenarios, it's crucial to establish principles that would stand on their own, and remain relevant across various contexts and scenarios. However, during the drafting of these principles, we found it necessary to provide some context to students, in order to effectively illustrate our point. Below is an example principle related to crowdsourcing with its potential pros and cons. The examples included within the parentheses in the sub-principles were crucial for fully demonstrating the meanings of the terms and showcasing the varied applications of crowdsourcing across different systems, making the principles more tangible and understandable.

> **The criteria used for determining the final results displayed to the users can either be crowdsourced (impacted by individual users) or be preset (perhaps by experts).**:
>
> 1. Crowdsource outputs can be timely (e.g. trending on Twitter), but could also be taken advantage of by bots.
> 2. Crowdsource can reflect collective wisdom (e.g. Wikipedia edits), but the crowd is not always right, and may collectively produce false information.
> 3. Crowdsource can spark conversation (e.g. Stack Overflow replies), but the conversation might get dominated by certain groups.

The challenge of designing a set of design principles for the tutor was similar to the challenge of refining the guiding questions in the threat modeling framework. In both instances, our initial goal was to establish a comprehensive set of references (be it design principles or guiding questions) for students to consult when they work on various scenarios. However, as we advanced, it became evident that to delve deeply into the intricacies of each scenario and encompass its interesting facets, a one-size-fits-all approach with a general set was inadequate. We shouldn't solely depend on a general set of references and expect seamless applicability across all contexts. Instead, we needed to develop context-specific scaffolding.

3.3 First Iteration of Contrasting Cases

Upon reflecting on the development process of the framework's axes and the creation of prompts in the automated tutor, it became clear that they share a common learning objective: identification. In the framework, students are directed to utilize the table to identify where the issues arise within the cells. In the tutor, students are asked to first identify the relevant issues before being presented with the applicable design principles. Now, let's consider the next steps. Once students have successfully identified the related issues within a given scenario, what types of support and scaffolding can we offer to inspire them to delve deeper and critically analyze the social impacts of these issues?

3.3.1 Contrasting Cases

Contrasting cases is an instructional design arising from perceptual psychology. A contrasting-cases problem gives learners multiple artifacts to inspect that have notable similarities as well as carefully-designed differences. The underlying theory is that seeing the similarities will also draw the learner's attention to the differences. According to Alfieri et al.'s meta-review on case comparison's effects on learning [3], case comparison exercises have been shown effective in a variety of contexts and types of implementations. They also discovered the intervention to be effective for learners with differing levels of experience. Indeed, we were able to find a diverse set of papers on successful contrasting case interventions. In a history class, Lin et al.'s study [32] demonstrated that presenting students with well-written and poorly written stories side-by-side (an example is shown in fig. 3.3 enhanced their understanding of crucial elements in effective story writing, such as a clear main thesis and teaching important lessons. In a math class, Lampert [29] showed students multiple solutions to the same math problem to prompt them to consider the efficiency of the methods. In a physics class, Salehi et al. [48] provided students with different scenarios about buoyancy, enabling them to notice important features related to buoyancy and invent rules encompassing features like mass, density, and volume. Thus, despite the absence of prior literature regarding the use of contrasting cases in teaching SRC, we were motivated by the extensive array of successful applications and see potential in adopting this approach.

Contrasting cases leverages both perceptual and discovery learning, as outlined by Professor Daniel Schwartz in his research talk [53]. Schwartz emphasizes the goal of perceptual learning is to "educate perception" and to "help people see the world better". He argues that without the ability to discern the details, students cannot effectively recognize and apply relevant knowledge. Consequently, contrasting cases are employed to highlight features in such a way that helps students more clearly discern differences that might have previously been overlooked.

Furthermore, contrasting cases utilize the advantages of discovery learning by avoiding what Schwartz terms "telling too early," which can result in a narrow focus on mimicking teacher-provided answers instead of encouraging independent exploration. A study conducted by Bonawitz et al. [9] illustrates this point. Their research found that children given a toy with multiple functions and no instructions explored more features than those who were demonstrated a specific function before receiving the toy, underscoring the value of discovery.

Contrasting cases have proven effective in teaching STEM subjects like math and physics. For instance, fig. 3.2 displays a worksheet used in a physics class designed to teach the concept of density. Students were asked to invent an index to describe the crowdedness of clowns in a bus, which subtly introduced them to the ratio structure of density. The researchers carefully designed the case studies to highlight both the deep structure, such as proportional ratios, and surface-level features, such as the clowns' clothing which had nothing to do with density. The findings revealed that students exposed to contrasting cases were more attuned to the deep structures of the problems compared to those in the control group, who received traditional instruction that simply explained how ratios function. Furthermore, the study indicated a significant advantage in terms of knowledge transfer. When faced with a new scenario, these students were four times more likely to successfully utilize their understanding of ratios gained from the clown exercise, compared to their peers who received conventional instruction.

Applying contrasting cases to teaching socially responsible computing presents unique challenges, given the abstract and complex nature of SRC concepts, which do not lend themselves to straightforward visual representation like the density example in fig. 3.2. In our approach, we aim to adapt the principles of contrasting cases to SRC education by developing case studies that demonstrate how these concepts manifest differently across various scenarios. Rather than merely teaching definitions and the importance of SRC concepts, these case studies are designed to encourage students to engage deeply with the material and uncover underlying structures and nuances on their own.

In addition to the visual format, contrasting cases can also be used in textual format. Figure 3.3 presents an example of this approach with a pair of stories—a good one and a bad one—handed out to students in a history class to teach them how to assess and write good stories. A week prior to the study, students were introduced to six criteria that define a good story: "(1) having a clear main thesis; (2) offering detailed examples that illustrate daily life, especially of children; (3) bringing historical facts to life; (4) presenting events and characters in a logical and connected manner; (5) teaching important lessons; and (6) raising questions for further inquiry" [32]. After distributing the stories, the teacher facilitated a class discussion to help students identify how each of these criteria was applied or neglected in the examples provided. Questions such as "Rubric criteria 1 states that the stories should have a main thesis. Do you think that the example stories have a main thesis? If so, how did the author achieve this?" were posed to guide the students. The study found that this method of highlighting the strengths and weaknesses of each criterion helped students gain a deeper understanding of what makes a narrative effective and improved their ability to self-assess their own writing.

This narrative approach to contrasting cases is intriguing, as the contrasts are established more subtly compared to the visual example in fig. 3.2. In that figure, students can quickly spot differences in the number of clowns per bus, but the stories in fig. 3.3 demand a more detailed analysis. Although both stories were designed to be similar in length and structure, a simple scan does not

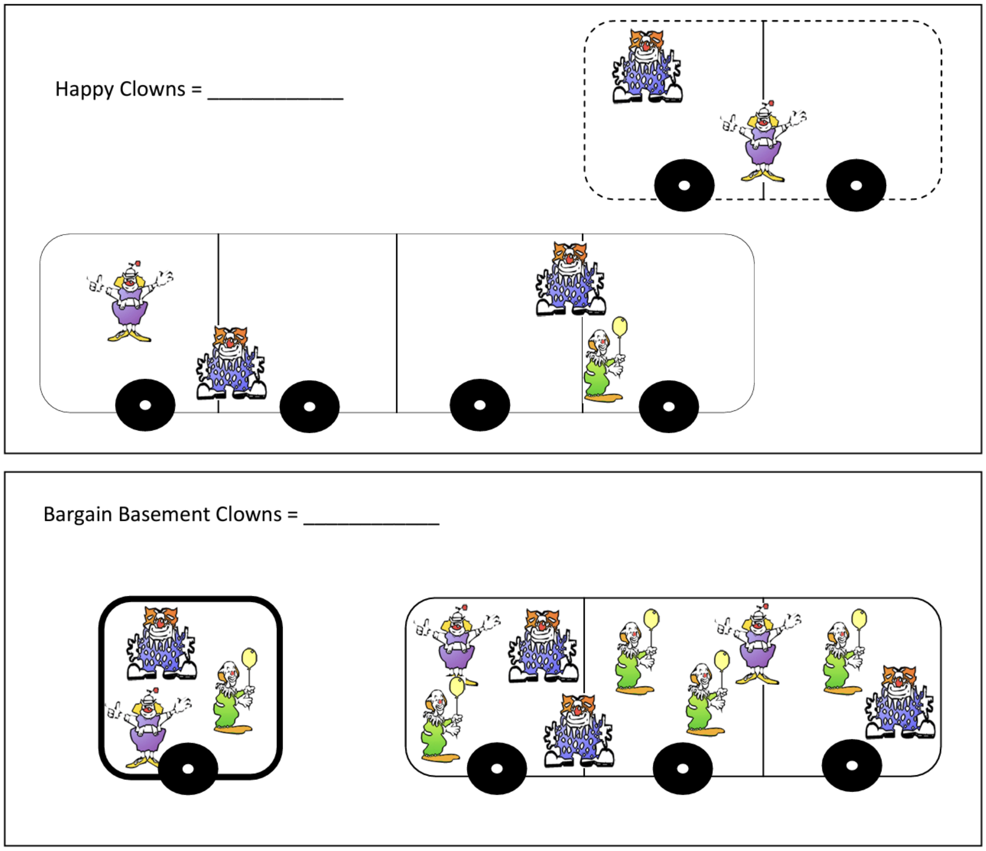

Happy Clowns = ____________
Bargain Basement Clowns = ____________

Figure 3.2: Contrasting Cases to Teach Density in a Physics Class[54]

suffice; students need to carefully evaluate each story against the six established criteria to fully grasp the nuances. This example illustrates how contrasting cases can be effectively used to teach more abstract and complex ideas, such as ”making historical facts come alive”. This has provided us with both confidence and inspiration to develop text-based cases for teaching socially responsible computing concepts.

3.3.2 Assignment Design

During the fall semester of 2022, we crafted an SRC assignment for the introductory course CS111, with the aid of Lizzie Kumar, the graduate teaching assistant of the course. The assignment followed a lab discussion format, where students began by reading two cases focused on the issue of targeted

An example of a well-written story:
Hello! My name is Sunia. I am 11 years old. I have three older brothers and a younger sister. I live in Tilonia. My village is on the banks of the Ganges River—a good place for farmers like my family to live because there is water even during the dry season.

We are not rich people. My parents, aunts and uncles get up very early every morning to get water from the town well and to milk the cows. Then they get the plow ready to go out to the fields. My brothers work during the day and go to school at night, but I do not go to school at all because I am a girl. A girl's proper place is at home, doing domestic work. This is because many people in my village feel that the benefits of a girl's education will be enjoyed by others, since a daughter, typically, leaves her family after marriage. I listen to the grownups talk about the news my uncles bring from the city. One day they saw a very respected man, Mahatma Ghandi. Ghandi is a strong believer in Hinduism. He was giving a speech about kicking the British out of India. People say we would be better off if we ran our own government. Ghandi said, "We can achieve our independence from the British without a war—without weapons and without hurting anyone. Non-violence is our weapon". "That's impossible", my uncle said. "He must be crazy to think the British will give up India without a fight". I hope Gandhi's right about not having a war—it scares me to think about a war right here, in my own village. Ghandi also thinks that we need to make some big changes in our society after we have our own government. He says we should stop discrimination against the untouchables—the families who for centuries have had the nastiest jobs. He says: "They are the Children of God". But other people think the untouchables are only able to do the dirty jobs that no one else wants.

I wonder what will happen in my country: Will people believe Ghandi and follow his lead? Why and why not? How can we win our independence from Britain without a violent revolution? Could the untouchables really be treated the same as the rest of us? It's an exciting time to be in India.

An example of a poorly-written story:
Hello! My name is Sunia. I am 11 years old. I have three older brothers and a younger sister. I have lots of uncles and aunts. I live in Tilonia, a small village in India. My village is on the banks of the Ganges River—a good place for farmers like my family to live.

I stay home helping my parents with household chores. We grow our own food, taking care of cows, etc. The cows are holy animals that cannot be harmed. My parents are very busy every day. Since I stay home most of the time during the day, I also listen to the grownups talk about the news my uncles bring from the city. One day they saw a very respected man, Mahatma Ghandi. He was born and raised in India and went to college in London. He later became a spiritual and political leader in India. He launched a movement of non-violent resistance to the Great Britain's ruling. Gandhi's political and spiritual hold on India was so great that the British rulers dared not to interfere with him. That day, he was giving a speech to a large crowed about his spiritual and political views: India should run itself and should be independent from the British ruling. However, we should earn the independence peacefully, not violently. People say that we would be better off if we ran our own government. I hope we do not have a war—it scares me to think about a war right here, in my own village.

My parents, aunts and uncles get up very early every morning to get water from the town well and to milk the cows. Then they get the plow ready to go out to the fields. My uncles go into the city to sell the milk at the market. We kids take care of the cows, water buffaloes, and goats then go out to the jungle to find food. My brothers go to evening school because they have to work in the daytime.

I think there is a lot going on in my country right now, but since we live here in the village, we don't hear much about it. The village just got its first radio a few weeks ago. I wonder what will happen to my country and my family.

Figure 3.3: Contrasting Cases to Teach Story Writing in a History Class [32]

Targeted Ads Case 1: Dr. Z, a doctor in a medium-size city, wants to test a new treatment protocol for drug addiction in people ages 18-30. Dr. Z. is particularly attentive to recruiting a set of participants that are representative of the city's residents. A wealthy local donor has offered modest initial funding, under the condition that Dr. Z. provide monthly updates on how the study is going. Dr. Z. is considering several ways to recruit participants, including putting ads on the public city buses, putting ads in the local newspaper (digital and print editions), and posting ads through the social media apps that are popular in his city. The available budget will only allow for one of these forms of advertising.

Targeted Ads Case 2: Ms. G. has an apartment to rent out. As someone who doesn't sleep well, she is wary of noise, so she prefers to rent the apartment to someone who is at least in their 40s, has at least started college, and has quieter musical tastes (like classical and jazz). She posts an ad on social media, using the ad-posting options to target people with her desired age, education, and musical constraints. Despite the affordable rent that she was offering, Ms. G was a bit surprised that everyone who contacted her about the apartment seemed to be a manager, doctor, or lawyer, with little racial diversity among the applicants.

Figure 3.4: Targeted Ads Case Studies

ads (see complete cases in fig. 3.4). Afterward, students were divided into small groups to engage in discussions and provide responses to a set of open-ended, free-response questions using a Google Form. Some questions were designed to prompt students to notice and elaborate on the details in individual cases, such as ”What are the strengths and weaknesses of each form of advertising in Scenario 1?” and ”What aspects of the limited applicant pool in Scenario 2 seemed due to how the ad software works?”. Other questions subtly encouraged students to draw connections between the two cases: ”What are some of the factors that help determine whether targeted advertising is a positive or negative choice in a specific scenario?”

3.3.3 Evaluations

After conducting an initial round of qualitative data coding, we identified two primary issues that prompted us to pause and reflect. First, some responses were very brief. For example, the last question was designed to serve as a summary question, prompting students to list out all the interesting factors they observed across two cases: ”What are some of the factors that help determine whether targeted advertising is a positive or negative choice in a specific scenario?” One student gave a succinct response without referring to any details in the cases: ”Who is being impacted, what is trying to be conveyed, how invasive it is.” Evaluating their ability to identify factors from the cases became challenging. Was this brevity due to a reluctance to reference case details, possibly stemming from laziness or deeming the cases irrelevant to the questions? Unfortunately, since we only had access to the students’ final submissions and were not physically present during classroom discussions, we cannot definitively determine the reasons behind this.

Another issue we observed was that students sometimes focused on questioning the legitimacy of the case setup, instead of focusing on the nuances of the issue. For example, in response to the question ”What aspects of the limited applicant pool in Scenario 2 seemed due to policies or regulations around online advertising?”, we expected students to talk about the human decisions regarding criteria used for targeted ads. We aimed to prod students to think about the issue of proxy variables, where even if sensitive criteria like race and ethnicity are not used, there might be other criteria acting as proxies to target specific racial groups. However, one student expressed confusion, stating ”I don’t get why these specific qualities are not considered to attract a diverse population. It feels like the question is using stereotypes to say these qualities are not going to inherently attract a diverse population.”

3.3.4 Lessons Learned

The primary challenge encountered in the study was the lack of detailed responses, which hindered our ability to conduct further analysis to understand students’ performance in identifying and analyzing nuances. To address this challenge, we should consider two avenues of exploration:

1. Revising the instructions provided to students to encourage more comprehensive and elaborate responses (further explored in section 3.4).

2. Finding ways to pose follow-up questions to students and gain insight into their thought processes (further explored in section 3.5).

3.4 Ontology Drawing

Drawing from insights gained in the previous study iteration section 3.3.3, a notable difficulty encountered with the free response assignment format lies in students frequently providing weak responses—either overly brief or heavily paraphrased from the assigned reading materials. To address this issue, we set out to find alternatives to free response. We explored the use of ontology and drawing, hoping to push students to engage deeper with the materials. Our hypothesis was that while it’s easy for students to compose brief responses (especially now in the era of ChatGPT),

Consent Case 1: At Hangzhou Wildlife Park in China, the entrance process for members is about to get an upgrade. When purchasing annual membership cards, members let the park collect their fingerprints and take their pictures. The members also signed a contract to voluntarily opt in or opt out of using their fingerprints for quick entrance to the park. This year, the park's management team decided to upgrade to using facial recognition (instead of fingerprints), arguing that facial recognition would help move the crowd faster during the peak seasons. The management team believes this will be a quick and smooth transition, since the members' pictures are already collected, and they've found a local third-party startup specializing in building facial recognition systems.

Consent Case 2: Because of Covid, lots of universities needed to offer online courses and test-taking. Thus, anti-cheating software became popular. Some software can detect keystrokes and collect feeds from a computer's camera and microphone, or even allow a person monitoring an exam to take control of students' devices. A university in Ohio adopted an anti-cheating software that prompted students to take a virtual scan of their room before taking the test. A link to terms and services is provided to students, and they need to click "yes" before proceeding to turn on the camera.

Figure 3.5: Consent Case Studies

creating a drawing that effectively represents some details of the issue would require more thoughts and efforts.

3.4.1 Drawing Design

Before proceeding to draft an assignment and asking students to create drawings, we needed to try out the idea on some concrete examples to evaluate the viability of using drawings to capture enough interesting details in a given scenario. We started off with two case studies both centered around the issue of consent. The complete cases are shown in fig. 3.5. Case 1 is adapted from an issue from the MIT Case Studies [10], and case 2 is adapted from a news article [24].

After reviewing several theoretical definitions of ontology, Jepsen [26] came up with a practical interpretation: "An ontology is a method of representing items of knowledge (ideas, facts, things—whatever) in a way that defines the relationships and classifications of concepts within a specified domain of knowledge". We viewed this as a useful lens to break down SRC concepts in a systematic way. Consequently, we set out to find existing ontology around consent, hoping to find a comprehensive ontology that covers all aspects of consent, providing us with the possible entities and relationships to build the drawings upon.

We came across several papers on informed consent, such as the work by Lin et al. [31], which proposes an informed consent ontology that covers various aspects within the context of the consent process in a clinical trial. Additionally, We encountered several papers focusing on the legal aspects of consent. For example, Pandit et al.'s GConsent ontology [38] analyzes and models information associated with consent under the General Data Protection Regulation (GDPR). Not surprisingly, none of these existing ontologies seems to align well with our target audience—introductory-level students who may not be familiar with the general idea of consent, let alone the terminology and details related to clinical trials or GDPR. Nevertheless, we did find GConsent [38] inspiring. The authors presented several graphs that broke down the extensive ontology, such as core ontology concepts, context of consent, and status of consent. With these graphs as a starting point, we managed to select a subset of elements and incorporate some of our own to outline the interesting features in the cases shown in fig. 3.5.

After numerous iterations, we arrived at the drawings shown in fig. 3.6 and fig. 3.7. The black-colored boxes and edges roughly correspond to core ontology concepts from GConsent, which "describe the essential entities and their relationship with consent" [38]. Within our drawings, the context of consent from GConsent was distinguished by two distinct colors: the purple segments pertain to the environmental attributes such as location and medium, while the green segments pertain to time-related attributes such as expiry and duration. Furthermore, the red segments in fig. 3.7 represent the consequences associated with refusing consent.

3.4.2 Evaluations

Our drawings display a high level of complexity, filled with abundant edges and a variety of colors. Nevertheless, there remain interesting nuances and questions that could be asked about the cases that weren't fully captured by the drawing.

For case 1:

- Members only agreed to provide their picture for the membership card creation (and probably for the occasional check by the park's staff), but now their picture is used as part of the training data for the facial recognition system.

- Can the park just notify members about the switch, or do they need to obtain new consent?

For case 2:

- Is a room scan necessary or legal for exam taking in the first place?

- The Power dynamics between the school and the students.

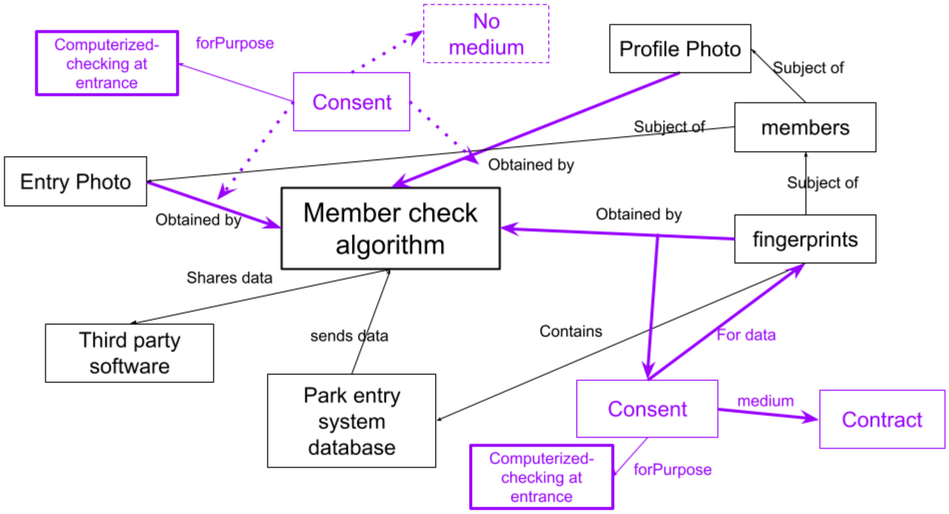


Figure 3.6: Ontology drawing for Case 1.

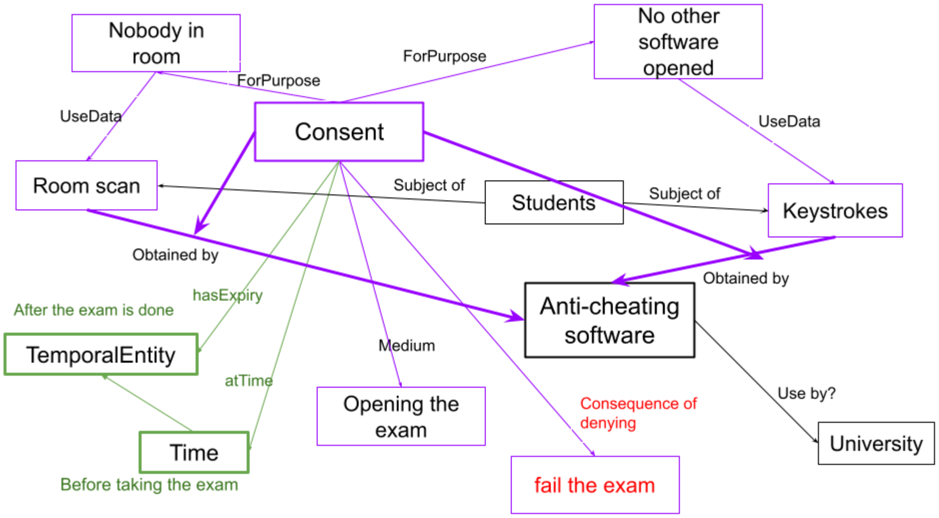


Figure 3.7: Ontology drawing for Case 2.

1. Who is the consent about?
2. Who is the consent given to?
3. What type of personal data are associated with the consent?
4. What type of purposes are associated with the consent?
5. What type of processing is associated with the consent?
6. How was the consent acquired/changed/created/invalidated?
7. What is the medium associated with consent?
8. What is the expiry of the consent?
9. Is the purpose or processing associated with a third party?
10. What is the role played by the third party in the purpose or processing?

Figure 3.8: Consent Subquestions Breakdown

- Rational ignorance: the students may not read/understand the consent form because they have to say yes to take the exam.
- Potential threats in the future: the school or the software could use the data later in a different context.

3.4.3 Lessons Learned

While we were dissatisfied with how our drawings turned out, the process of creating them helped us recognize that our fundamental goal was to dissect the issue into various sub-questions. Thus, we reviewed our drawings and compiled a list of questions, as shown in fig. 3.8, intended to facilitate the breakdown of the issue of consent (including some questions directly lifted from GConsent's competency questions table [38]). This strategy of dissecting the core issue into subquestions proved highly beneficial in our subsequent iterations of the contrasting cases study.

3.5 Second Iteration of Contrasting Cases

As outlined in section 3.3.4, we aimed to explore two different approaches to improve on our contrasting cases study and elicit more detailed responses from students. The first is to revise the instructions provided to students, and the second is to find ways to understand students' thought processes. We made an attempt at the first approach with ontology drawing. Now, we are looking into the second approach, and we conducted the second iteration of contrasting cases study by transitioning from collecting students' responses via Google Form to conducting Zoom interviews.

3.5.1 Study Design

During the summer of 2023, we ran a think-aloud study with 6 undergraduate students. They were all participants of a summer undergrad research experience program in the artificial intelligence lab at Brown, and were all computer science majors from other universities. Since these students did not take any CS classes at Brown, they weren't exposed to any of our SRC program. Therefore we consider them to have roughly the same knowledge and skills in the subject area compared to the students in the intro class in the last study iteration. The study took the format of 30-minute one-on-one interviews on Zoom, where students were asked to read two cases on misinformation. Case 1 discussed the challenges of identifying misinformation in the video format on TikTok, while case 2 talked about bots and human accounts spreading misinformation at the same rate on Twitter. The complete cases can be seen in fig. 3.9. After finishing reading the cases, students were prompted to compare and contrast the cases, and talk about various aspects of misinformation and content moderation policy. Compared to the first iteration where assignments were submitted through a Google Form, the interview format here gave us a better understanding of students' thought processes, and we were able to ask follow-up questions when students brought up new ideas. The complete interview script can been found in fig. 3.10.

3.5.2 Evaluations

This iteration served as a pilot to assess the effectiveness of the Zoom format and to gauge students' responses to the contrasting cases approach. Consequently, the evaluations are divided into two components: understanding student performance and identifying areas for enhancement in our study design.

Student Performance After an initial round of data analysis using thematic coding (see appendix A, we found that students were able to identify and discuss a wide range of nuances around misinformation and content moderation policy, such as:

- factors mentioned in the TikTok case that make detecting misinformation difficult
  - rich media format

Misinformation Case 1: The micro-video format of TikTok makes it more difficult to detect deceptive information. [Study author] Lundy said. ”The information is passed through such rich media objects—you have sound, visuals, text, body language, captions, and meme elements that require context, and all these factors interact at once to create the ’meaning’ or (mis)information that is being shared.”

Lundy used [...] methods of searching for ”community language” rather than expected terms, to get a much more representative and useful picture of how misinformation looks on the platform. According to Lundy, the incredible reactiveness of TikTok’s algorithms’ collaborative filtering poses a particular challenge to containing the infodemic. ”The more misinformation you interact with, the more that you see—you can quickly find yourself immersed in massive numbers of TikTok videos relating to COVID-19 vaccine misinformation just after liking a few videos,” she said.

Lundy learned that TikTok users who oppose the COVID-19 vaccine use intentionally coded language, misspelled words, and alternate hashtags to evade anti-misinformation efforts. She found that misinformation topics featured in previous COVID-19 vaccine hesitancy literature—parodies of vaccine side effects, concerns about vaccine production and approval, conspiracies about governments and vaccine contents, and claims that COVID-19 is not dangerous—are still prevalent despite public health efforts. Her research illustrated how COVID-19 vaccine misinformation often appears in the form of logical fallacies, where some information may be true but misleads to false conclusions.

Misinformation Case 2: From Russian ”bots” to charges of fake news, headlines are awash in stories about dubious information going viral. You might think that bots—automated systems that can share information online—are to blame. But a new study shows that people are the prime culprits when it comes to the propagation of misinformation through social networks. And they’re good at it, too: Tweets containing falsehoods reach 1500 people on Twitter six times faster than truthful tweets, the research reveals.

Bots are so new that we don’t have a clear sense of what they’re doing and how big of an impact they’re making, says Shawn Dorius, a social scientist at Iowa State University [...]. We generally think that bots distort the types of information that reaches the public, but—in this study at least—they don’t seem to be skewing the headlines toward false news, he notes. They propagated true and false news roughly equally. [...]

[Researchers] compare[d] the spread of news that had been verified as true with the spread of stories shown to be false. They found that whereas the truth rarely reached more than 1000 Twitter users, the most pernicious false news stories [...] routinely reached well over 10,000 people. False news propagated faster and wider for all forms of news—but the problem was particularly evident for political news, the team reports today in Science. At first the researchers thought that bots might be responsible, so they used sophisticated bot-detection technology to remove social media shares generated by bots. But the results didn’t change: False news still spread at roughly the same rate and to the same number of people. By default, that meant that human beings were responsible for the virality of false news.

[...] That got the scientists thinking about the people involved. It occurred to them that Twitter users who spread false news might have more followers. But that turned out to be a dead end: Those people had fewer followers, not more. Finally the team decided to look more closely at the tweets themselves. As it turned out, tweets containing false information were more novel—they contained new information that a Twitter user hadn’t seen before—than those containing true information. And they elicited different emotional reactions, with people expressing greater surprise and disgust. That novelty and emotional charge seem to be what’s generating more retweets.

Figure 3.9: Misinformation Case Studies, Summer 2023

Consent form and overview
I have received your consent form. Do you have any questions about it? Is it okay if I start recording now? I'll send you a Google Doc with instructions on it. Feel free to write on it, or talk out loud, whichever you prefer. There's no correct answer. The study is not meant to be a test on how you perform on the task. It's a pilot study because we are still fine tuning the material. So if there's anything that dosen't make sense to you, or if you have any questions, let me know! Anything you say will be super helpful!

Scenario and instructions
You are working at a social media company's content moderation policy team (some content would not be shown to at least some users). You are reading the case study to understand the topic of misinformation. What are some factors that should be taken into account when defining the policy? Which ones seem particularly interesting about this case?
Feel free to highlight or comment on the case.

Questions after students finish reading case 1

1. What are some factors that should be taken into account when defining the policy?
2. Which ones seem particularly interesting about the case you saw?
3. What role might each of human and algorithmic decision making play in content moderation?
4. What are the issues or challenges to using each of these modes?

Question after students finish reading case 2 How does seeing this second case study change your responses to the first question? You can add, remove, or refine parts of your answers.

Key factors For which ones do the two cases offer interesting and different nuances? Are there any other interesting nuances in either of the cases?

1. What kind of information is potentially being moderated?
2. What kind of harm can be caused?
3. How and how easily does the information spread?
4. How easy is it to verify the accuracy of information?
5. Who or what should decide whether a specific piece of content needs to be moderated?
6. Should content be moderated before or after being shown to users?
7. What knowledge or expertise is needed to determine which pieces of information should be moderated?

Summary

1. What has been your experience using TikTok and Twitter? Have you seen any misinformation mentioned in the case studies?
2. Did these cases bring up anything new that you didn't know before? (about misinformation, content moderation, the social media platforms, etc.)

Figure 3.10: Interview Script, Summer 2023

  - the use of community language and slang
  - logical fallacies of using true information but misleading to false conclusions
  - intentional misspelling of certain keywords

- differences between the two cases
  - how the format of information and content are different on two platforms (e.g. text

    based tweets vs. short videos on TikTok; Twitter has more political debates)
  - how the recommendation algorithm and the sharing/spreading models work differently (e.g. Tweets get pushed to the followers' timeline vs. trending videos get pushed to users' feed based on their viewing history)
  - how the user bases are different (e.g. younger audience on TikTok; academic community use Twitter)

- factors and strategies to consider when designing and implementing content moderation policies
  - prioritizing popular posts
  - using automated tools to flag potential misinformation and then passing it to human fact-checkers
  - getting a group of people with diverse backgrounds and expertise to check the content in order to reduce bias

Case Design When writing up the cases, we had concerns about the potential difficulty in comparing cases due to their distinct narratives. The cases also contained an abundance of intricate details. Leveraging the insights gained from the attempt at creating the ontology drawings section 3.4.3, we made a list of factors that help break down the issues. During the study, we showed students the factors and posed the question "For which ones do the two cases offer interesting and different nuances?". We thought the list of factors touched on most of the interesting details in the two cases, making it a useful tool to guide students in dissecting the cases.

Following the study, we reviewed the cases by outlining the manifestation of key factors in each case. The table shown in table 3.1 lists these factors and summarizes how they showed up in the two cases. Note that certain cells were marked as "not mentioned", implying that the corresponding factor wasn't directly mentioned in the case studies, although for some of these factors, we expected students to infer from the cases and make arguments related the factor. While there are some interesting comparisons to be made between the two cases, and sometimes students were able to provide additional examples themselves (such as comparing the young users on TikTok to the academic users on Twitter), we must acknowledge that the cases did not align closely enough to fully illustrate the key factors. Since helping students identify and analyze the key factors are the main objective of the exercise, and such alignment of key factor and cases is a crucial part of contrasting cases design, we realized we would need to follow a different design process for the next study.

3.5.3 Lessons Learned

Crafting Cases For the first iteration of the contrasting cases study, we crafted two scenarios. However, rather than delving into a discussion about the nuances of the central topic on targeted ads, some students were preoccupied with challenging the authenticity of the case setup (see section 3.3.3). To resolve this issue, we considered using real-world events instead of crafted scenarios. We believed that this approach would enable us to swiftly substantiate the case setup or clarify any queries students might have by referring to these real-world events. Consequently, for the second iteration of the study, we were very driven to shape the cases on real-world instances. However, real-world events are inherently complex, and we encountered significant difficulty in locating a pair of such examples that exhibited a similar structure while effectively illustrating the same set of key features, as depicted in table 3.1.

Drawing upon our insights gained from both iterations, our conclusion is that in order to harness the advantages of both realms — the authenticity of real-world examples and the controlled structure in crafted cases — it's important to start with gathering a collection of real-world examples, followed by extracting the interesting features and details, and subsequently crafting new cases by incorporating these details to ensure the cases look realistic.

Preparing for Students' Reactions Despite structuring the interview script for us to ask questions and students to answer, there were instances when students sought clarification by asking questions such as, "What does content moderation mean?", "Is content moderation the same thing as blocking?", and "What does automated tools mean?". In these situations, it's helpful to give students some concrete examples. However, it's also equally important to strike a balance and avoid excessive explanation, since we'd like students to explore the nuances on their own. We learned that responding effectively to these clarifying questions on the spot was quite challenging, highlighting the importance of anticipating and preparing for such inquiries in advance.

Considering Similarities Upon reviewing the interview script (see fig. 3.10), we realized that we didn't structure the prompt effectively to guide students to fully leverage the potential of comparing contrasting cases. When instructing students to compare the two cases, we prompted them to discuss the differences between the two cases after presenting them with a list of key factors: "For which ones do the two cases offer interesting and different nuances?" There were no specific instructions regarding identifying similarities. In retrospect, we also failed to consider similarities when setting up the cases, as evident in table 3.1 with the unbalanced breakdown of key factors, making the comparison process even more challenging for students.

Considering Students' Background Knowledge When reviewing the interview transcript, we observed varying levels of background knowledge among students. Some inquired about the meaning of content moderation and expressed surprise at the existence of such policies in social media companies. Conversely, others demonstrated familiarity with multiple platforms' practices. For instance, one student said ”As far as I know, no big company uses AI for content moderation. When Facebook doesn't, Twitter doesn't. [...] Instagram is community based. [...] Users can report on content. They also work with independent fact-checkers, so I would say Instagram system is the best one.” While we appreciated the students' enthusiasm and their engagement in deeper conversations, we recognized the importance of considering the influence of differing background knowledge to comprehensively understand the effect of contrasting cases.

Key Factors	**Case 1**	**Case 2**
What kind of information is potentially being moderated?	Content: covid-19 vaccine misinformation. Format: micro-video, including sound, image, caption, body language.	Content: fake news, particularly political news. Format: mostly text, sometimes images.
What kind of harm can be caused?	Undermines public health efforts to get more people to get vaccinated.	Not mentioned.
How and how easily does the information spread?	Not mentioned.	Bots and human users both spread fake news, roughly at the same rate. Novel information spreads faster among human users.
How easy is it to verify the accuracy of information?	Very hard. Users have figured out various ways to around the rules by using "community language".	Not mentioned.
Who or what should decide whether a specific piece of content needs to be moderated?	Not mentioned.	Not mentioned, but we'd expect students to talk about the platform taking responsibility for doing account verification and taking out bots spreading misinformation.
Should content be moderated before or after being shown to users?	Not mentioned, but we'd expect students to talk about the impact of letting the vaccine misinformation spread, and the difficulty of verifying new information and keeping the database up to date.	Not mentioned, but we'd expect students to talk about that misinformation spreads faster than truthful content, and the difficulty of verifying novel tweets.
What knowledge or expertise is needed to determine which pieces of information should be moderated?	Not mentioned, but we'd expect students to mention factoring in authoritative opinions from scientists and expert in covid, in addition to TikTok's own team.	Not mentioned.

Table 3.1: Case Studies Review: Post-Study Key Factors Breakdown

Chapter 4

Proposal for a Larger Scale Study on Contrasting Cases

We wish to draw on the various lessons learned through our previous prototypes to design and conduct a more detailed assessment of contrasting cases for use in SRC. Our overarching research question, as stated in the Introduction, is ***In what ways does the use of contrasting cases help students understand and analyze nuances of SRC concepts?***. This section serves to document the evolution of our study design, beginning with our initial plan and subsequently addressing the adjustments made based on various feedback along the way.

4.1 Research Questions

To answer the central research question, we need to first examine the quality of the students' responses, as a threat to validity check, leading us to

RQ. *1 Do students produce beyond shallow answers when doing the contrasting cases exercise?*

According to contrasting cases theory (as discussed in section 3.3.1), when students are prompted to articulate similarities and differences among a set of thoughtfully crafted cases, they should perform better at identifying key features. Therefore, if we can establish that students do produce high-quality responses when using contrasting cases, we can proceed to address the other part of the central research question regarding how contrasting cases helps. However, if the responses we obtain are predominantly shallow, with students failing to mention the key features, we may need to consider three courses of action: 1) refining the study protocol and conducting a reiteration of the study, 2) employing a filtering process to exclude low-quality responses for further analysis, or 3) revising our central research question to explore the reasons why contrasting cases may not be effective for SRC.

When examining the effect of contrasting cases, we want to include students' background in the topic as a variable in our analysis, which brings us to

RQ. *1.1 How does the level of background knowledge influence the quality of students' answers?*

To account for varying levels of background knowledge, we plan to ask students about their background, so we can analyze the results while controlling for this factor. Should we conclude that students with richer background knowledge exhibit better performance, it would imply that our study primarily assesses the influence of background knowledge levels rather than the effectiveness of contrasting cases. In such a scenario, we would be compelled to redesign the cases for a different subject matter and conduct a subsequent iteration of the study.

Another variable we want to consider is students' prior SRC experience:

RQ. *1.2 How does prior SRC experience influence the quality of students' answers?*

We could potentially gain valuable understandings regarding whether contrasting cases is more suitable for students with a certain experience level. This might also suggest that employing contrasting cases to craft exercises could be beneficial for both introductory and advanced courses, or possibly for one level in particular.

After checking the quality of students' responses, we are now ready to examine the effects of contrasting cases:

RQ. *2 In what ways do students delve deeper when prompted to use techniques from contrasting cases?*

The long-term effect is another aspect of evaluating the effects. Ideally, we would run some longitudinal studies over the years, such as from freshman year to senior year, in order to observe the outcomes of retention [19](assessing how effectively students can recall their knowledge to use in the future), and transfer [40] (assessing how effectively students can apply their knowledge in different situations). However, due to the time constraints within the scope of this dissertation, our primary focus lies in taking the initial step to investigate the potential for long-term effects following students' completion of the contrasting cases exercise. Therefore, we have formulated the following research question:

RQ. *3 How well can students apply their knowledge and skills learned to a different scenario?*

4.2 Study Design

For this study, our initial focus is on the topic of content moderation, crafting a specific set of cases around it. However, from the outset, we envisioned a broader plan to extend our investigation to other SRC concepts, aiming to deepen our understanding of the effects of contrasting cases through subsequent expansions.

Drawing upon the insights gained from the last zoom study section 3.5.3, it is evident that crafted cases work better than relying solely on direct citation from real-world events. Before we can start crafting cases, we should first gather a range of real-world examples and also establish a curated list of key features that we intend to highlight through the cases.

One valuable resource we utilized was The Santa Clara Principles On Transparency and Accountability in Content Moderation [1]. Crafted by "a group of human rights organizations, advocates, and academic experts", these principles aimed to provide guidelines for companies "to obtain meaningful transparency and accountability" in their content moderation practices. Its open consultation report contains quotes from various companies and organizations, offering us a rich source of real-world examples related to content moderation. Additionally, the principles were broken down into smaller, more detailed sub-principles. After reviewing the sub-principles, we picked a subset of the high-level things that we found would be most relevant and understandable to our students who might not have much experience with the subject, and dropped the bits that would make more sense for guiding the company to set up the correct mechanism (e.g. what data needs to be collected to ensure transparency).

The pair of cases we crafted are shown in fig. 4.1. At the time, we also planned to create a third case, in order to address RQ 3, which aimed to measure students' ability to transfer the knowledge acquired from the contrasting cases exercise to a new scenario. This plan was later changed as we made adjustments to the overall study. However, at the time of the proposal, this third case was included as part of our data analysis plan.

Making use of the insights obtained from the last iterations in section 3.5.3, we reviewed the two cases by constructing a key factors breakdown table (shown **??**) in to ensure that the cases outlines the key factors clearly for easy comparisons. Every row in the table showcases a key factor that we extracted from The Santa Clara Principles [1]. Each cell within the table contains excerpts from the cases that demonstrate the key factors. It's important to note that some table cells remain empty, indicating the absence of certain key factors in those cases. This intentional design allows us to examine scenarios where the contrast is between "presence vs. absence" rather than "two variations of presence." Some of the absent factors might still be distinctive enough for students to recognize.

Content Moderation Case 1: Annie is a content moderator at BuzzBuzz, a US-based social media platform. The company's automated content-checking tools removed a post and, following company policy, flagged it for Annie to review. The post in question was written by a grandmother who affectionately called her grandchild with dark skin a "black diamond". Annie assumes (but can't be entirely sure) that the automated tool flagged the post as being racist, but notices that the grandmother is from Bangladesh, which doesn't share the American concept of race. Annie wishes to reinstate the post, but her American boss prefers to err on the side of caution and upholds the deletion decision.

Content Moderation Case 2: Julie is on the content moderation team at FooChat, a large social media platform. Her job is to help maintain the machine-learning tool (model) that underlies their content-moderation algorithms. The model reviews and flags potentially violent posts (a mark that is visible to users of the platform), collecting them into a designated repository. Additionally, the model will automatically downrank the flagged posts based on their severity. Julie's team then analyzes this repository, identifying patterns to adjust parameters in the model, and enhancing the model's accuracy in identifying similar posts in the future. FooChat's customer-service team forwards a request from a user who is trying to understand why their posts appear with a "potentially offensive" mark. Julie uses the internal exploration feature on the content-moderation tool and discovers that some of the posts use language that suggests potential violence to women. This information is communicated back to the requesting user, who can choose to edit the post content to eliminate the mark.

Figure 4.1: Content Moderation Case Studies (Proposal Stage)

Key Factors	**Case 1**	**Case 2**
Use of automated processes	The company's automated content checking tools [...] flagged it for Annie to review	The model reviews and flags potentially violent posts
Downranking vs. removal	removed a post	the model will automatically downrank the flagged posts based on their severity
User awareness of the action taken		a mark that is visible to users of the platform
Cultural context	the grandmother is from Bangladesh, which doesn't share the American concept of race her American boss	
Explainability	Annie assumes (but can't be entirely sure)	Julie uses the internal exploration feature on the content-moderation tool

Table 4.1: Pre-Study Key Factors Breakdown

4.2.1 Refinements Based on the Last Iteration

Based on insights gained from the last Zoom study section 3.5.3, we refined the interview prompts, and the full script is shown in fig. 4.2.

Providing Example Responses We understand that the talk-aloud study and contrasting cases format might be unfamiliar to some students. It's important to set clear expectations at the beginning of the study to ensure that we collect high-quality responses. Therefore, before showing students the case studies, we will provide example responses to encourage detailed and thoughtful responses throughout the study.

Similarities then Differences As mentioned in section 3.5.3, we need to improve the prompts by taking into account similarities. In their meta-analytic review of implementing case comparison activities in various learning environments, Alfieri et al. [3] argued for the importance of starting with similarities. Their theory was that focusing on differences might cause a higher cognitive load, making it difficult for students to extract the critical features, possibly leading them to focus on superficial features that aren't relevant. Therefore, they recommended that "it might be prudent for instructors to begin with the objective of finding the similarities between cases before shifting attention to the differences and how the cases exemplify other, more nuanced categories." Accordingly, in our interview script, when students are assigned two cases, we would guide them to do the contrasting cases exercise by talking about similarities first and then moving on to differences.

Background Knowledge Ideally, we would like to directly inquire where students applied their background knowledge during the study. However, we recognize that this may be an impractical expectation as students may not remember all the specific details. Therefore, we have crafted several questions centered on students' experiences with social media and content moderation tools.

Preparing for Clarifying Questions The final part of the interview script doesn't consist of the prompts meant for students. Instead, it comprises a compilation of expected clarifying questions along with their corresponding answers that we'll have ready in case they're needed. We'll continuously update this list as we encounter additional clarifying questions during the study. As discussed in section 3.5.3, formulating a well-thought-out answer in the moment can be challenging and may disclose too much information. Hence, having this curated list allows us to effectively steer students in the correct direction without prematurely revealing certain nuances that we aim for students to uncover on their own.

- **General instructions and expectations**

  This is a think-aloud study. It's important that you verbalize your thoughts in real-time as you work through the questions. And it's important to express all your thoughts, even if you find them irrelevant or unimportant.

  Please substantiate your answers with as many details from the case or other examples as you can think of. For example, if you notice an issue related to bias, avoid providing simple answers like "the system could be biased". Instead, provide thorough answers with details such as "I noticed that the dataset mentioned in case 1 only contains data from English-speaking countries, potentially introducing bias. In case 2, the dataset has a very small sample size, which could also introduce bias but in a different way."

- **Questions about the case(s)**

  1. What issues do you notice at a high level?
  2. *Additional Questions for students who get two cases:*
     - What similarities do you notice between the two cases?
     - What differences do you notice between the two cases?
  3. What are some variables or factors that should be considered when deciding how to implement content moderation for an organization like a social media company?
  4. Is there any additional information or insights you would like to add regarding the topic of content moderation?

- **Questions about background knowledge**

  1. Do you use social media?
  2. How much have you previously read about content moderation?
  3. Have you previously run into issues (general or personal) with content moderation (on social media or elsewhere)? We don't need personal details, but we're curious about the nature of your experience as it might have influenced your responses to this activity.
  4. Have you ever been involved in building content moderation tools?
  5. Have you lived in a country other than the US for longer than 2 months?

- **A new case**

  1. What issues do you notice at a high level?
  2. What are some variables or factors that should be considered when deciding how to implement content moderation for an organization like a social media company?

- **Possible clarifying questions from students and prepared examples**

  1. What does content moderation mean?

     Content moderation is the process of reviewing and monitoring user-generated content on online platforms to ensure that it meets certain standards and guidelines. In other words, when a user submits content to a website, that content will undergo a screening process (known as the moderation process) to ensure that the content upholds the website's regulations and is not illegal, inappropriate, harassing, etc.
  2. Is content moderation the same thing as blocking?

     Not necessarily. Content moderation focuses on identifying concerning content. What to do after identification (e.g. blocking) is a separate question.
  3. What does "automated tools" mean?

     Use of a machine learning algorithm or AI, for example, to identify content of concern.
  4. What does "downrank" mean?

     The content will be demoted by the algorithm. It's not removed, but becomes less visible, like appearing on a second page of search engine results.

Figure 4.2: Interview Script for Content Moderation Case Studies

4.3 Study Logistics

We planned to conduct the study by recruiting students from introductory classes at Brown. The study was to be structured as an outside-of-class activity, rather than an in-class assignment. We also intended to involve the SRC teaching assistants (STAs) in the study. These undergraduate students, interested and experienced in SRC content, collaborate with course staff to develop lecture slides, assignments, and other materials for integrating SRC content into various courses.

The study was to take the form of one-on-one interviews on Zoom. This format would allow us the opportunity to ask follow-up questions when students raised interesting points or needed further guidance. The interview script, automatically recorded by Zoom, was to be analyzed qualitatively, providing insights into the students' thought processes. Before initiating the study, we planned to have all participants sign a consent form that would give them comprehensive information about the study procedure, including our data collection and analysis process.

The participants were to be divided into 3 groups, and their performance in the study compared across these groups to address different research questions:

- **Group 1** would consist of students from the introductory course. They were to read two case studies on misinformation, compare and contrast the cases, and discuss various aspects of misinformation and content moderation policy.

- **Group 2** would include a different cohort of students from the introductory course. Instead of receiving two cases, they would be given only one, but during the interview, they would receive similar guiding questions and speak for the same amount of time.

- **Group 3** would consist of STAs, undertaking the same contrasting cases exercise as Group 1.

4.4 Analysis Plan

4.4.1 Quality Rubric

To assess the quality of the students' responses (RQ 1), we planned to use a quality rubric to assign two numeric scores to each student. The first score, which focuses on coverage, would consider the quantity of key factors addressed by the students. The second score, related to depth, would evaluate the level of detail in the students' responses.

This rubric was not related to contrasting cases but aimed purely at assessing the quality of the response. We hypothesized that, compared to Group 1, students in Group 2, while expected to exhibit a similar skill level in SRC, would not benefit from the advantage of comparing and contrasting two cases against each other. Therefore, we anticipated that Group 2 would receive lower scores on the quality rubric compared to Group 1, demonstrating the effectiveness of contrasting cases (RQ 2).

In fig. 4.3, the score levels and criteria outlined in the rubric were depicted, along with illustrative example responses that showcased varying levels of depth in alignment with the rubric's criteria. Our quality rubric was inspired by an internal document within the SRC program that offers guidance and resources for STAs to create SRC rubrics in their classes, with credit due to the STA leadership team for creating this document.

Criteria for Coverage

- How many of the 5 factors did students bring up?
- Did students bring up any other factors or interesting nuances outside of our list?

Levels and Criteria for Depth

- **Good**
  Construct an argument that extends beyond the details and arguments presented within the case. Back the argument with either details from the cases, or examples from experience and background knowledge. Provide reasoning on how the nuance impacts content moderation policy.
- **Getting There**
  Identify the nuance along with the relevant details from the case study, but not able to build on the details and construct an argument.
- **Bad**
  Talk about the case as a whole rather than bringing up nuances.
  Or focus on details that aren't relevant to the process differences in the case.

Example Responses Focusing on Automated Tools Factor

- **Good**
  - *Extracting details from the cases:* Human moderators can detect mistakes or enhance the automated tools. In case 1, BuzzBuzz's AI flagged something and then passed it to human moderators for review. If the human moderators don't agree with the AI's decision, they can revert it. In case 2, the human moderators can find patterns to adjust the model to more accurately flag similar content in the future.
  - *Bringing up examples from background knowledge:* the use of automated tool can impact the content moderation policy. Depending on how the tools work, such as whether they simply detect keywords like racist slur or use NLP to understand the meaning and flag racist content based on context, the definition of a rule violation will change accordingly.
- **Getting There**
  In case 1 BuzzBuzz's AI flagged something and then passed it to human moderators to review. In case 2, FooChat's AI flagged something and the human moderators adjust it to flag more similar content in the future.
- **Bad**
  Both cases mentioned using ML to flag things. One tool catches racist content and the other catches violent content.

Figure 4.3: Quality Rubric

4.4.2 Knowledge Scale

To capture students' background knowledge in the topic as one of the variables in the analysis, we planned to build a knowledge scale based on open coding of students' responses, and then assign each student a numeric score to represent the level of their background knowledge.

To gauge the spectrum of students' background knowledge, we considered two different aspects. During the interview, participants would be asked at the end to share what they already knew about content moderation before doing the exercise, and how much background knowledge they used to answer which questions. In addition to analyzing the responses to these particular questions, we also planned to review the entire transcript, and identify and flag key phrases in students' responses, such as "I know companies that currently do . . . " and "as far as I know, this has been a popular policy". Our last round of study (section 3.5) found that these phrases often indicate that students are drawing upon their background knowledge.

We hypothesized that STAs in Group 3 would score higher on the knowledge scale compared to Group 1. We also wanted to compare Group 3's and Group 1's scores on the quality rubric, to see if there's any significant difference (RQ 3).

4.4.3 Analysis Plan Summary

The table in fig. 4.4 offers a summary of how we planned to address each research question. It covers the setup stage of the study, including the features of the cases (as detailed in the "Case Design" column) and the features of the interview questions (as detailed in the "Prompt Design" column). The "Analysis" section describes our data analysis plan. For all the research questions, we planned to thoroughly examine students' responses to all interview questions to ensure that we don't miss anything that may not have been directly prompted.

4.5 Incorporating Feedback

4.5.1 Refinements Based on a Trial Run with the Lab

After developing the first draft of the cases, we conducted a trial run in the lab. We requested lab members to simulate being undergraduate students enrolled in a computer science class, adhering to the provided instructions as closely as possible.

Structure When designing the cases, our goal was to maintain consistency in structure and narrative, ensuring each case encapsulated a similar set of key factors while showcasing the nuance

Research Question	**Case Design Considerations**	**Prompt Design Considerations**	**Analysis Plan**
RQ 1. Do students produce beyond shallow answers when doing the contrasting cases exercise?		Give examples beforehand to show students what kind of detailed answers we are looking for. Predict the kinds of answers we'd get from students, and develop follow-up questions accordingly.	Look at the distribution of Group 1's and Group 3's (both get contrasting cases) quality rubric scores. Make sure that we are getting sufficient high-quality responses for the rest of the analysis to make sense. See the specific action plans outlined in section 4.1.
RQ 1.1. How does the level of background knowledge influence the quality of students' answers?	The cases shouldn't contain much real-life reference. This way students can't say things like "oh I know that Google currently does this thing".	At the end of the interview, ask students about background knowledge. During the interview, encourage students to back up their responses with reference to the case studies or other examples.	For Group 1 and Group 3 (both get contrasting cases), check for potential correlation independently for students with high and low scores on the knowledge scale.
RQ 1.2 How does the prior SRC experience influence the quality of students' answers?			Compare Group 1's (intro students getting contrasting cases) and Group 3's (STAs getting contrasting cases) quality rubric scores.
RQ 2. In what ways do students delve deeper when prompted to use techniques from contrasting cases?	The cases should be structured in a similar way to make them easy to compare.	There should be explicit instructions in the interview script to push students to compare and contrast.	Compare Group 1's (intro students getting contrasting cases) and Group 2's (intro students getting single case) quality rubric scores.
RQ 3. How well can students apply their knowledge and skills learned to a different scenario?	Construct a third case around the same topic.	Ask students to analyze this third case after the contrasting cases exercise.	Qualitative analysis to find out whether students (in Group 1 and Group 3 who both get contrasting cases) can bring up key features they had identified from the contrasting cases.

Figure 4.4: Summary of Research Questions in Terms of Case Design Considerations, Prompt Design Considerations, and Analysis Plan

of how these factors manifested differently across scenarios. We believed that such a design would facilitate a straightforward comparison, enabling students to not only recognize the overarching key factors but also appreciate the subtle differences in how they are presented across the cases.

However, feedback from the lab illuminated unforeseen discrepancies between the two cases, primarily concerning their length and the depth of narrative. Case 1, centered around Annie's ethical quandary and her contention with her superior, was noted to be considerably less elaborate and lacked the technical specificity found in Case 2, which provided an in-depth examination of Julie's role in content moderation. Moreover, a critique on narrative coherence emerged. Case 1 offered only a cursory mention of a violation detection tool without delving into its operation, whereas case 2 presented a detailed description of the content moderation workflow.

In response to this feedback, we revised both cases so they aligned better structurally. We took out some of the technical details from case 2, while enriching case 1 with some additional insights into how the human moderator interacts with the automated moderation tool. These adjustments aimed to balance the length of the cases and align their storytelling focus, now both telling the story of a content moderator examining the decision made by automated tools.

Wording When crafting the cases, we were careful about our choice of words, aiming to keep the language straightforward and accessible. Our goal in designing this exercise for a computer science class was to ensure that the language used would engage students with the core computer science concepts and key factors, without imposing an unintended challenge in reading comprehension.

Despite these intentions, we received feedback indicating room for improvement in our language choices. One prominent issue was the complexity and length of the company names used in the scenarios, which some found cumbersome and distracting, to the extent of having to repeatedly refer back to the cases. Additionally, suggestions were made to avoid names that closely mimic those of real-world companies, in order to minimize confusion. Furthermore, the use of terms such as "chat" and "social media platform" was identified as potentially distracting. Although these details were not intended to be pivotal, such distinctions could be picked up by students and lead them to dwell on the functionalities specific to different types of social media platforms, thus diverting their attention from the key factors we aimed for them to focus on.

To address these issues, we replaced the original company names with the simpler, more generic identifiers "AAA" and "BBB". These shorter, more memorable names not only reduce the cognitive load on students, but also clearly signal that the companies and scenarios are fictional, thereby avoiding any confusion with real-world entities. Additionally, these names helped us remove the unnecessary distinction between "chat" platform and "social media platforms".

Other Factors To understand how the cases were perceived, we asked participants to share the details in the cases that they found interesting. While it was encouraging to see that most of our intentionally embedded key factors were identified—hardly a surprise given our lab's deep involvement in computer science education—certain unintended details captured more attention than anticipated. Notably, the use of exclusively feminine names (Annie and Julie) was perceived as a significant detail by some, sparking discussions on gender representation. Similarly, the explicit mention of Bangladesh led to reflections on the realism of the scenario and considerations about how students' personal backgrounds might influence their interaction with the case studies. For instance, questions arose about the potential reactions of students from Bangladesh.

Recognizing that such details, although unintentional, could divert focus from the core learning objectives, we decided to modify the cases accordingly. We replaced the original names with gender-neutral alternatives and adopted they/them pronouns, aiming to neutralize the unintended emphasis on gender. Furthermore, we omitted the reference to Bangladesh, while keeping the more generic description of "a country that doesn't share the American concept of race." This adjustment maintains the necessary context for students to grasp the cultural context factor without leading students to fixate on specifics that might detract from the topic of content moderation.

4.5.2 Changing the Study Setup after Thesis Proposal

During the thesis proposal, we had lots of discussions about the potential validity threats and limitations associated with conducting the study as an outside-of-class activity. We initially believed that conducting interviews via Zoom outside of the classroom settings would provide a controlled environment where it's easier for students to verbally express their thoughts in real time, allowing for immediate follow-up questions and a deeper understanding of each student's reasoning. This approach was deemed superior to written exercises, which posed the risk of students misunderstanding the prompts, responding minimally, or failing to provide the depth of insight we sought. However, the necessity of recruiting volunteers for the outside-of-class activity introduced a significant threat to external validity through volunteer bias. We faced the risk of attracting only those students who had either a pre-existing interest in the subject matter (sufficient to invest 30 minutes of their time), or those motivated by the incentive of a gift card as compensation. Unfortunately, neither group is likely to be adequately representative of the broader introductory-level class population from which we aimed to recruit.

There were also concerns about the ecological validity: how much can the study be reused and reapplied in a real classroom setting later? The interview format, while beneficial for tailoring questions to each student's thought process and garnering rich information from 30-minute sessions, presents a significant challenge in scalability. Transforming this approach into an exercise suitable for all students in an introductory-level class highlights scalability concerns, underscoring the study's limitations on its feasibility to a more generalized setting.

Furthermore, questions were raised about our decision to recruit STAs as participants. We planned to investigate this group through RQ 1.2: how does the prior SRC experience influence the quality of students' answers? Our plan was to compare the intro students' answers with STAs' answers, with the assumption that STAs would have more SRC experience. However, feedback from my committee suggested that the diverse range of skills among both introductory students and STAs, which could be highly dependent on the topic. For example, some introductory students, despite lacking prior computer science experience, might have extensive familiarity with content moderation tools from personal use of social media platforms. Conversely, some STAs, while proficient in course content and current SRC discussions, might have limited understanding of content moderation concepts. This variability challenges the assumption that "prior SRC experience" can be easily defined or generalized for our study's purposes, especially given the topic-specific nature of the cases.

In response to these concerns, we revised the study design into a two-part structure: Part 1 consists of a mandatory written exercise administered to all students in a class, while Part 2 involves conducting Zoom interviews with a select group of students. This adjustment ensures engagement with the entire class, thereby avoiding the bias inherent in a self-selecting sample. The written responses serve as a preliminary indicator of the students' familiarity with the topic and their perception of the study as we envisioned it. From these responses, we are equipped to identify and select a diverse group of students for the Part 2 interviews to gain more insights.

As a result of these revisions, our research questions also changed:

1. In what areas do students respond deeply when engaged in the contrasting cases exercise?
   (a) How does the level of background knowledge relate to the quality of students' answers?
2. In what ways do students delve deeper when getting two cases rather than one?
3. How well can students apply their knowledge and skills learned to a different scenario?

4.5.3 Redesigning the Quality Rubric

Another major area of feedback we received was about the quality rubric. The proposed rubric (in fig. 4.3) puts a heavy emphasis on how students engage with the cases. The criteria defined a "good" response as one that "extends beyond the cases," a "getting there" response as incorporating "relevant details from the cases," and a "bad" response as including "irrelevant details". This approach inadvertently shifted the focus towards students' ability to incorporate case details into their responses rather than evaluating the depth of their understanding.

In order to improve the quality rubric, we first had a discussion about the core objective we wanted to measure, and came to the conclusion that we intended to gauge students' ability to construct a coherent argument regarding a key factor. This goal actually aligns with the learning objective of some CS courses, such as CS112 Computing Foundations: Program Organization. This course, which follows CS111, expands on programming and data handling, emphasizing efficient structuring, algorithms, and ethical considerations. Inspired by CS112's grading rubrics, we redesigned our criteria to assess whether students could articulate the significance of the key factor and relate it to the broader issue of content moderation. Furthermore, we realized that labeling responses lacking a strong argument as "bad" was not accurate, and decided to use "weak" as label instead. The redesigned rubric is shown in table 4.2.

Tag	**Criteria**	**Example Response**
Good	*Mentions for a single factor: Explain a highlighted detail with information that was not explicit in the case regarding why it matters or how it connects to content moderation policy.*	"The automated tool needs to be facilitated by human moderators, because human moderators can detect mistakes. As illustrated in case 1, if the human moderators don't agree with the tool's decision, they can revert it, making sure the decision follows the company's content moderation rules."
Getting There	*Merely mentions the highlighted details, but does not expand on that detail with additional insights.*	"In case 1, the tool flagged something and then passed it to human moderators to review."
Weak	*Bring up details that aren't relevant to the central issue.*	"Users can edit posts as many times as they want. User interaction."
Missing	*Didn't bring up the key factor.*	

Table 4.2: Updated Quality Rubric

Chapter 5

Implementing the Study

Drawing on the insights gathered from earlier iterations and integrating the feedback from lab mates and committee members, we developed an expanded study that aimed to evaluate the effect of contrasting cases at a larger scale. This study was conducted in two computer science courses. The study was structured in two phases: part 1 involved a written assignment featuring case studies that were presented to all students in the class, and part 2 involved follow-up interviews with selected students, aiming to gain deeper insights into students' thought processes and engagement with the case studies.

5.1 Study Setup

We gathered data in two distinct computer science courses at Brown University during the Spring 2024 semester. The first course, CS111 Computing Foundations: Data, is an introductory courses open to all students and does not require prior programming experience. The second course, CS32 Introduction to Software Engineering, is a project-based course that focuses on designing and building software systems. Enrollment in CS32 requires completion of an introductory programming course as a prerequisite.

Pedagogic Context We selected CS111 and CS32 for our study based on several criteria. First, both courses had previously incorporated SRC in their curriculum, providing an ecological setting where students were already familiar with SRC assignments. And this familiarity allowed for seamless integration of our study within the existing course structures. Second, the courses catered to distinct student demographics: CS111 was composed of beginners with minimal or no prior exposure to computer science or SRC, whereas CS32 students had completed at least one semester of computer science, including some SRC content. Studying these diverse student populations enabled us to explore the differential impacts of contrasting cases across varying levels of prior CS knowledge. Lastly, although other courses at similar levels were considered, CS111 and CS32 were selected for their pedagogical flexibility, which enabled the incorporation of our study without disrupting the normal course progression.

For the CS111 course, the subject of content moderation was selected as the central topic for the case studies. The primary learning objective was to enable students to identify and analyze the various considerations involved in developing content moderation tools. Although this topic was not originally covered in the course syllabus, it was deemed highly relevant and engaging for students, given its prominence in the current social media landscape—a context with which most students are probably familiar from personal experience. After discussions with the course instructor, we decided to integrate this study into the course at the start of the semester, as part of the first assignment. This timing was strategically chosen to capture students' initial reactions before they had substantial exposure to SRC content, thereby allowing for a clearer assessment of the case studies' effect.

For CS32, we selected CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart), seeing it as an apt subject to extend beyond the conventional presentation of ADA standards as mere checklists. The process for developing these case studies mirrored our previous approach: beginning with some fundamental design principles, researching real-world applications to see how these principles manifest differently, and then crafting narratives to establish a pair of contrasting cases that highlight these nuances. We grounded our cases in the 7 Principles of Universal Design [33], selecting specific guidelines most relevant to CAPTCHA scenarios. To ensure the cases reflected actual challenges, we referenced various sources on CAPTCHA and critiques of their accessibility issues ([12, 65, 52]).

Our objective is for students to appreciate how these design principles materialize in software development, recognizing potential conflicts and the trade-offs required. Upon reviewing last semester's course syllabus, we observed that the focus on accessibility was predominantly centered around the use of screen readers. Consequently, we determined that introducing our case study prior to the screen reader exercise would be beneficial. This arrangement aims to prevent students from being overly influenced or primed by the screen reader topic, ensuring they consider a broader range of accessibility issues without narrowing their focus solely to aspects highlighted by screen reader use.

Group Allocation In both courses, students were randomly allocated into one of three distinct groups, each receiving a different version of the case studies. Group A received a pair of contrasting cases, Group B received only Case 1, and Group C received only Case 2. This experimental design was chosen not only to observe the effectiveness of contrasting cases but also to explore whether exposure to a single case could be as effective. The underlying aim was to discern whether students who engaged with contrasting cases performed better compared to their peers who analyzed a single case. The complete written assignments for Group A, which include both cases, are detailed in appendix B for CS111 and in appendix C for CS32.

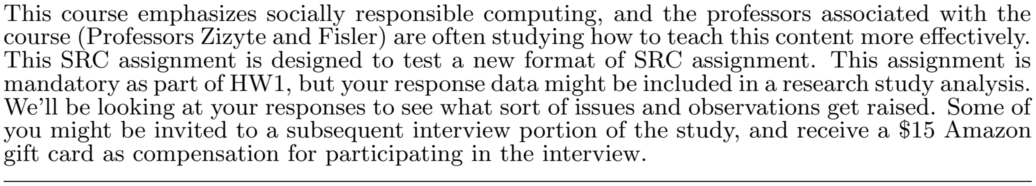
Written Tasks All students were asked to complete the same two tasks. Task 1 involved reading and highlighting interesting details in the cases, while Task 2 introduced a new scenario related to the cases' topic, and prompted students to list relevant factors for consideration. The complete task prompts are shown in fig. 5.1. For Task 2, students were instructed to organize their responses in a table format, dividing each factor into three columns: the factor to consider, what's interesting about this factor, and the source of their knowledge about the factor. In the source column, students indicated the origin of their thoughts by selecting checkboxes for ”case”, ”previous knowledge”, or ”personal experience”.

- **Task 1**: Please read the case study below, and use the highlighting tool of Google Docs to mark any parts that look interesting or notable to you.
- **CS111 Task 2**: Imagine you are a software engineer designing content moderation tools for a social media startup CCC. What are some factors you need to consider?
- **CS32 Task 2**: Imagine you are a software engineer designing the user-facing part of a CAPTCHA system for an online voting platform CCC. What are some factors you need to consider?

Figure 5.1: Written Tasks Prompts

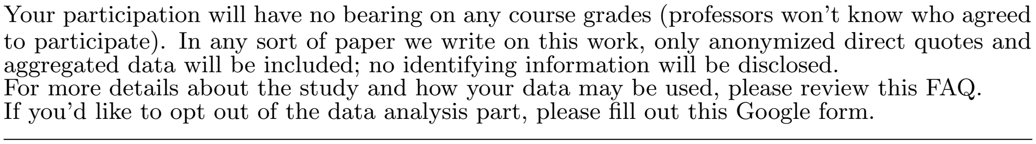
Assignment Handout Design In the assignment handout, we included a section that outlined the overarching objectives of the study (as shown in fig. 5.2), deliberately omitting specifics of the study's design to avoid influencing the students' responses. This approach was intended to elicit authentic reactions rather than responses tailored to what students might perceive as desirable outcomes. Specifically for CS32, which normally promotes an open collaboration policy encouraging the use of any available resources and peer sharing, an additional paragraph was incorporated. This paragraph underscored that, uniquely for this task, students were required to complete everything independently to ensure the integrity of the data collected.

Student Consent: The assignment handout emphasized that while completion of the assignment was mandatory, consent to include their data in the study was voluntary. An opt-out option was provided for students who chose not to participate in the data collection, as detailed in the handout. As shown in fig. 5.3, students were also provided with a list of frequently asked questions regarding the study and data use, and the full list is available in the appendix for further reference (appendix D).



This course emphasizes socially responsible computing, and the professors associated with the course (Professors Zizyte and Fisler) are often studying how to teach this content more effectively. This SRC assignment is designed to test a new format of SRC assignment. This assignment is mandatory as part of HW1, but your response data might be included in a research study analysis. We'll be looking at your responses to see what sort of issues and observations get raised. Some of you might be invited to a subsequent interview portion of the study, and receive a $15 Amazon gift card as compensation for participating in the interview.

Figure 5.2: Assignment Handout Section on Study Explanation



Your participation will have no bearing on any course grades (professors won't know who agreed to participate). In any sort of paper we write on this work, only anonymized direct quotes and aggregated data will be included; no identifying information will be disclosed.
For more details about the study and how your data may be used, please review this FAQ.
If you'd like to opt out of the data analysis part, please fill out this Google form.

Figure 5.3: Assignment Handout Section on Data Use for Research Study

5.2 Winter Pilot

5.2.1 Study Context and Participants

To validate the design and methodology of our study before its full-scale implementation in regular courses, we conducted a pilot round during the winter break immediately prior. We invited students who had completed the same courses in the fall semester via email, resulting in participation from 8 students from CS111 and 12 from CS32 for the written tasks. Additionally, follow-up interviews via Zoom were conducted with 5 students from CS111 and 7 from CS32. The primary focus of this pilot was to assess whether students understood and engaged with the cases and prompts as intended, rather than to analyze their performance. The latter aspect was meant to be explored in greater depth during the formal execution of the study.

5.2.2 Refinements for CS111

Student responses to the cases largely aligned with our expectations, requiring minimal changes, which was anticipated due to several rounds of prior feedback. However, we identified a point of confusion in Case 2 concerning the phrase ”with a mark that is visible to users of the platform.” One student was uncertain whether the mark would be visible only to the content poster or to everyone on the platform. To clarify this ambiguity, we revised the wording to ”visible to all users.” The final version of the cases for CS111 is illustrated in fig. 5.4, where different colored highlights indicate various pre-planted key factors. Additionally, fig. 5.5 presents this information in a tabular format for clear visual comparison.

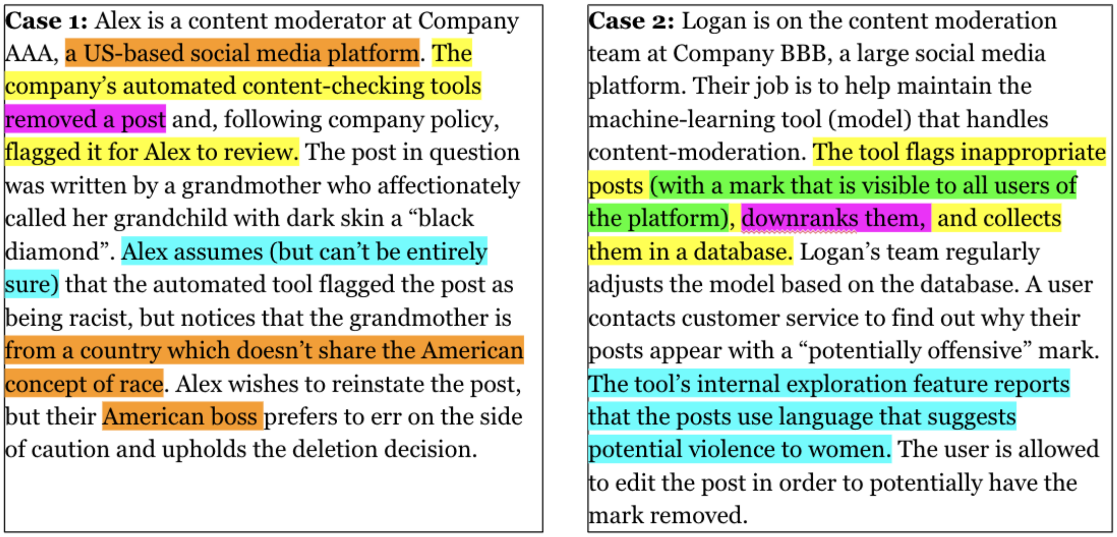
5.2.3 Refinements for CS32

We implemented several revisions to both the cases and the prompts to enhance clarity and focus. Figure 5.6 and fig. 5.7 illustrate how the cases were modified, with strikethroughs indicating removed content and italicized words representing additions or replacements.

For case 1 (as shown in fig. 5.6), we initially described the interactive challenge as a ”simple logic problem,” intended to signify straightforward questions requiring minimal cognitive effort. However, feedback from interviews revealed confusion among students about the term ”logic,” with some questioning its inherent simplicity and others diverging into discussions on user testing rather than focusing on universal design principles, which was our study's aim. To mitigate this, we simplified the language to ”simple questions” and only kept a clear example: ”what's 1+six”. This change aimed to prevent misinterpretations about the nature of the tasks and intelligence assessments, particularly following concerns that the phrase ”be easily solvable by someone with the intelligence of a seven-year-old child” led some students to debate intelligence definitions rather than the CAPTCHA's functionality.

Additionally, the original case 1 ended with a mention of an alternative, non-interactive machine learning option available only to users who registered with an email address. This detail was intended to spur discussion on the trade-off between convenience and privacy. However, students expressed a desire for more information on how the machine learning aspect functions and some felt unprepared to discuss it adequately. In response, we revised this section to provide more details about the machine learning process, thereby making the email requirement more prominent and understandable.

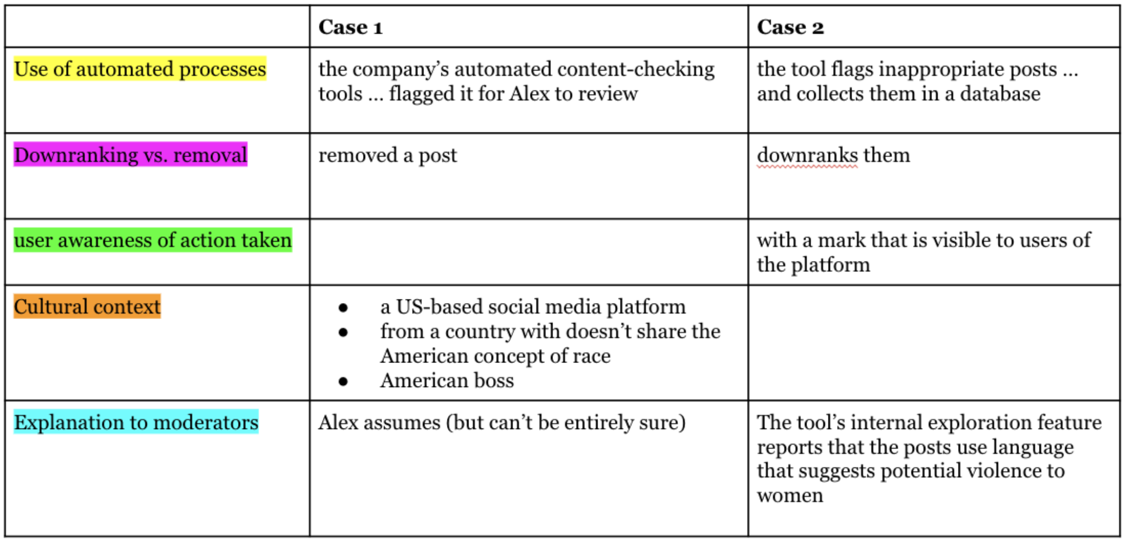
In case 2, as depicted in fig. 5.7, the phrase ”allows users to request a new task” initially led to confusion among some students. They interpreted this as indicating a different type of CAPTCHA task, not merely a different instance of a previously mentioned image or audio challenge. To clarify



Case 1: Alex is a content moderator at Company AAA, a US-based social media platform. The company's automated content-checking tools removed a post and, following company policy, flagged it for Alex to review. The post in question was written by a grandmother who affectionately called her grandchild with dark skin a "black diamond". Alex assumes (but can't be entirely sure) that the automated tool flagged the post as being racist, but notices that the grandmother is from a country which doesn't share the American concept of race. Alex wishes to reinstate the post, but their American boss prefers to err on the side of caution and upholds the deletion decision.

Case 2: Logan is on the content moderation team at Company BBB, a large social media platform. Their job is to help maintain the machine-learning tool (model) that handles content-moderation. The tool flags inappropriate posts (with a mark that is visible to all users of the platform), downranks them, and collects them in a database. Logan's team regularly adjusts the model based on the database. A user contacts customer service to find out why their posts appear with a "potentially offensive" mark. The tool's internal exploration feature reports that the posts use language that suggests potential violence to women. The user is allowed to edit the post in order to potentially have the mark removed.

Figure 5.4: CS111 Contrasting Cases with Key Factors Marked



	Case 1	**Case 2**
Use of automated processes	the company's automated content-checking tools ... flagged it for Alex to review	the tool flags inappropriate posts ... and collects them in a database
Downranking vs. removal	removed a post	downranks them
user awareness of action taken		with a mark that is visible to users of the platform
Cultural context	• a US-based social media platform • from a country with doesn't share the American concept of race • American boss	
Explanation to moderators	Alex assumes (but can't be entirely sure)	The tool's internal exploration feature reports that the posts use language that suggests potential violence to women

Figure 5.5: CS111 Pre-Planted Key Factors Table

this, we revised the wording to ”a new CAPTCHA challenge.”

Another issue emerged when we asked students to compare and contrast the two cases between the two cases: some students fixated on the terms ”current CAPTCHA system” and ”upgrade” used in case 2. They spent considerable effort trying to distinguish between the existing system and proposed upgrades, which was not the intended focus, as both cases were designed to discuss the design of a new CAPTCHA system. To eliminate this confusion and reduce cognitive load, we simplified the language in Case 2 to emphasize that it concerns the development of a new CAPTCHA system, aligning it more clearly with the narrative in case 1.

> **Case 1 (used in winter pilot)**: Jordan is an engineer working at AAA, an e-commerce platform. As the platform grows, the team notices increasing bot traffic, and has asked Jordan to propose a CAPTCHA system that would be easy to use. Jordan's system uses simple ~~logic~~ questions ~~that can be easily solvable by someone with the intelligence of a seven-year-old child~~ (e.g., “what's 1+six”), but the problems are only available in English. The ~~logic~~ questions are fully compatible with screen readers and can be navigated using a keyboard. Users are prompted to answer the questions within two minutes, but can press a “give me two more minutes” button to get extra time. Users are allowed up to three wrong attempts. There's also an alternative, non-interactive option where the system uses machine learning to detect human users, but this option *requires users to provide their email address as a unique identifier within the system.*

Figure 5.6: CS32 CAPTCHA Case 1 Modification

> **Case 2 (used in winter pilot)**: Charlie is an engineer at BBB, an online ticketing platform. The current CAPTCHA system has been effective preventing bots from bulk buying tickets, but *isn't particularly user-friendly* ~~the company wants to upgrade the system to be more user friendly~~. Charlie proposes using a third-party tool in which users who are already logged into their Google account in the same browser will be verified as human without filling in the CAPTCHA. Under the hood, cookies track users' online activities and share that data with Google. Those not logged into Google, will be asked to type out a single word shown in an image. An audio alternative is also available. Both the images and audios are from database built from old English books. The system also allows users to request a new *CAPTCHA challenge* ~~task~~ if they find the current one too difficult. To prevent brute force attacks by bots, the proposed tool requires users to finish the task within a minute.

Figure 5.7: CS32 CAPTCHA Case 2 Modification

The final version of the cases for CS32 is illustrated in fig. 5.8, where different colored highlights indicate various pre-planted key factors. Additionally, fig. 5.9 presents this information in a tabular format for clear visual comparison.

We also modified the Task 2 prompt for CS32 based on observations during the interview, where some students extensively explored the technical aspects of the CAPTCHA system, including methods to differentiate bots from human users, strategies to prevent bot attacks, and metrics for monitoring bot success rates. Recognizing that these discussions diverged from the primary learning objectives of the study, we added a specific emphasis on the ”user-facing part” to the task prompt (the modification is illustrated in fig. 5.10). This adjustment was intended to steer students’ focus toward considering factors related to universal design.

> **CS32 Task 2**: Imagine you are a software engineer designing the *user-facing part of a* CAPTCHA system for an online voting platform CCC. What are some factors you need to consider?

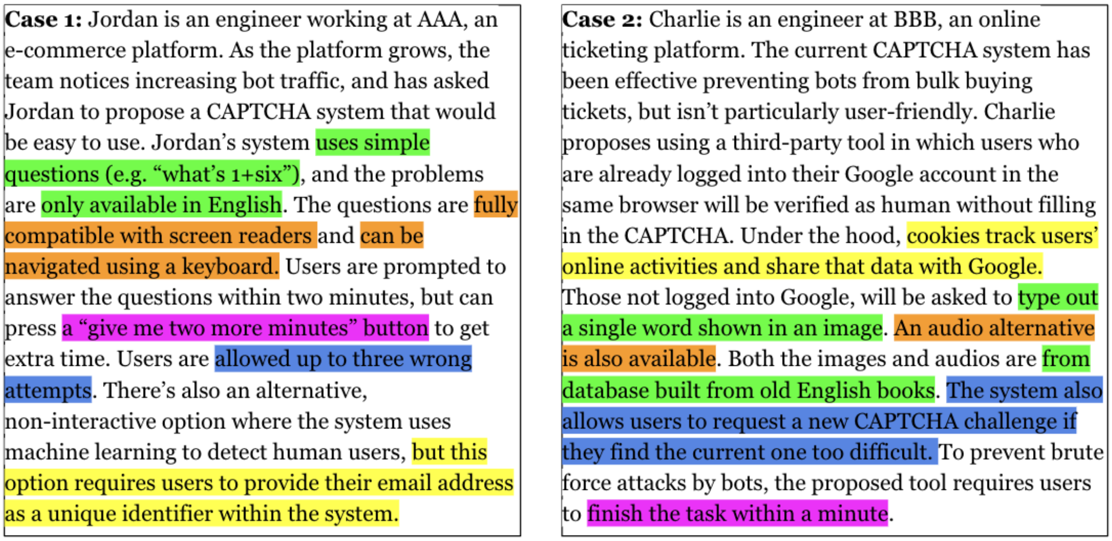
Figure 5.10: CS32 CAPTCHA Task Prompts Modification

5.2.4 Refining the Interview Script

After experimenting with various interview prompts and allowing students to freely expand on their thoughts, we refined the questions and developed a structured interview script. This script was segmented into several sections with predefined questions designed to encourage students to elaborate on their written responses. Additionally, it included space for new follow-up questions and allowed for adjustments during the interview to tailor the discussion more closely to each student’s responses. The finalized script for CS111 is available in appendix E, and the script for CS32 is detailed in appendix F.

5.2.5 Updated Research Questions

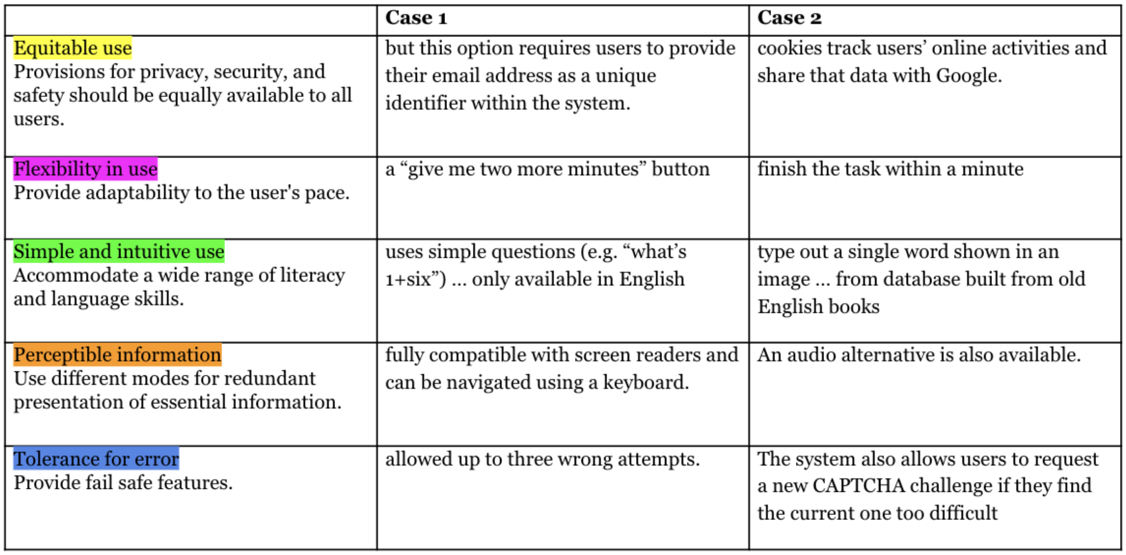
After running the pilot and finalizing the details of the study content and study protocol, we updated the sub research questions and analysis plan, as shown in table 5.1. Note that the data from the interviews are not included here, since the interviews are semi-structured, we would process all the interview data and see if we get any different or additional insights compared to the written response analysis.



Case 1: Jordan is an engineer working at AAA, an e-commerce platform. As the platform grows, the team notices increasing bot traffic, and has asked Jordan to propose a CAPTCHA system that would be easy to use. Jordan's system uses simple questions (e.g. "what's 1+six"), and the problems are only available in English. The questions are fully compatible with screen readers and can be navigated using a keyboard. Users are prompted to answer the questions within two minutes, but can press a "give me two more minutes" button to get extra time. Users are allowed up to three wrong attempts. There's also an alternative, non-interactive option where the system uses machine learning to detect human users, but this option requires users to provide their email address as a unique identifier within the system.

Case 2: Charlie is an engineer at BBB, an online ticketing platform. The current CAPTCHA system has been effective preventing bots from bulk buying tickets, but isn't particularly user-friendly. Charlie proposes using a third-party tool in which users who are already logged into their Google account in the same browser will be verified as human without filling in the CAPTCHA. Under the hood, cookies track users' online activities and share that data with Google. Those not logged into Google, will be asked to type out a single word shown in an image. An audio alternative is also available. Both the images and audios are from database built from old English books. The system also allows users to request a new CAPTCHA challenge if they find the current one too difficult. To prevent brute force attacks by bots, the proposed tool requires users to finish the task within a minute.

Figure 5.8: CS32 Contrasting Cases with Key Factors Marked



	Case 1	**Case 2**
Equitable use Provisions for privacy, security, and safety should be equally available to all users.	but this option requires users to provide their email address as a unique identifier within the system.	cookies track users' online activities and share that data with Google.
Flexibility in use Provide adaptability to the user's pace.	a "give me two more minutes" button	finish the task within a minute
Simple and intuitive use Accommodate a wide range of literacy and language skills.	uses simple questions (e.g. "what's 1+six") ... only available in English	type out a single word shown in an image ... from database built from old English books
Perceptible information Use different modes for redundant presentation of essential information.	fully compatible with screen readers and can be navigated using a keyboard.	An audio alternative is also available.
Tolerance for error Provide fail safe features.	allowed up to three wrong attempts.	The system also allows users to request a new CAPTCHA challenge if they find the current one too difficult

Figure 5.9: CS32 Pre-Planted Key Factors Table

Research Question	Analysis Plan
RQ 1. In what areas do students respond deeply when engaged in the contrasting cases exercise?	Qualitatively analyze the categories (or themes) of the factors that students brought up. Use the quality rubric to rate the quality of the reasoning. Find out which factors have more high-quality reasons, and which ones have low-quality reasons.
RQ 1.1. How does the level of background knowledge relate to the quality of students' answers?	Across individual students, check for any patterns between background knowledge and their response.
RQ 2. In what ways do students delve deeper when getting two cases rather than one?	Compare the results across groups A, B and C to see if there are any differences or patterns in the quality of responses.
RQ 3. How well can students apply their knowledge and skills learned to a different scenario?	Examine the categories of factors that students touched on in their response and the source. Summarize which factors are based on the cases.

Table 5.1: Research Questions and Analysis Plan

5.3 Data Analysis Process

Data Collection and Processing We collected students' written submissions by uploading them to a designated folder within Brown's Google Drive, ensuring that access to this folder was strictly limited to authorized members of the research team, all of whom are affiliated with Brown as faculty or students. Each work was associated with a unique anonymous participant ID, safeguarding the anonymity of the data throughout our analysis. Furthermore, we securely stored students' emails separately, using them solely for study-related communication and scheduling purposes. This allowed us to associate emails with specific IDs only when necessary, such as to potentially invite participants to the interview portion of the study or to facilitate the removal of data for those opting to withdraw from the study.

To process and analyze the data collected via Google Docs, we leveraged the capabilities of Google Sheets' App Script for batch processing, utilizing the automation and formatting options it offers. In this context, ChatGPT played a crucial role in generating scripts. Our typical workflow involved providing ChatGPT with a detailed task description based on the data contained in a single row. For example, we might instruct ChatGPT to access a Google Doc linked in cell B2, parse a table within that document, and then organize the extracted data into a separate, predefined ”factor table,” ensuring each entry is uniquely identified by a combination of the ID from cell A2 and the row number from the table in the document.

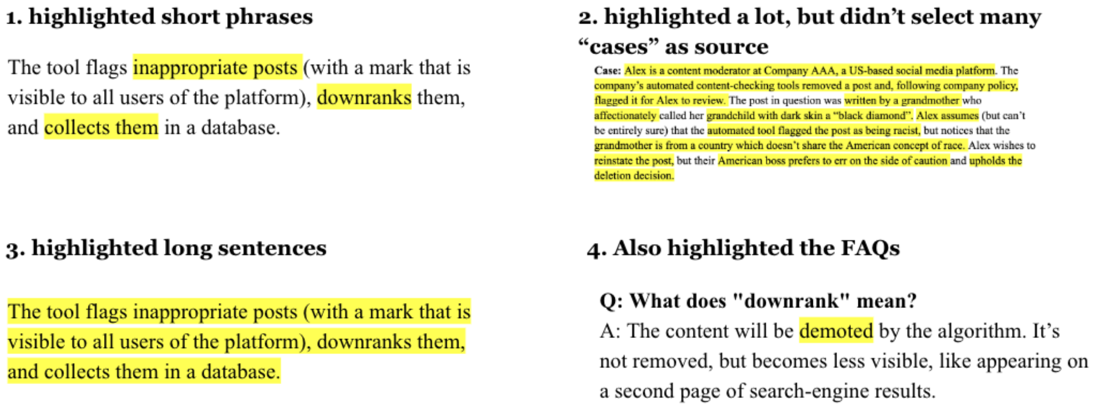
Following the generation of code by ChatGPT, our next steps were to manually review the code for any bugs, correct them, and then test the code on a single row or a small batch of data. This allowed us to verify that the outputs matched our expectations. Once satisfied with the functionality on the smaller batch, we then tasked ChatGPT with modifying the script to process the entirety of the data. The final stage involved randomly selecting several rows of the processed data to manually inspect and validate their correctness. Employing ChatGPT in this way ensured our thorough understanding of each function’s logic, expected inputs and outputs, and ultimately instilled confidence in the accuracy of the final processed data.

Highlight Analysis Task 1 of the written portion asked students to read and highlight the case: ”Please read the case study below, and use the highlighting tool of Google Docs to mark any parts that look interesting or notable to you.” This prompt was designed with two objectives: firstly, to ensure that students thoroughly read the case, and secondly, to capture their perceptions of the case through the open-ended nature of the wording ”interesting or notable.” This method engaged students in active reading and provided us with valuable insights to refine the case studies for future iterations of the research.

During the analysis phase, we initially attempted to parse out the highlights to compare the parts that captured students’ attention with our pre-planted factors in the cases. However, we encountered challenges in distilling and matching students’ highlights with ours, as highlights are continuous rather than discrete. Even within our own highlighting, we used different colors to mark different factors (as shown in fig. 5.4 and fig. 5.8). Without the color distinctions, the highlights tended to blend together, making it difficult to determine which parts corresponded to specific factors.

As a result, we shifted from a quantitative to a qualitative analysis approach, looking for patterns in the highlights rather than trying to quantify exact matches with pre-planted factors. Figure 5.11 presents some examples of these patterns, which we used as criteria for selecting interview participants.

Our findings from the interviews reinforced the complexity of the highlighting data. Students reported various reasons for highlighting: some highlighted to note the motive and context of the narrative; others marked sections they found problematic or agreed with; some highlighted parts they did not understand; and a few even used different colors to signify multiple reasons.



1. highlighted short phrases

The tool flags inappropriate posts (with a mark that is visible to all users of the platform), downranks them, and collects them in a database.

2. highlighted a lot, but didn't select many "cases" as source

Case: Alex is a content moderator at Company AAA, a US-based social media platform. The company's automated content-checking tools removed a post and, following company policy, flagged it for Alex to review. The post in question was written by a grandmother who affectionately called her grandchild with dark skin a "black diamond". Alex assumes (but can't be entirely sure) that the automated tool flagged the post as being racist, but notices that the grandmother is from a country which doesn't share the American concept of race. Alex wishes to reinstate the post, but their American boss prefers to err on the side of caution and upholds the deletion decision.

3. highlighted long sentences

The tool flags inappropriate posts (with a mark that is visible to all users of the platform), downranks them, and collects them in a database.

4. Also highlighted the FAQs

Q: What does "downrank" mean?
A: The content will be demoted by the algorithm. It's not removed, but becomes less visible, like appearing on a second page of search-engine results.

Figure 5.11: Examples of Students' Highlights

Interview Participant Selection To enrich our understanding of participants' thought processes and enhance insights from their written responses, we developed specific criteria for selecting interviewees, aiming to capture a diverse cross-section of perspectives. Based on the patterns identified in the highlights and a preliminary scan of the data, four distinct criteria were identified (detailed in table 5.2). We issued 20 invitations for each course — double the number anticipated based on the acceptance rate from the winter pilot — to ensure adequate participation. The recruitment process was carefully managed to achieve a balanced representation across the established criteria and within the three participant groups. Ultimately, this resulted in a total of 9 interviewees for each course: in CS111, 4 from Group A, 3 from Group B, and 2 from Group C; in CS32, 2 from Group A, 4 from Group B, and 3 from Group C.

Selection Criteria	**N (CS111)**	**N (CS32)**
Highlighted a lot, but didn't list many factors	2	3
Didn't highlight much	2	3
Used multiple colors for highlighting	2	1
Had a lot of factors, but very brief reasoning	3	2

Table 5.2: Interview Selection Criteria and Number of Responses

Data Analysis Approach Working closely with our undergraduate research assistant Annika Singh, we conducted an initial round of analysis by independently coding the written responses from the 9 interviewees. During this process, we each took notes on how we applied and adapted the original codebook and rubric. This was followed by a collaborative session where we discussed our findings, resolved any discrepancies, and integrated our methodologies into a unified codebook and rubric. Subsequent to our analysis of the written responses, we shifted our focus to the interview transcripts. This examination was carried out with the intention of identifying unique insights and additional details not evident in the written responses. These findings were regarded as supplementary qualitative data, providing depth and context that the written responses alone did not fully convey.

Following the interview analysis, we selected a representative sample of written responses for further examination. From each group, we randomly chose 10 participants whose number of factors closely aligned with the group's average, based on a range within one standard deviation. This stratified sample of 30 written responses was then collaboratively coded by the undergraduate researcher and the author, employing the finalized codebook to ensure consistency and depth in our analysis.

Finally, for the overall dataset, we conducted quantitative analyses and made comparisons across the groups. As shown in table 5.3, our approach to data analysis involves a strategic combination of carefully selecting samples for in-depth qualitative analysis and conducting a quantitative examination of the overall dataset. This method enabled us to explore the nuances of individual thought processes while also grasping the broader trends and patterns across groups.

Dataset	**Sample Size (CS111 | CS32)**	**Data Analyzed**	**Type of Analysis**
Interview	9 | 9	written response + interview transcript	qualitative
Selected Sample	30 | 9	written response	qualitative + quantitative
Overall Data	161 | 88	written response	quantitative

Table 5.3: Overview of Dataset and Analysis Methods

5.3.1 Developing a Codebook

In Task 2 of the written part, students were instructed to list out the factors for consideration and provide a reason why each of them was interesting. Our initial plan was to code only the factor column to categorize the issues students discussed. However, we soon noticed discrepancies between the factors listed and the accompanying reasons, with some students providing arguments that extended beyond the scope of the factor they identified. For instance, one student wrote down "AI learning" as a factor, and for the reason, they wrote *"The automated tools could be fetching* *their information about racism from outdated sources, or not checking the country from where a comment originated from."* This response actually addressed two distinct issues: the impact of data quality on automated tools' performance and the cultural considerations necessary when evaluating data quality. Had we focused solely on the factor column, the cultural dimension would have been overlooked. To capture the full breadth of students' insights, we opted to code both the factor and the reason together. We then assigned one or multiple tuples of (tag, quality) to each row, where 'tag' represents the type of issue identified, and 'quality' assesses the strength of the argument surrounding that tag.

We initially constructed our codebook with five top-level codes, or what we called "categories", each corresponding to one of the pre-planted factors outlined during the case design. As our analysis proceeded, we enhanced these initial categories by adding more specific "tags" under each to capture a broader range of topics and keywords identified in the responses. Additionally, we recognized the emergence of entirely new types of issues or factors raised by students that were not included in our original planning. To include these insights, we expanded our codebook by introducing new categories. This division resulted in two main sections within our codebook: the pre-planted categories, which were anticipated based on the study design, and the new categories, which were developed in response to unforeseen topics brought up by the students. Each category, whether pre-planted or new, includes various tags that detail more specific elements related to the broader issue or factor. The complete codebook for CS111 is available in appendix G, and for CS32, it can be found in appendix H.

5.3.2 Updating the Quality Rubric

To assess the quality of students' responses, we started by applying the quality rubric outlined in table 4.2. This rubric was crafted based on criteria emphasizing the students' ability to construct good arguments using evidence from the case studies. During the coding process, each coder independently took notes on how the rubric was applied, including example responses to illustrate their decisions. We then compared these notes and engaged in discussions to summarize the main characteristics of each rating and the distinctions between them.

An important insight emerged during this process: while we anticipated that students would base their answers predominantly on the cases, we observed that some were able to introduce new categories and construct compelling arguments without directly referencing the case materials. Consequently, we revised our quality rubric to accommodate these findings, with the updated version available in appendix I.

The revised rubric offers clearer differentiation between ratings: a "good" rating indicates that the response effectively connects the factor to the central topic. In contrast, a ”getting there” rating suggests that the discussion is relevant but lacks the connection between the factor and the central theme. Furthermore, a ”weak” rating is assigned to responses that are more generic, often missing specific details to support the argument or clear relevance to the main topic.

5.4 Findings

Having detailed the methods employed in our data analysis process, we now proceed to describe our findings. This section presents the results of our data analysis, organized in alignment with the research questions that framed our study.

5.4.1 RQ 1. Do students produce beyond shallow answers when doing the contrasting cases exercise?

This research question serves as both a foundational check and a prerequisite for further analysis. Our goals here is to grasp the space of responses: what details are students noticing from the case studies, and do they align with our expectations? This examination is critical for several reasons. First, it helps us assess whether students are producing high-quality responses. Such an evaluation is fundamental before delving into additional research questions. If a significant portion of the responses are found to be weak, it may indicate that students either did not take the assignment seriously or failed to grasp its main ideas. On the other hand, a predominance of high-quality responses might suggest that the students already possessed a strong understanding of the topic before undertaking the written tasks and didn't benefit much from the contrasting cases approach.

Course	Average	Median	Mode	Std Dev
CS111	4.91	5	5	2.05
CS32	4.91	5	4	1.9

Table 5.4: Statistics on the Number of Factors Students Listed (Overall Data)

How many factors did students list? We first examined the Overall Data from both courses to gauge the number of factors students identified. As shown in table 5.4, students in both courses demonstrated an average, median, and mode around 4 and 5. Since when we designed the cases, we had planted 5 factors in each set of the cases, these numbers roughly match our expectations. Considering that we embedded five factors within each set of case studies during the design phase, these statistics are in alignment with our expectations.

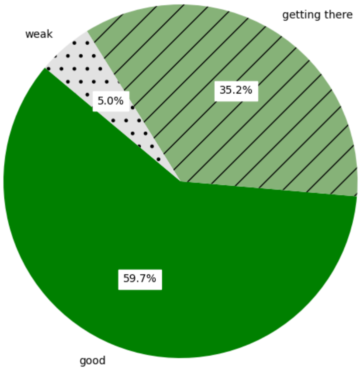


Figure 5.12: CS111 Quality Rating Breakdown (Selected Sample)

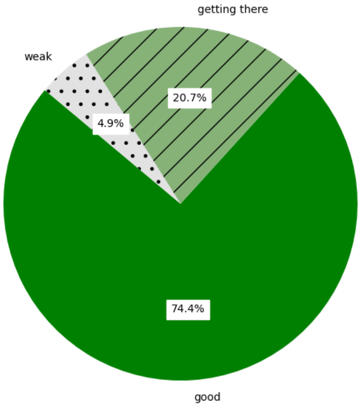


Figure 5.13: CS32 Quality Rating Breakdown (Selected Sample)

How was the quality of the students' responses? Subsequently, we looked into the quality of students' responses by checking the result of applying the Quality Rubric to Selected Sample. As shown in fig. 5.12 and fig. 5.13, the majority of responses in both courses were categorized as ”good,” with 59.7% in CS111 and 74.4% in CS32. The ”getting there” category comprised 35.2% of responses in CS111 and 20.7% in CS32. Notably, both courses exhibited a very low incidence of ”weak” responses, each accounting for approximately 5%.

Reflecting on the breakdown, it becomes evident that our initial concerns regarding the potential for uniformly weak responses were unfounded. Instead, the data reveals an interesting mixture of response qualities, with a significant portion classified as ”good” across both courses. This diversity in the quality of responses suggests that the contrasting cases approach did not result in monotonous outcomes. Instead, it prompts a deeper investigation into the underlying reasons for this variation. Such findings encourage us to further explore the details of these responses to uncover patterns or explanations for the observed mixture.

Were students consistently getting “good” because they were good writers? The prevalence of ”good” ratings in both fig. 5.12 and fig. 5.13 prompts an investigation into whether some students got consistent ”good” ratings and the quality rating was more indicative of inherent writing proficiency than students' performance in this specific study. Figure 5.14 and fig. 5.15 illustrate the distribution of each student's quality ratio in their responses. For example, a student is categorized as ”only good” in the charts if all their responses received a ”good” rating. As the data from

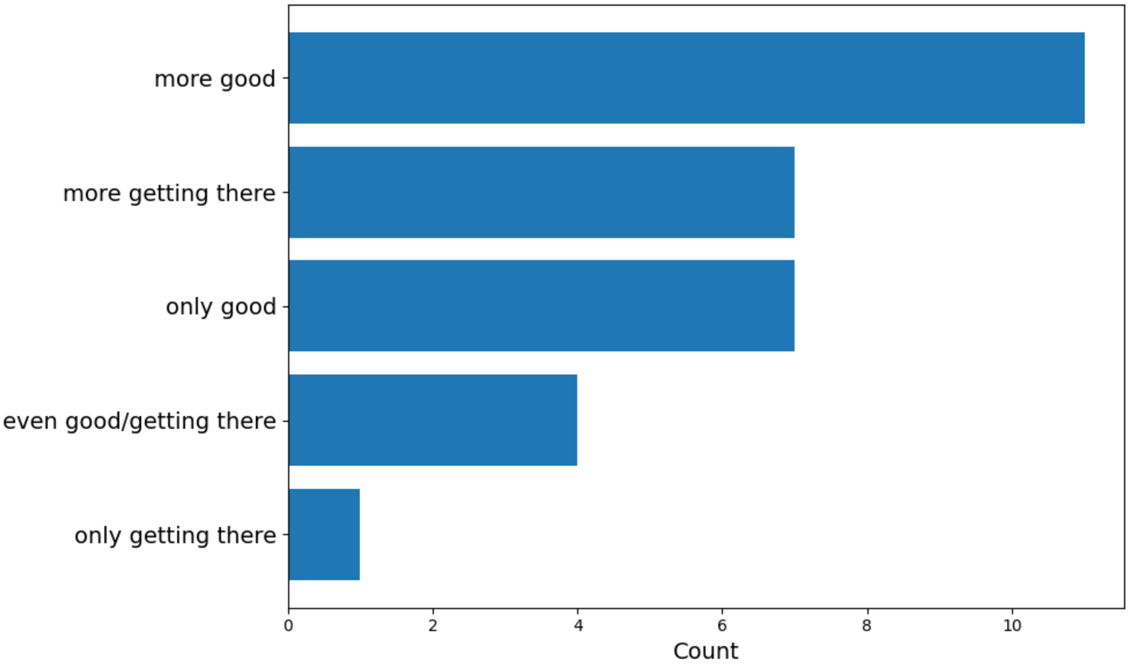


Figure 5.14: CS111 Frequency of Quality Ratio Description (Selected Sample)

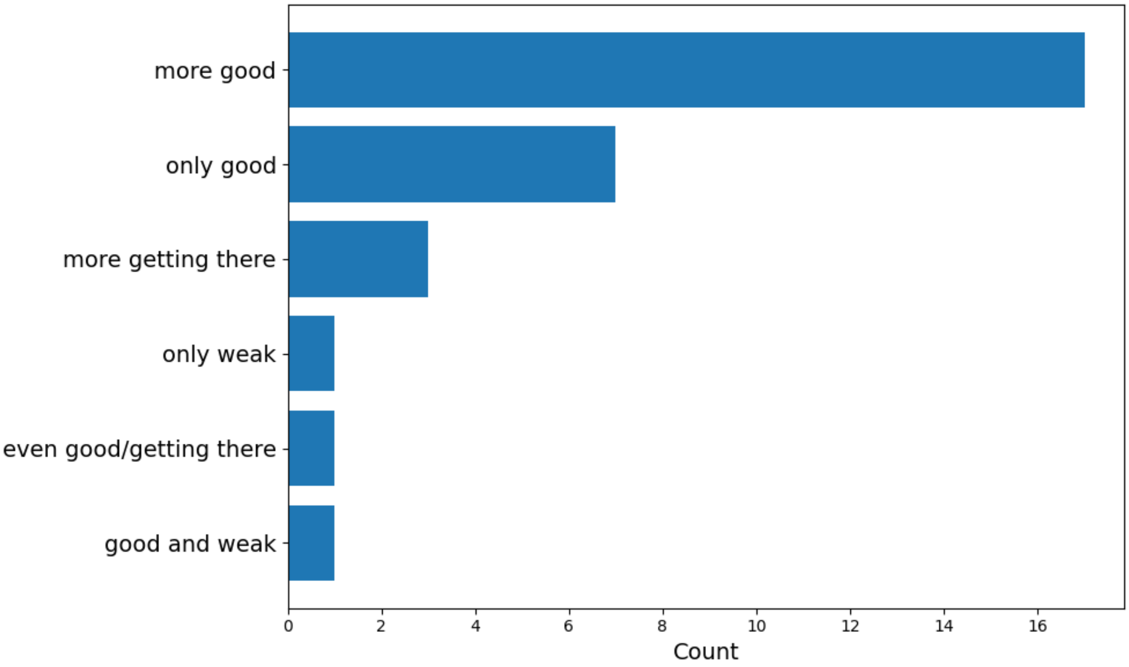


Figure 5.15: CS32 Frequency of Quality Ratio Description (Selected Sample)

fig. 5.14 and fig. 5.15 reveal, only 26.7% (8 out of the total 30) of students in both groups consistently received either ”only good” or ”only weak” quality ratings. The majority exhibited mixed quality in their responses, suggesting that factors other than mere writing skill might influence the quality of their response.

These checks ensured that we collected enough interesting responses with variation in quality, enabling us to proceed confidently to explore the underlying patterns of students' responses.

What categories did students bring up? Figure 5.16 and fig. 5.17 illustrate the frequency of each category's occurrence within the Selected Sample. Blue bars represent the categories that were pre-planted during the case design, whereas red bars represent new categories that emerged from students' responses. Additionally, each bar is segmented into patterns that denote the quality ratings: solid patterns indicate ”good”; lined patterns indicate ”getting there”; and dotted patterns indicate ”weak”.

Pre-Planted Categories in CS111 In fig. 5.16, the top two bars, automated tool and culture, both have a count over 30, which indicates that some students' responses got tagged multiple times under the category. For example, one student had a total of three factors listed, and all of them touched on some aspects of automated tool:

1. *Limited sample of data could cause false correlation” touches on the quality of data that's fed into the tool*
2. *A faulty content moderation system can dissuade users to continue using the platform” talks about the consequences of the tool making an error*
3. *Strict moderation systems, even if they are effective, can harm daily users” talks about the consequences of the tool being too strict*

Similarly, another student's response mostly focused on culture and language:

1. *People from different cultures could consider different things offensive.*
2. *One person using certain language could be perfectly acceptable while someone else using the same words could need to be monitored.*
3. *Someone could make an honest language mistake if they are not using their native language*

It's also worth noticing that for this student, all their responses got rated as ”getting there” for quality, because their arguments were missing the connection of how the factors affect the design of content moderation tools. In contrast, this response was rated as ”good”: *Depending on where someone is posting from, that could have an influence on their culture and language. The same marks that need to be moderated in one region might not be the same in another. Content moderation should have a way to be specific to different regions.*

Explanation to moderators was another pre-planted category within our case studies, and all the responses under this category received a ”good” rating. Notably, students referenced specific cases to support their arguments regarding explanations to moderators. For instance:

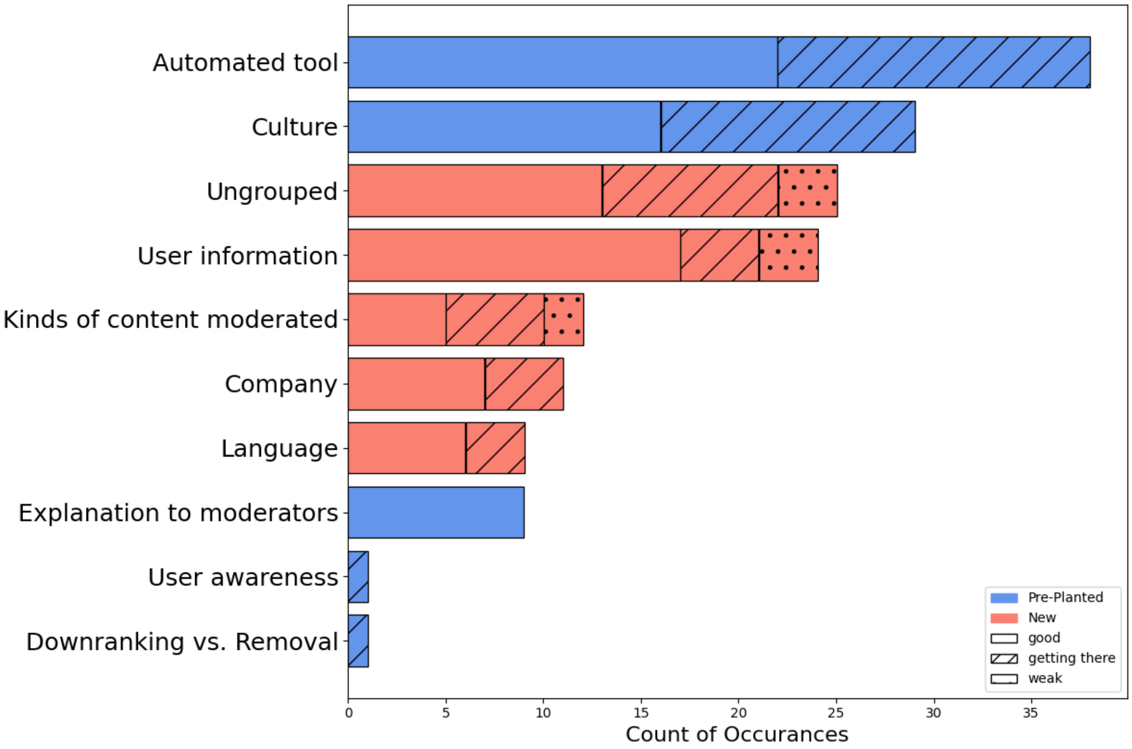


Figure 5.16: CS111 Category by Popularity and Quality (Selected Sample)

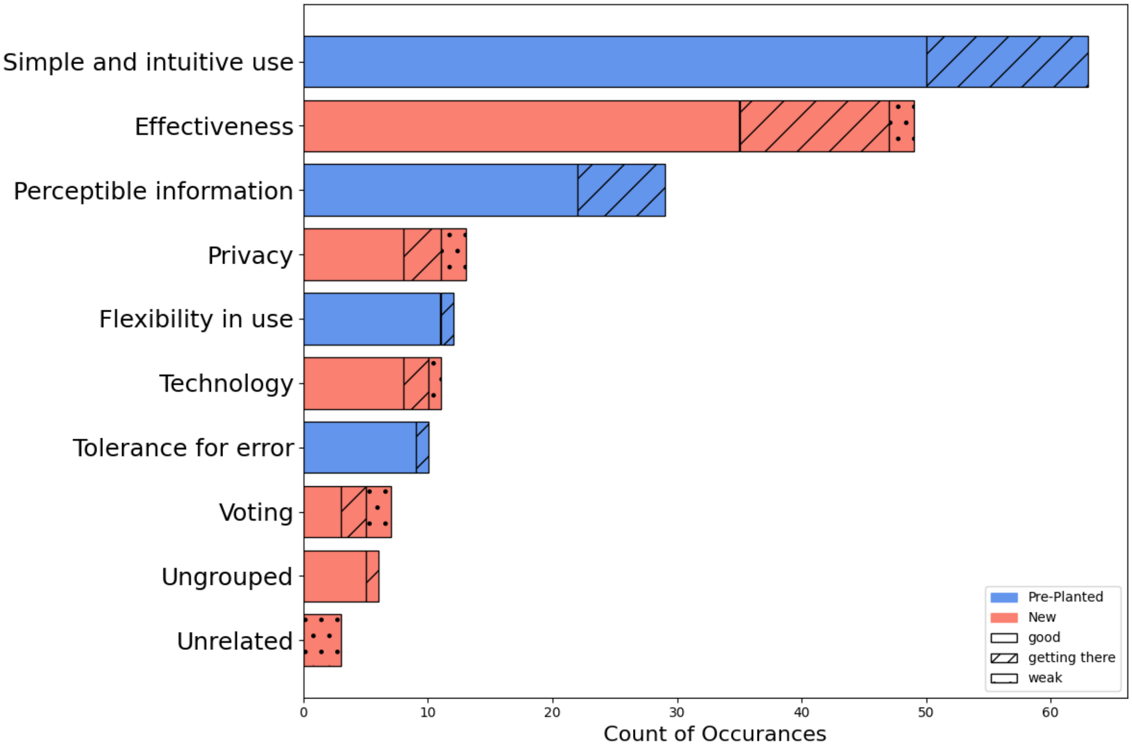


Figure 5.17: CS32 Category by Popularity and Quality (Selected Sample)

- One student mentioned case 1 explicitly: *Moderation tools should be transparent to users so that if content is justifiably flagged, users can prevent future offenses on the platform. In the case study above, both Alex and their boss are unclear on why the grandmother's post was flagged (this ambiguity could have adverse consequences on users, such as distrust and disengagement from the platform).*
- Another student mentioned case 2 explicitly: *As Case 2 shows, and as I've heard from learning about how AI moderation systems work, it can sometimes be a problem that human software engineers are unable to see the exact decision process of an AI.*

These instances suggest that the strategically embedded details in the cases were successfully noted by some students. However, the clarity of these details was not uniform across all participants. For example, in their written response, one student didn't have the critical "assume" part highlighted in case 1. But during the interview, when prompted to read the case again and talk about what additional details they found interesting, they picked up on this: *"I think it's kind of interesting that the automated flagging tool doesn't have any way of indicating why it flagged the post."* and further commented that *"It seems like it would be helpful in both evaluating and improving their automated tool, and also for people reviewing content manually to be able to see why a post was flagged."*

For the last two pre-planted categories, "user awareness" and "downranking vs. removal", each was identified by only one student. The perspective of "user awareness" was subtly incorporated only in case 2, and it was more suggested rather than explicitly stated: "the tool flags inappropriate posts (with a mark that is visible to all users of the platform)". Therefore, we weren't too surprised by the few mentioning of this category. On the other hand, the limited recognition of "downranking" was particularly unexpected, given its explicit mention in the case text and its emphasis as a keyword in the FAQ section, which we anticipated would enhance its visibility and make an impression on students. However, a subsequent review of the Overall Data revealed that "downranking" was seldom mentioned; a keyword search indicated that only 20 out of 161 students (approximately 12.4%) used the term "downrank" or its misspelled variant "down rank".

New Categories in CS111 As illustrated in fig. 5.16, 'user information' emerged as the most prominent new category, represented by the longest red bar. And in the codebook, we have developed several tags under 'user information' to capture the nuanced aspects of this category:

[age] age of the user, parental controls, consideration for kids or teens

[ux] user experience, user reaction, how content moderation impacts these

[feed] user feedback, incorporating user feedback into content moderation practices

[user] user's info, background, history of offenses, previous post, and using these information to do content moderation

Below are representative quotes from each tag:

1. **Age:** *"Children or minors could be exposed to sexual or violent content that is inappropriate and disturbing."*
2. **User experience:** *"An ongoing effort to eliminate bias in moderation systems is important as to not be counter-productive and anger users."*
3. **User feedback:** *"In order to build a platform that allows for transparency and user engagement, there must be a convenient system in place that encourages users to speak up when they view content that they think should be taken down."*
4. **History of offenses:** *"In ambiguous circumstances, checking the history of the user on the platform may lead into some insight on whether or not to flag an item. For example, someone who tweets something that comes off as offensive, and has a history of flagged tweets in the past, should be more intensely monitored and more quickly flagged."*

Ungrouped was another popular new category, representing a collection of diverse tags that did not fit into any other groups:

[crct] whether a correct answer or ground truth exists, difficult to get the 'right answer'

[fsp] freedom of speech, issues of censorship

[law] legal issues, regulations, country or state level policies

[priv] privacy, how user privacy is protected or violated

[ind] individual standards of what content should be moderated, what is considered "offensive", "inappropriate", or "immoral" (this is not linked to culture because it's based on individuals)

[intr] interactions with the content moderated, what responses/comments the content provoke, how large the size of audience is for a given post

[intn] user's intention (e.g., for educational purposes, malicious, political, etc.)

[bot] ways to deal with content generated by bots

[eth] whether the content is ethical, whether the moderation tool or human moderator is making ethical decisions.

Among these tags, freedom of speech, privacy, and user's intention emerged as the more popular ones. Below are representative quotes from each of these tags:

1. **Freedom of speech:** *"You can eliminate harmful content, but blocking certain content and eliminating freedom of speech is also not okay."*
2. **Privacy:** *"Make sure the content moderation process is not infringing on user privacy."*
3. **User's intention:** *"Many people may try to say certain things educationally or include certain keywords that could block them from expressing themselves."*

An interesting pattern emerged in the distribution of quality ratings in fig. 5.16: all responses classified as "weak" were associated with the newly introduced categories. This observation may stem from our quality rubric's setup. Specifically, the criteria for a "weak" rating included a focus on elements that were "not obviously a part of content moderation tools". As a result, the pre-planted categories, which were explicitly designed to align with the content moderation tools, were more likely to receive higher quality ratings compared to the new categories. This indicates that our evaluation methodology may have inadvertently favored pre-planted categories.

Pre-Planted categories in CS32 In fig. 5.17, the two longest blue bars represent the pre-planted categories "simple and intuitive use" and "perceptible information", each with counts exceeding 30. This suggest that multiple tags were assigned to responses within these categories, reflecting a range of perspectives from students. Under 'simple and intuitive use,' we developed several tags based on the diverse arguments presented by students. Below are these tags along with representative quotes:

1. **Language:** *"Users who are non-native English speakers could have a more difficult time identifying and typing out the word."*
2. **Deterrence from usage:** *"If the CAPTCHA takes a long time, users may not want to waste more time to vote."*
3. **Context:** *"Make sure that the training data includes multiple cultures in the case that it is an image, for instance, that has a different meaning for different groups. Alternatively, choose universally recognizable items to identify."*
4. **Instructions to user:** *"Need to make sure the completion steps are clear. I think of Captchas where I select all squares with motorcycles, and I am not sure whether to click the square with a tiny piece of the tire in it."*
5. **Age (old):** *"If the CAPTCHA is too difficult, it will prevent not only bots but also real humans from using the site, especially as technological rifts between generations grow with those who are older and may struggle to follow complex instructions."*
6. **Age (young):** *"I wonder if age will affect a user's ability to perform the task well, especially for younger users."*

In fig. 5.17, only four blue bars are visible, one fewer than the five pre-planted categories we introduced during the case design phase. Notably, the category "equitable use" was not identified by any student in Selected Sample. While several students raised concerns related to privacy, "equitable use" encompasses a broader principle: "provisions for privacy, security, and safety should be equally available to all users." This aspect was subtly suggested in the cases through setups involving non-interactive options for CAPTCHA, which required more data from the user than the interactive alternatives. Despite its subtlety, during the interviews, some students were able to articulate this tradeoff: *"The tradeoff is that you're giving up your own personal data, which in this case includes cookies. [...] it's something to think about. If it's worth that tradeoff."*

New categories in CS32 The most prominent new category represented by the longest red bar in fig. 5.17 is ”effectiveness”. This category frequently appeared together with pre-planted categories such as ”simple and intuitive use” and ”flexibility in use”. Students frequently discussed the critical balance and tradeoffs involved in designing CAPTCHA tasks that are sufficiently straightforward for human users but challenging enough to deter bots. We suspect that this emphasis on efficiency and concerns about bot attacks was heightened by the repeated mentions of bots and brute force attacks in the case studies, as well as the direct questions posed in the FAQs. This exposure likely prompted students to prioritize these aspects in their responses.

Below are two representative quotes that exemplify these combined concerns:

- **Effectiveness and simple and intuitive use:** *“Question difficulty - The questions should be easy enough that a variety of users can pass them, but not easy for robots to bypass.”*
- **Effectiveness and flexibility in use:** *“Limited time, infinite attempts - In case 2, they mentioned only giving the user 1 minute to complete the captcha successfully, but a bot could likely submit thousands or even hundreds of thousands of attempts within a minute, thus allowing them to complete the captcha successfully. I think limited attempts is better than limiting the time in terms of bot prevention.”*

Among the new categories, ”voting” was notably unexpected. This category arose because students noticed the context specified in the task 2 prompt: ”for an online voting platform”. Although this detail was intended merely as an example of another potential application for CAPTCHA, not a focus of the study, it unexpectedly captured some students’ attention, underscoring their sensitivity to the context of where the system is employed. Interestingly, recognizing the ”voting” context enabled some students to formulate relevant arguments concerning CAPTCHA design. For instance, one student noted, *”If the CAPTCHA involves viewing images, in the specific context of an online voting platform, it’s essential to ensure these do not indicate support for any one candidate or group. Even something subtle like the color scheme could indicate some form of bias.”* Conversely, other students got sidetracked, focusing their discussions exclusively on aspects of the voting platform unrelated to CAPTCHA, such as *”Having someone vote multiple times if you only want them to vote once. Aka, could be using a VPN or a different IP address.”*

Technology was another new category that emerged unexpectedly. It primarily resulted from students elaborating on specific technological details mentioned in the cases, such as the necessity to accept various types of email addresses (not limited to Google) and ensuring compatibility with multiple web browsers (not exclusively Chrome). A selected student quote highlights this focus: *”It is important to consider which browsers the CAPTCHA will cover, and the justifications need to be clear to match the target customer. In case study 2, for example, a user must use Google.”*

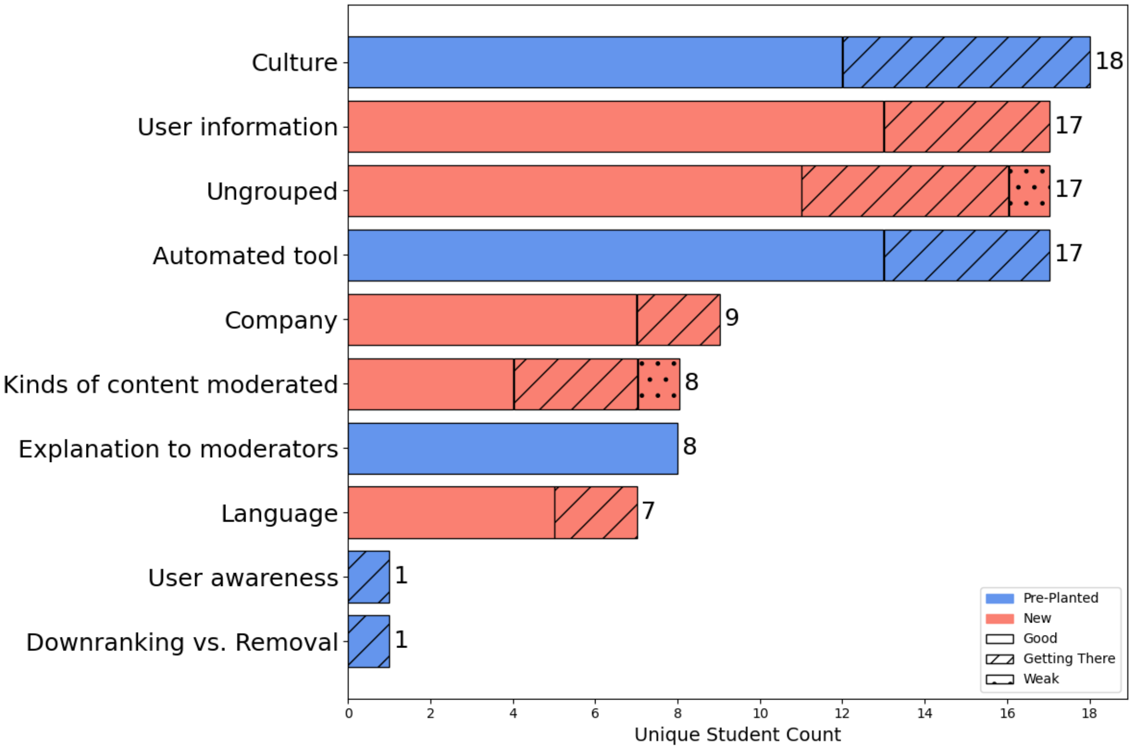


Figure 5.18: CS111 Category by Unique Student Count (Selected Sample)

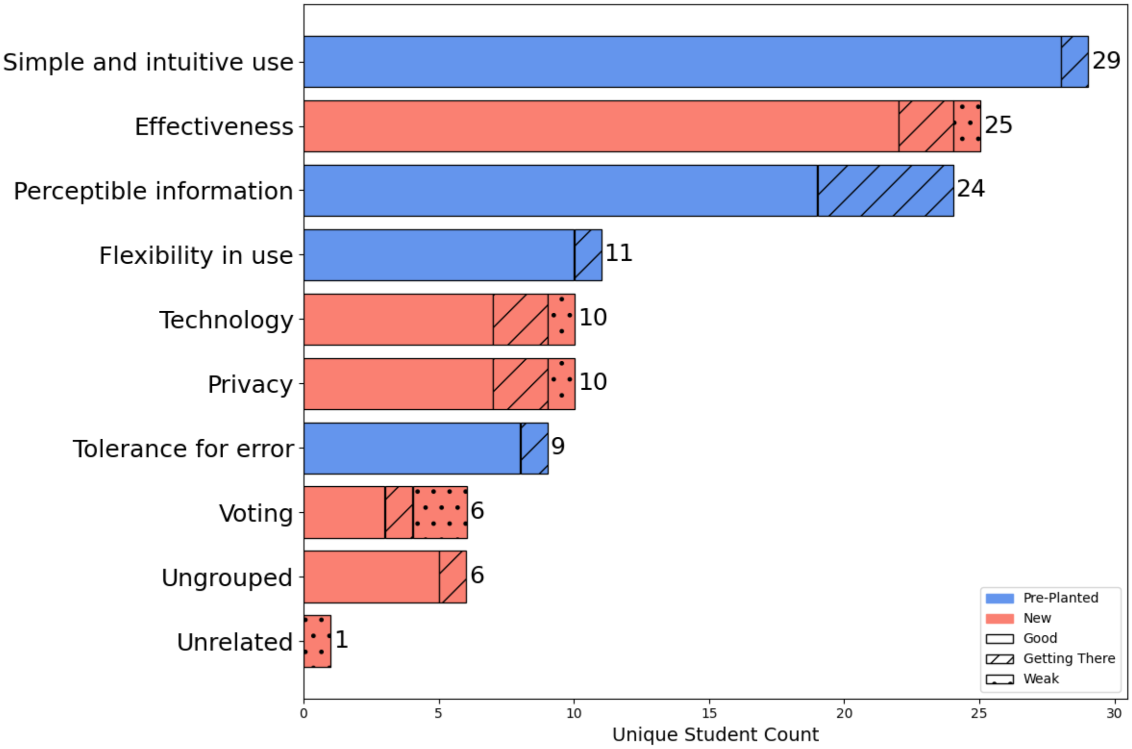


Figure 5.19: CS32 Category by Unique Student Count (Selected Sample)

How many unique students brought up each category? Figure 5.18 and fig. 5.19 offer a different perspective on category counts by focusing on unique student rather than total occurrences. Similar to fig. 5.16 and fig. 5.17, blue and red bars represent pre-planted and new categories respectively, with internal segment patterns indicating different quality ratings. In cases where a student received multiple tags under the same category, only their highest rating is reflected in these graphs. For example, if a student received a "good" and a "getting there" rating under the "culture" category, they are counted in the "good" segment for that category.

According to fig. 5.18, approximately half of the students in CS111 were tagged under the categories "culture", "user information", and "automated tool". Similarly, fig. 5.19 shows that nearly all students in CS32 engaged with the categories "simple and intuitive use", "effectiveness", and "perceptible information", with a majority achieving at least one "good" rating in these areas.

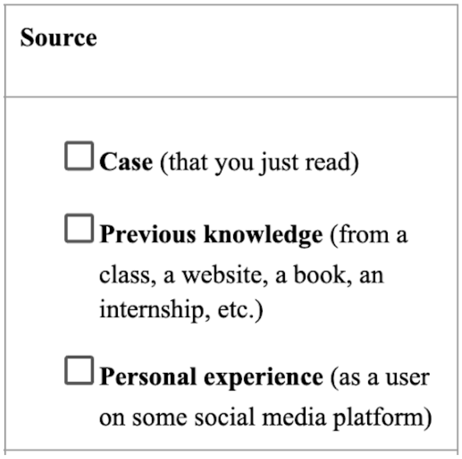
RQ 1 Summary To address Research Question 1—Do students produce beyond shallow answers when engaged in contrasting cases exercises?—our analysis revealed that students from both courses typically listed four to five different factors on average. The quality of responses varied both across and within individual students. In both courses, students prominently identified two out of the five pre-planted categories, while the remaining pre-planted categories did not resonate as strongly as we had anticipated. Notably, several pre-planted categories were elaborated upon extensively, demonstrating that students formulated arguments from diverse perspectives. Additionally, some new categories were also popular and were developed extensively with numerous tags underneath. Overall, while students did not mention all the pre-planted categories, they introduced several unexpected and interesting new categories.

5.4.2 RQ 1.1. How does the level of background knowledge relate to the quality of students' answers?

How much background knowledge did students claim to use? In the written task 2, students were asked to fill out a table of factors, and for each factor, they were instructed to select one or multiple checkboxes to indicate how they thought of it, as shown in fig. 5.20.

Table 5.5 and table 5.6 detail the counts and ratios of the checkbox combinations selected by students in the Overall Data in CS111 and CS32.

To conduct a deeper analysis on whether students based their responses on the case studies, we refined our dataset by filtering out some rows from the table. This was done to focus exclusively on entries that clearly distinguished between "from case" versus "not from case" responses. As seen in table 5.5, the last four rows of the table were excluded for further analysis, as these did not



Source

□ **Case** (that you just read)

□ **Previous knowledge** (from a class, a website, a book, an internship, etc.)

□ **Personal experience** (as a user on some social media platform)

Figure 5.20: Source Checkboxes Presented to Students

allow for a clear determination of whether responses were derived from the case studies or from the students' own background knowledge.

The rows in table 5.5 and table 5.6 are organized by popularity of the response sources. By comparison, ”Cases” was a significantly more popular source in CS32 (accounting for 39.8%) than in CS111 (27.3%). The combination of ”Previous knowledge + Personal experience” ranked as the fourth most popular source in CS111, accounting for 10% of responses, whereas in CS32, it was only the sixth most popular, contributing to 3.9% of responses. These findings suggest that students in CS32 were more inclined to base their responses on the case studies. Additionally, while coding the Selected Sample, we observed a similar pattern: there were very few direct mentions of the cases in CS111, in contrast to CS32, where direct citations were much more frequent.

What kind of background knowledge did students have? To gain deeper insights into the students' background knowledge, all participants were asked during the interview to provide context about what they knew before engaging with the exercise, and were encouraged to elaborate on their answers and provide specific examples from their experiences. Table 5.7 presents a summary of the students' answers during the interview.

As shown in the ”Class Experience” row, while students from both classes cited experiences from coursework, the majority in CS32 referenced courses related to computer science, with a notable exception noted in the last row—a psychology class. Conversely, students in CS111 reported taking relevant courses across a diverse range of departments.

Reflecting on personal experiences as users, students in CS111 frequently discussed their own

Source Checkboxes Picked	**Count**	**Ratio (%)**
Previous knowledge	213	27.3
Cases	181	23.2
Personal experience	151	19.4
Previous knowledge + Personal experience	78	10.0
Cases + Previous knowledge	66	8.5
Cases + Personal experience	46	5.9
Cases + Previous knowledge + Personal experience	31	4.0
Blank	13	1.7

Table 5.5: CS111 Source Checkboxes Picked by Count and Ratio (Overall Data)

Source Checkboxes Picked	**Count**	**Ratio (%)**
Cases	174	39.8
Previous knowledge	77	17.6
Personal experience	58	13.3
Cases + Previous knowledge	52	11.9
Cases + Personal experience	34	7.8
Previous knowledge + Personal experience	17	3.9
Cases + Previous knowledge + Personal experience	14	3.2
Blank	11	2.5

Table 5.6: CS32 Source Checkboxes Picked by Count and Ratio (Overall Data)

and their friends' experiences as content creators and viewers on various social media platforms. In contrast, students in CS32 revealed two notable patterns when discussing their use of CAPTCHA. Firstly, many students acknowledged frequent encounters with CAPTCHA but admitted they had not considered its implications deeply. Secondly, some students emphasized that while they had no personal difficulties in completing CAPTCHA challenges, they recognized that these tasks could be more challenging for certain user groups.

Regarding prior knowledge, a significant number of CS111 students cited news articles as their information source. In contrast, students in CS32 tended to reference their technical knowledge, particularly related to software and front-end design.

Do we see any connections between quality rating and source? Figure 5.21 and fig. 5.22 depict the counts of each quality rating and the breakdown of sources, distinguishing between "from case" and "not from case." A striking observation between the two courses is the nearly inverted color segmentation in the figures, which indicates that the majority of responses in CS111 were "not from case," whereas in CS32, "from case" responses prevailed. This trend aligns with our earlier findings: students in CS32 frequently cited the case studies directly in their written responses.

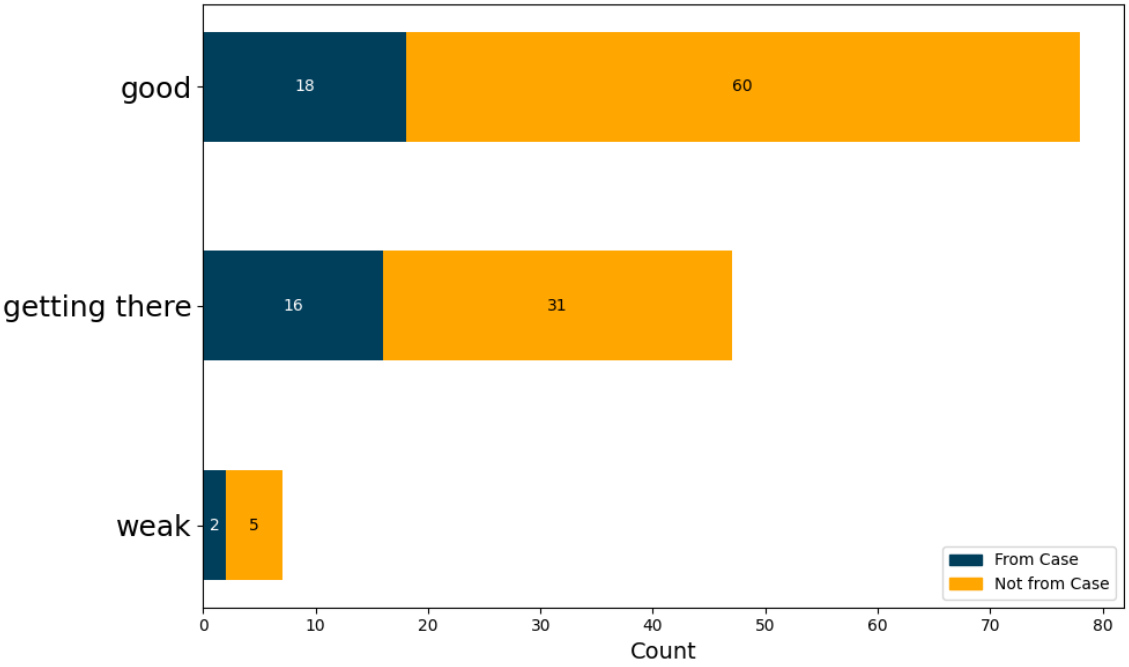


Figure 5.21: CS111 Quality Rating and Source Breakdown (Selected Sample)

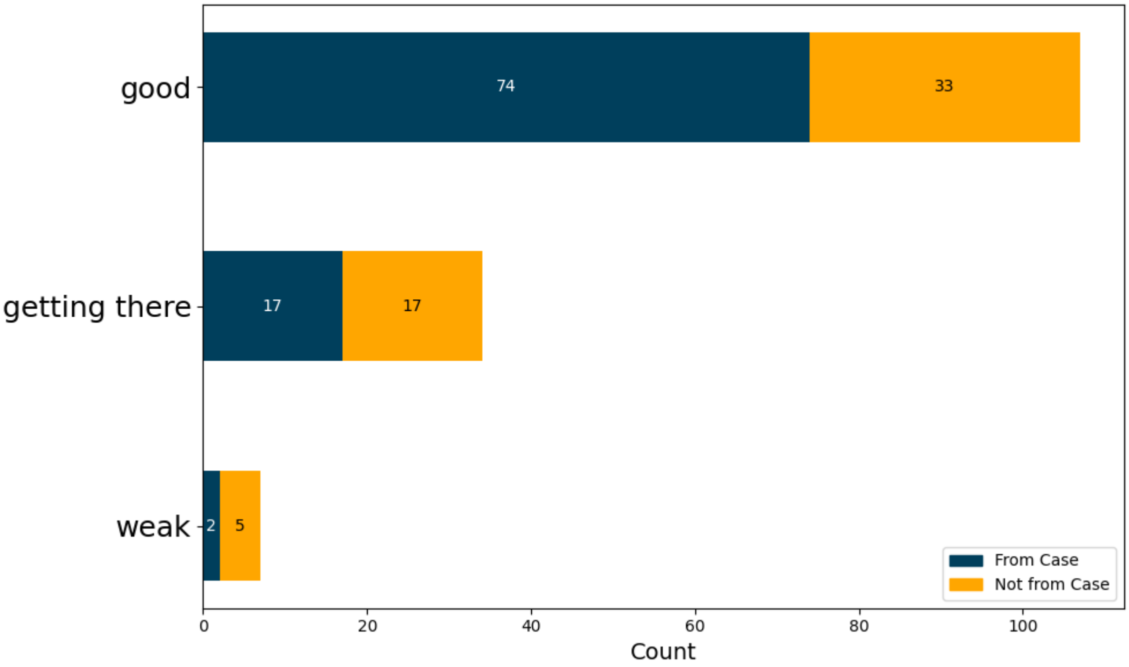


Figure 5.22: CS32 Quality Rating and Source Breakdown (Selected Sample)

	CS111	**CS32**
Class Experience	• Targeted ads exercise in CS111 • Philosophy class, Ethics of Digital Technology • History class • A class in the language department which talked a lot about social media	• SRC assignment on accessibility from CS15 • Cyber security ethics class • UIUX class • Psychology class on how the choice of words impacts decision making
Other Experience	• Instagram and content warning label • YouTube and demonetization • Twitter and free speech • Covid and misinformation • Social media in both Korea and the US	• Using various CAPTCHA, finding some easier to use than others • Experience with front-end and software development in general

Table 5.7: Example Sources of Students' Background Knowledge (Interview)

Additionally, during the interview, most students in CS111 reported having prior knowledge of content moderation and building their responses based on these knowledge, while many CS32 students noted that despite frequently encountering CAPTCHA as a user, they had not given it much thought before doing the exercise.

We conducted a Chi-Square test to determine if there was a statistical relationship between the two categorical variables depicted in fig. 5.21 and fig. 5.22: quality rating and source. The results were different for the two courses. In CS111, the test returned a p-value of 0.41 and an effect size of 0.12, indicating a small and statistically insignificant association. Conversely, for CS32, the Chi-Square test yielded a p-value of 0.02 and an effect size of 0.22, suggesting a moderate and statistically significant association. These findings are consistent with our previous observations that the case studies had more impacts on students in CS32, though the effect was not substantial.

RQ 1.1 Summary We posed the research question: How does the level of background knowledge relate to the quality of students' answers? We hypothesized that the impact of contrasting cases would be robust enough to render the level of background knowledge a non-decisive factor in response quality. Specifically, we anticipated observing a pattern where responses derived "from cases" would yield higher quality due to the students' ability to construct well-formed arguments using the carefully designed details from the cases.

Although the actual pattern observed (as shown in fig. 5.21 and fig. 5.22) did not perfectly align with our hypothesis, the findings did reveal that students in CS111 and CS32 have very different levels of background knowledge. And notably, CS32 students, who generally had less background knowledge, tended to rely more heavily on the cases when constructing their answers.

5.4.3 RQ 2. In what ways do students delve deeper when getting two cases rather than one?

When setting up the study, students in both courses were randomly assigned to one of three groups, each receiving different versions of the case studies but similar prompts. This design allowed us to observe the actual effects of using contrasting cases:

- Group A received a pair of contrasting cases.
- Group B received only Case 1.
- Group C received only Case 2.

The subsequent analysis explores various aspects of how the groups compare to one another, examining whether students in Group A demonstrated better performance and providing deeper insights into the effectiveness of contrasting cases.

Did the number of factors mentioned vary among the groups? Figure 5.23 and fig. 5.24 display scatter plots where each dot represents an individual student, positioned on the X-axis according to the number of factors mentioned. Most vertical lines across the groups did not show a predominant color, indicating similar distributions in the number of factors among the groups. However, group A in CS32 distinguished itself with the blue dots clustered more towards the right, suggesting that students in this group mentioned slightly more factors compared to the other groups. This pattern suggests that the contrasting cases may have encouraged students in CS32 to identify more features. However, further analysis is required to determine the magnitude of this impact.

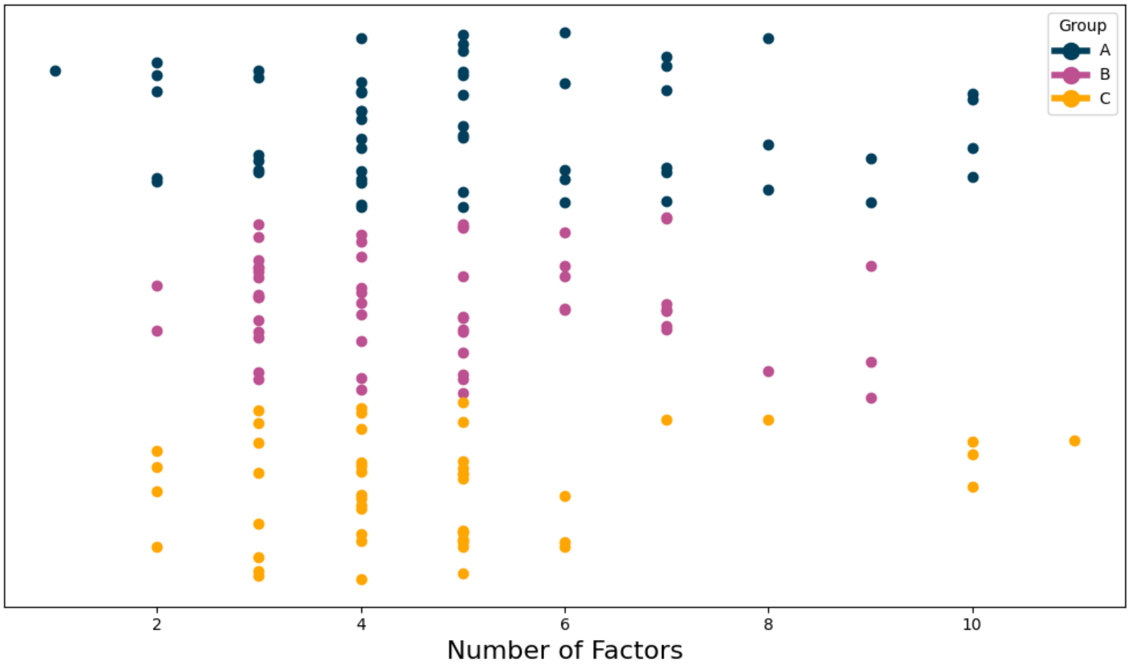


Figure 5.23: CS111 Number of Factors by Group (Overall Data)

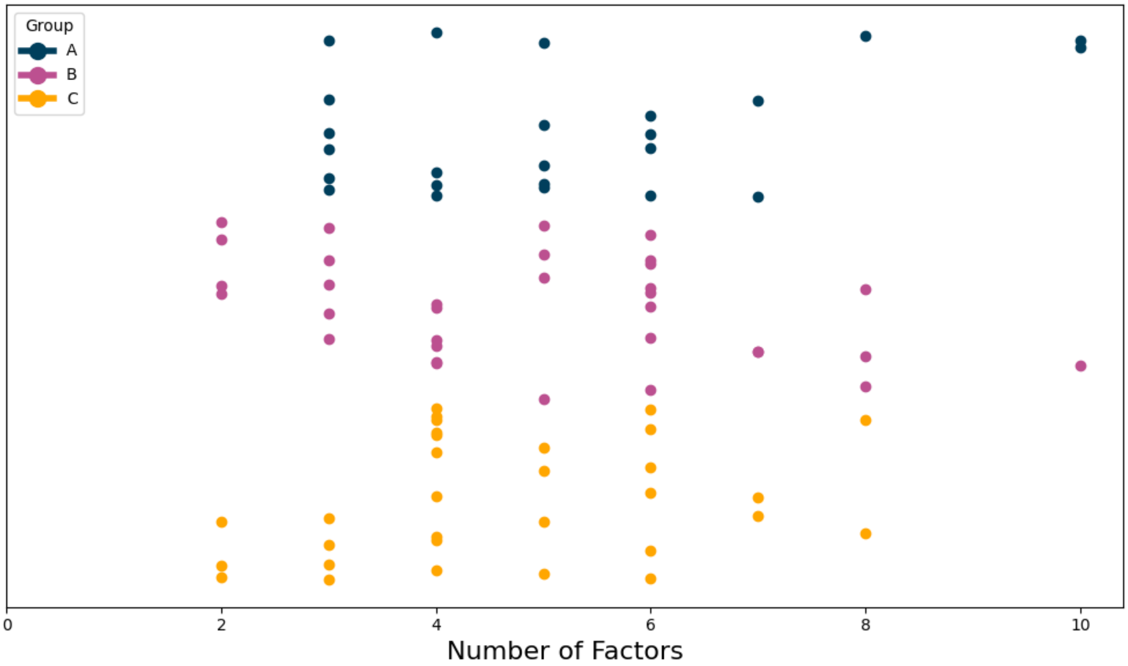


Figure 5.24: CS32 Number of Factors by Group (Overall Data)

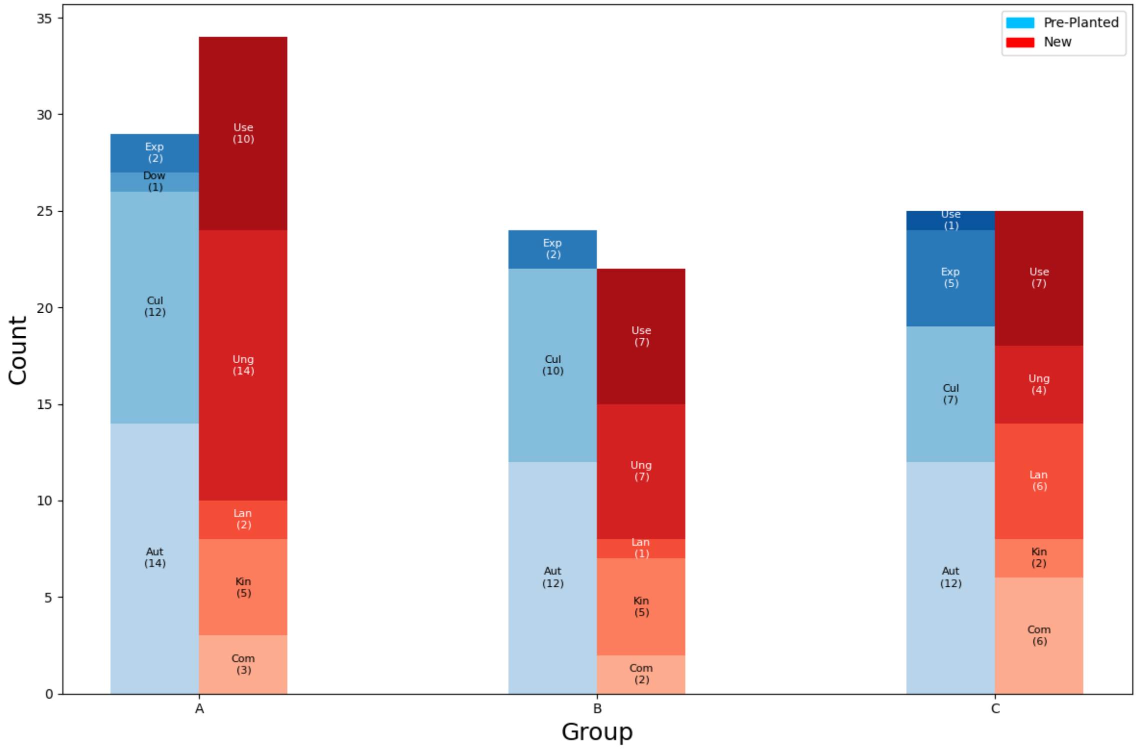


Figure 5.25: CS111 Category Breakdown in Each Group (Selected Sample)

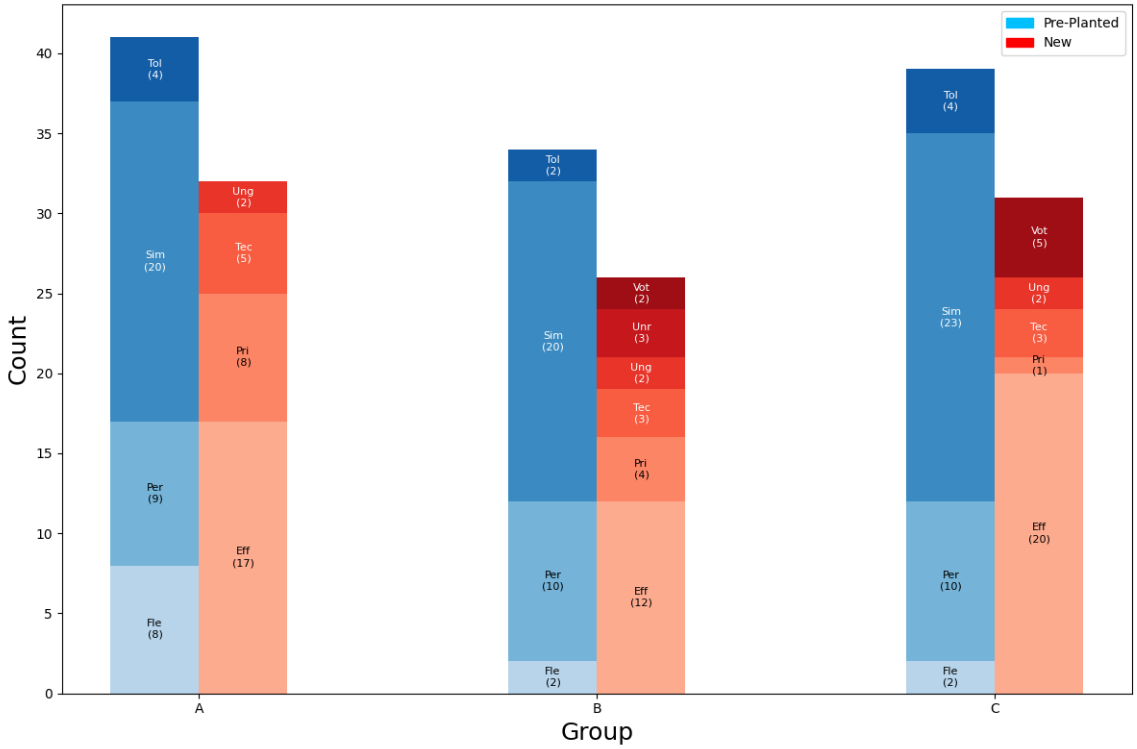


Figure 5.26: CS32 Category Breakdown in Each Group (Selected Sample)

What categories did each group bring up? Figure 5.25 and fig. 5.26 illustrate the count of occurrences for each category within the groups. Blue bars represent the pre-planted categories, and red bars represent the new categories. Within these bars, different shades of blue and red segment the individual categories, providing a visual breakdown of their frequency.

In fig. 5.25, the bars representing Group A are taller than those for Groups B and C, suggesting that students in Group A from the Selected Sample produced responses that covered more categories. Further analysis of the average number of characters in the Overall Data supports this: in CS111, Group A's responses averaged 210 characters, compared to 192 in Group B and 175 in Group C. This indicates that students in Group A in CS111 tended to write slightly longer and more detailed responses than those in the other groups.

Conversely, in fig. 5.26, although the bars for Group A are also taller, the character count analysis presents a different picture: Group A averaged 108 characters, while Group B averaged 169 and Group C 253. This suggests that in CS32, Group A's responses were not longer, challenging the assumption that taller bars correlate with more comprehensive content. This discrepancy suggests that the Selected Sample in CS32 might not be fully representative of the Overall Data, though it does not eliminate the possibility that Group A's shorter responses could still be richer in content.

In fig. 5.25, the distribution of different shades across groups does not show significant variation, indicating that each group touched on roughly similar categories in CS111. However, there were some variations in CS32. In fig. 5.26, the segment representing ”vote” (darkest red on top) is present only in groups B and C. This could be attributed to sampling variations, as the Overall Data shows that some students in group A also mentioned ”vote” and ”voting.” Another notable difference is observed in the category ”effectiveness” (light pink at the bottom), which occupies a larger proportion in group C (a total of 20 mentions) compared to group A (17 mentions) and group B (12 mentions). This pattern suggests that exposure to only Case 2 might have prompted students in group C to particularly focus on designing an effective CAPTCHA system capable of distinguishing humans from bots.

Was any category mainly brought up by one group? Figure 5.27 and fig. 5.28 offer an additional perspective to compare the categories across the groups. In these graphs, each bar represents the total number of unique students who mentioned a specific category, segmented by group. In both figures, no single group color dominates any of the bars, indicating that the distribution of category mentions is relatively even across all groups in both courses.

How did the groups compare in terms of quality? Figure 5.29 displays the distribution of quality ratings among each group in CS111, showing a similar ”good” versus ”getting there” ratio

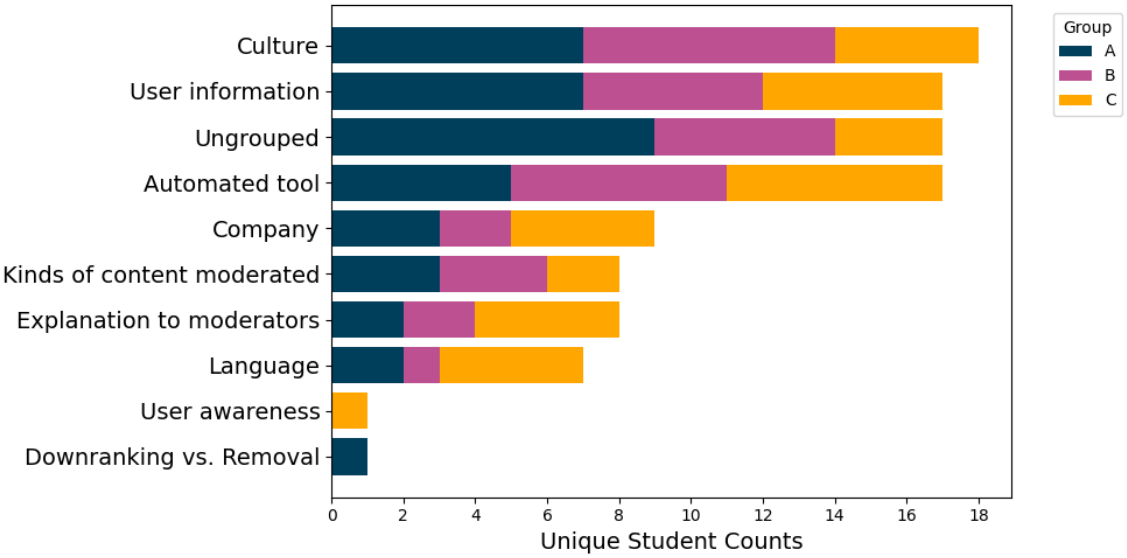


Figure 5.27: CS111 Category Breakdown by Group (Selected Sample)

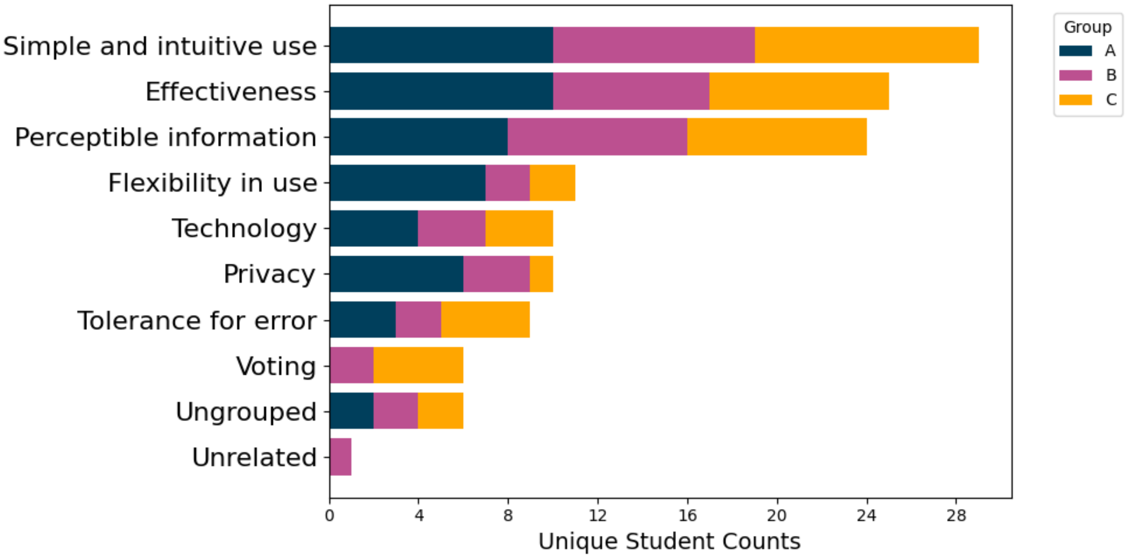


Figure 5.28: CS32 Category Breakdown by Group (Selected Sample)

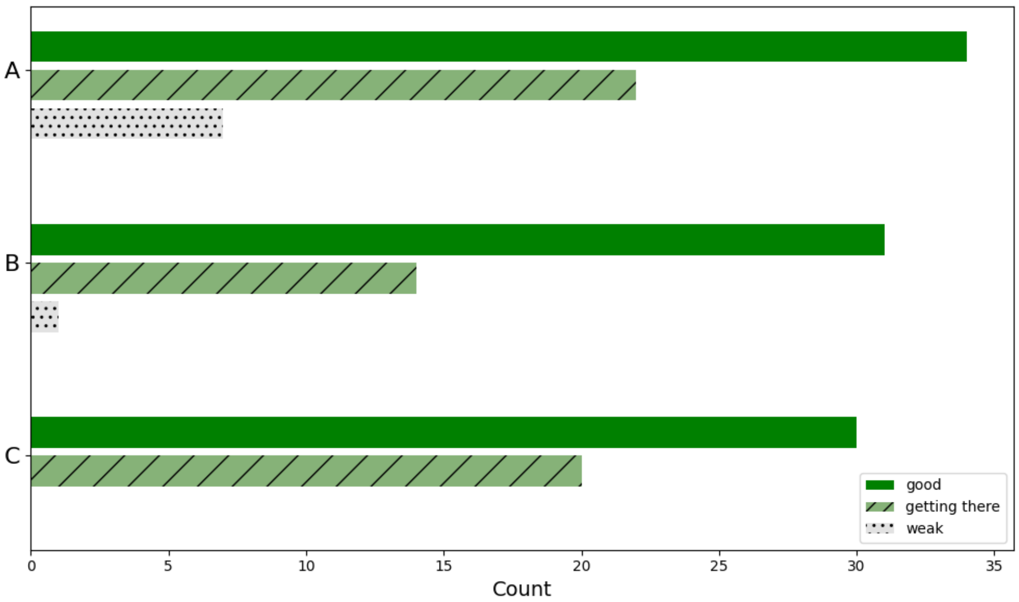


Figure 5.29: CS111 Quality Rating by Group (Selected Sample)

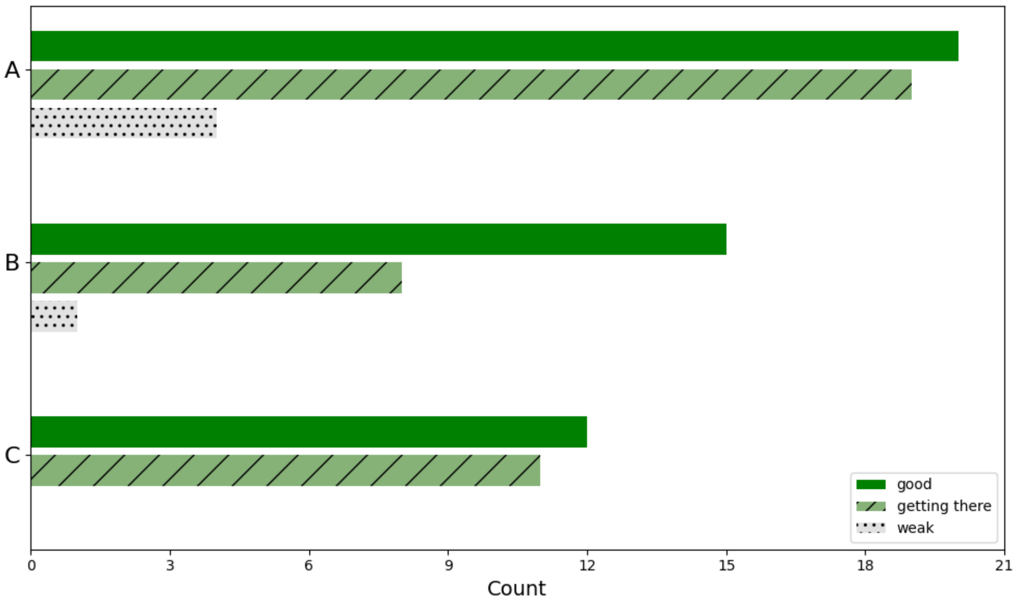


Figure 5.30: CS111 Quality Rating by Group (Filtered)

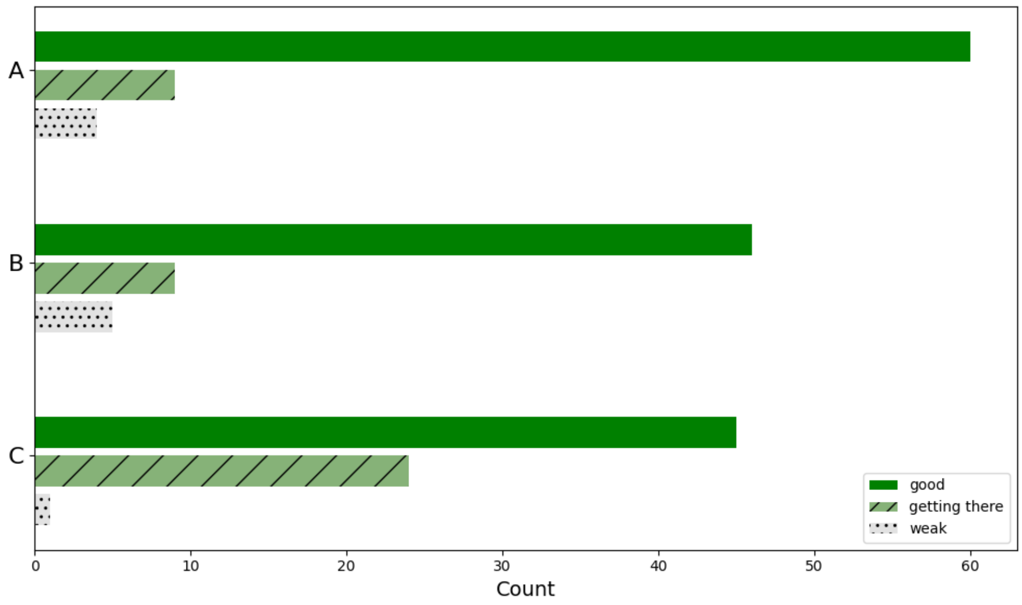


Figure 5.31: CS32 Quality Rating by Group (Selected Sample)

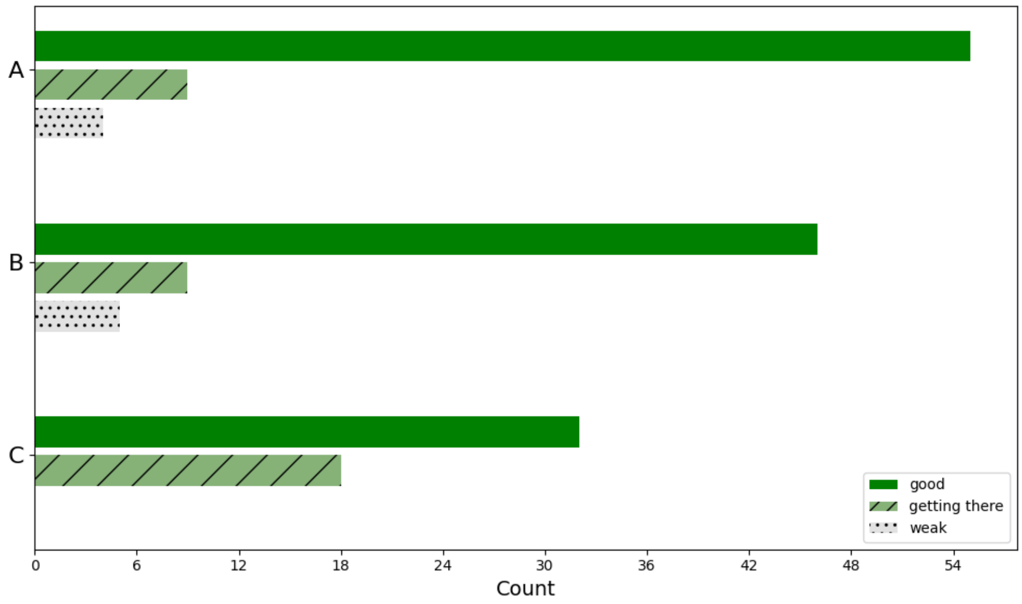


Figure 5.32: CS32 Quality Rating by Group (Filtered)

across the groups. Figure 5.30 presents a modified version of this graph, excluding students who checked many ”not from case” options, indicative of a high level of background knowledge. This filtering aims to isolate and better compare the responses that made use of the case studies. In this refined analysis, the ratio of ”good” to ”getting there” shifts, with Group A showing a relatively higher proportion of ”getting there” ratings compared to ”good,” suggesting that group A may be performing slightly worse than groups B and C.

Similar to the graphs for CS111, fig. 5.31 illustrates the distribution of quality ratings among each group in CS32. Notably, Group C is distinguishable from the other two groups as the ”getting there” bar is much longer. Figure 5.32 provides a filtered view, excluding students with a high incidence of ”not from case” responses to focus on those relying primarily on the cases. After this adjustment, the ratio of ”good” to ”getting there” shows minimal change, indicating that Group C might still be underperforming relative to Groups A and B.

Did any group pick more “from case” as source? Figure 5.33 and fig. 5.34 display the counts of responses classified as ”from case” versus ”not from case” within each group. In CS111, as depicted in fig. 5.33, the distribution across all groups is relatively consistent, with ”not from case” responses forming the majority. In contrast, fig. 5.34 for CS32 shows a different pattern; Groups A and B predominantly feature ”from case” responses, while Group C exhibits a more balanced ratio of ”from case” to ”not from case.”

RQ 2 Summary We explored the research question: In what ways do students delve deeper when provided with two cases rather than one? We hypothesized that group A, which received contrasting cases, would outperform groups B and C, who were given only a single case. This expectation was based on the assumption that group A would leverage the carefully constructed contrasts within the cases to produce more in-depth responses. However, upon examining the differences among the groups in terms of the number of factors listed, the variety of categories mentioned, quality ratings, and sources of information, we did not find any patterns that strongly support this hypothesis.

5.4.4 RQ 3. How well can students apply their knowledge and skills learned to a different scenario?

Task 2 in the written assignment presented students with a new scenario, asking them to list factors for consideration related to the central topic (complete prompts can be found in fig. 5.1). We opted against directly prompting students to analyze the cases to prevent the task from becoming merely an exercise in reading comprehension. Instead, we kept the prompt open-ended to see if students

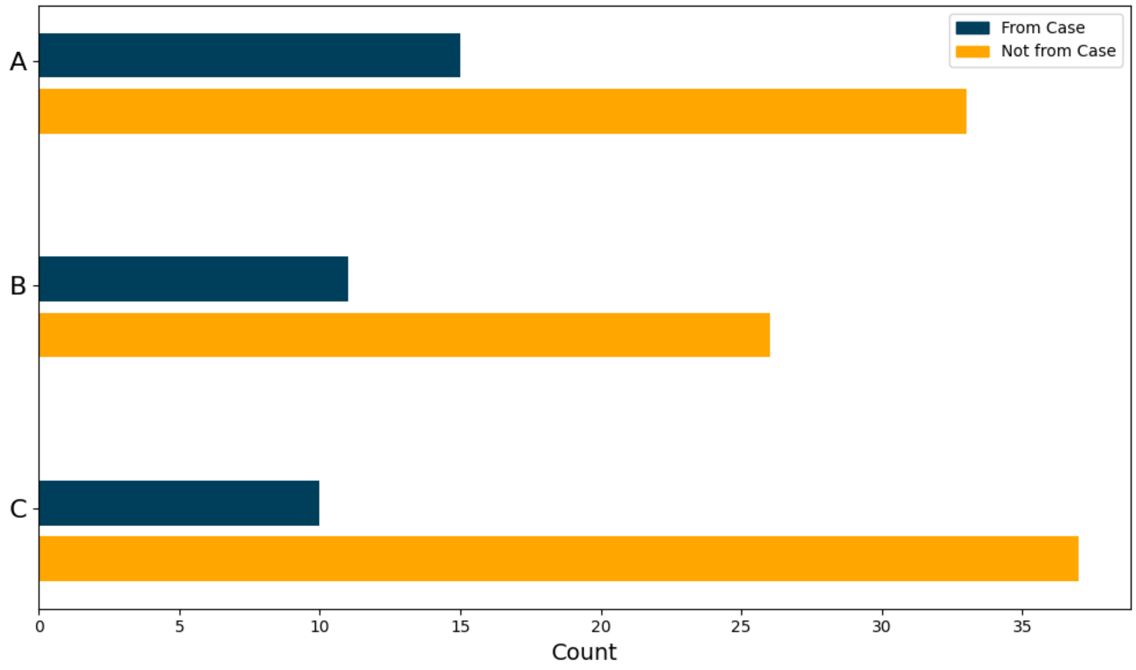


Figure 5.33: CS111 Source by Group (Selected Sample)

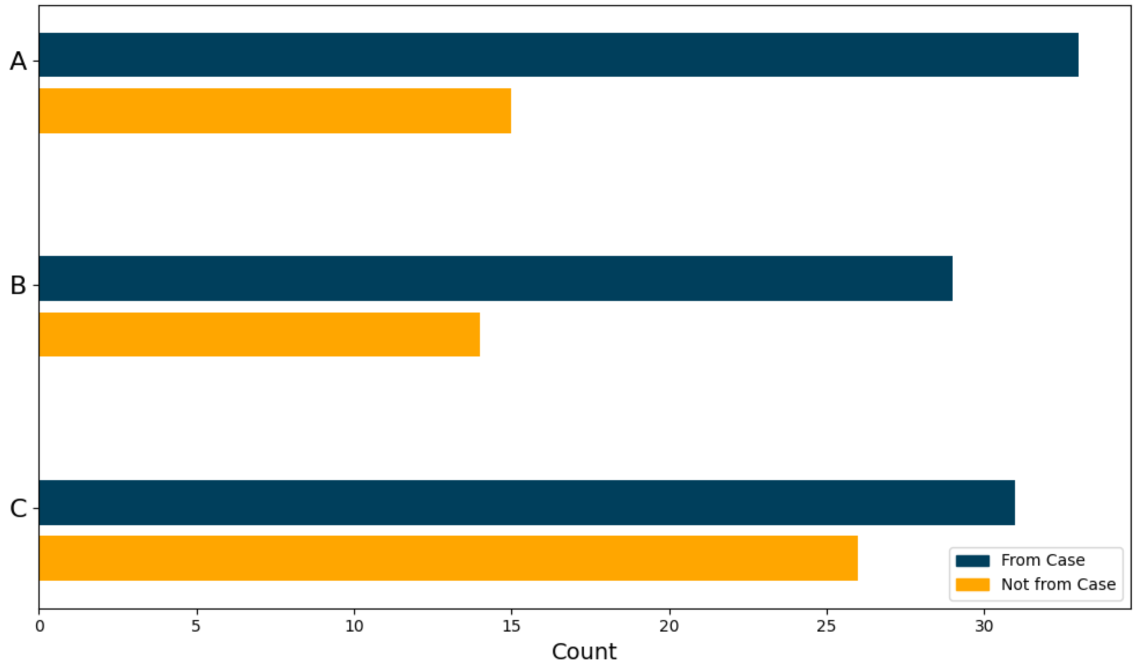


Figure 5.34: CS32 Source by Group (Selected Sample)

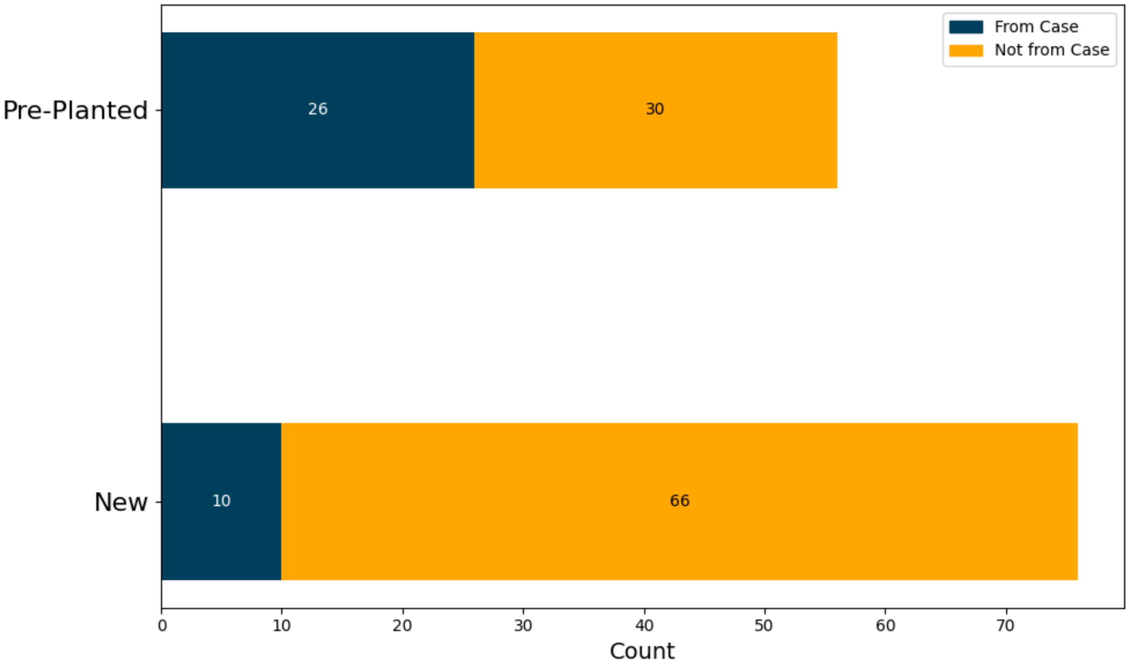


Figure 5.35: CS111 Source-Category Breakdown (Selected Sample)

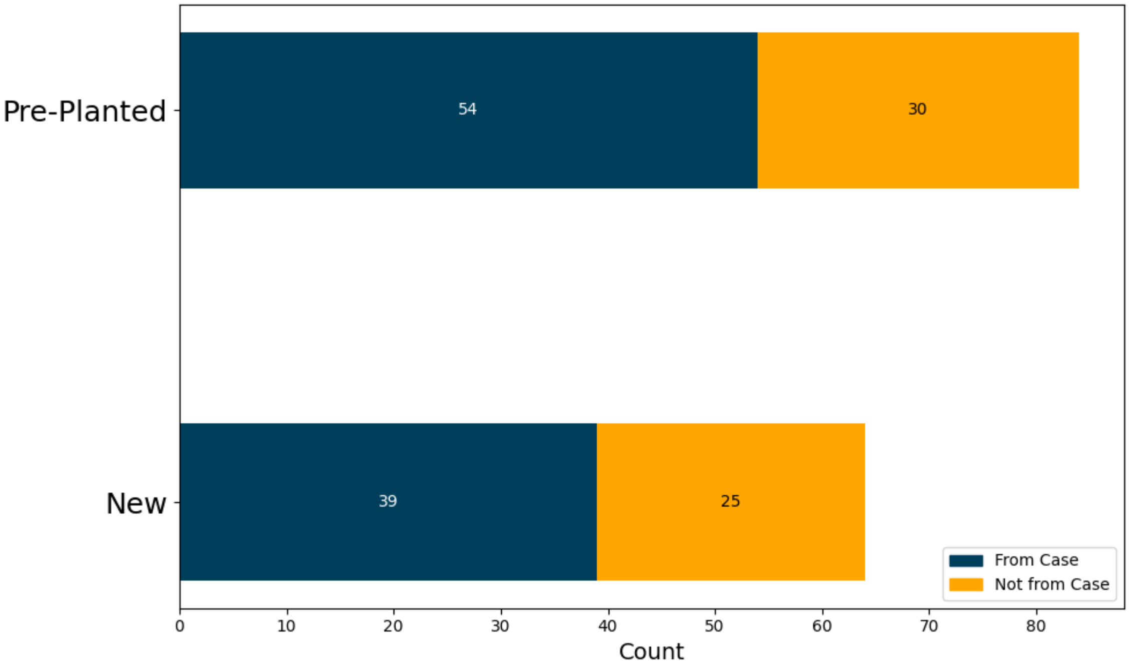


Figure 5.36: CS32 Source-Category Breakdown (Selected Sample)

would naturally extract nuanced details from the cases, given that they had just encountered them, and leverage the details to formulate their own arguments. This approach was designed to evaluate how effectively students could engage with and apply the case material in a fresh context. To evaluate students' performance in this aspect, we need to determine how frequently students cite the pre-planted factors from the cases. A strong correlation between students' responses and the case content would suggest that the cases are being effectively utilized as intended.

Figure 5.35 and fig. 5.36 display the counts of pre-planted versus new categories, further segmented by the sources labeled as "from case" and "not from case." In fig. 5.35, the distribution within the "pre-planted" category bar shows a roughly even split between sources, suggesting that students did not predominantly rely on the cases to mention these pre-planted categories. For the "new" category, the orange segment representing "not from case" predominates, indicating that most of the new categories were independently thought of by the students, rather than being influenced by the cases, which suggests limited construction of novel arguments derived from the cases themselves.

In fig. 5.36, the distribution within the "pre-planted" category predominantly features the blue "from case" segment, which is larger than the orange "not from case" segment (54 vs. 30). This indicates that in CS32, students primarily referenced the cases when discussing pre-planted categories. For the "new" categories, "from case" also constitutes the larger portion, but this predominance can be largely attributable to the "effectiveness" category. As previously analyzed in RQ 1, "effectiveness" is a new category that was frequently mentioned in conjunction with pre-planted categories such as "simple and intuitive use" and "flexibility in use." This alignment supports our earlier findings, confirming that students in CS32 were engaging with the cases as anticipated.

In summary, the analysis indicates that CS111 students did not heavily rely on the cases when identifying relevant factors, whereas CS32 students did. A plausible explanation for this pattern is the influence of the topic covered by the cases. As previously discussed in RQ 1.1, students in CS111 typically possess a higher level of background knowledge compared to those in CS32. This existing familiarity might explain why CS111 students were less dependent on the cases, whereas the reliance observed among CS32 students suggests a stronger impact of the case studies on those with less prior knowledge.

5.5 Conclusion: did contrasting cases work?

Our study, conducted across two different courses, aimed to explore how contrasting cases impact student engagement with pre-planted factors within case studies. The findings, however, did not strongly support our initial hypothesis that contrasting cases would significantly enhance the students' ability to identify and analyze these embedded factors. In CS111, students tended to bring up more new categories rather than focusing on the pre-planted ones, whereas CS32 students showed the opposite trend. Furthermore, within each course, students were randomly assigned to one of three conditions—Group A received contrasting cases, while Groups B and C received a single case. The analysis revealed no substantial differences in the performance across these groups: quality ratings were similar, and the group receiving contrasting cases did not identify more pre-planted categories than those who received a single case.

Upon reflection, we recognize that our assumptions in task design might have influenced these outcomes. The tasks were structured to encourage students to engage with the cases and then independently list factors related to the case topics. However, the connection between the cases and the task requirements may not have been as clear to the students as we assumed. This was particularly evident in responses to an interview question that asked students if they would like to add or change anything in their list of factors after discussing the cases and relevant content. Surprisingly, many students saw no need for additions or modifications, despite having discussed relevant points during the interview. For example, one student saw no additions needed even though they had discussed various content moderation methods and trade-offs. Another mentioned nothing was missing after a conversation on transparency and the collaboration between human and machine decision-making. Similarly, another student responded with ”Not really” when asked about adding to their factors, despite just having identified the trade-offs between the convenience offered by non-interactive options and the potential loss of privacy involved.

These responses indicate a potential disconnect in how students link the detailed discussion of cases to the broader analytical tasks. This insight suggests that while the case studies did foster engagement and elicited ”good” responses, the intended use of contrasting cases to deepen understanding might not have been as effective as anticipated.

Despite these challenges, the use of case studies proved promising. Across both courses, students frequently produced ”good” responses that delved into the complexities of the given factors, showcasing a depth of understanding and reasoning. The open-ended nature of the tasks also encouraged students to bring forward new and interesting categories. For less familiar topics, such as CAPTCHA, students were able to draw upon the case studies to formulate strong arguments.

In conclusion, while our study did not conclusively demonstrate the effectiveness of contrasting cases in fostering a deeper understanding of pre-planted factors, it did affirm the value of using case studies to present the nuanced aspects of design principles effectively. The quality of student responses, particularly their ability to discuss nuances and bring in new categories, highlights the potential of case studies as a teaching tool. Further research is still needed to determine the most effective structure for presenting case studies and instructions to students.

Chapter 6

Reflection and Future Work

6.1 Limitations

This section outlines the key limitations identified in our study, providing a transparent overview that can guide future research efforts and refine methodological approaches to enhance the validity and reliability of similar studies. To effectively address the validity threats in our study, we consulted the papers ”Internal, External, and Ecological Validity in Research Design, Conduct, and Evaluation” [4] and ”How Do Computing Education Researchers Talk About Threats and Limitations?” [50] for definitions and comprehensive checklists to consider.

External Validity Threats Andrade [4] defines external validity as the extent to which study findings can be generalized to other contexts. Our study was limited to two courses within Brown University, which presents a ”limited setting” challenge as described in [50]. Additionally, the volunteer basis of our interview recruitment process could introduce volunteer bias, potentially making the interviewees unrepresentative of the broader student population. This concern was underscored during interviews where some students explicitly expressed their enthusiasm for participating in the study and their general interest in SRC. These expressions of interest may not accurately reflect the attitudes of the broader student body, potentially skewing the representativeness of our findings.

Internal Validity Threats According to Andrade [4], internal validity assesses whether the study design, conduct, and analysis can address the research questions without introducing bias. The semi-structured nature of our interviews posed a challenge in avoiding confirmation bias, as there was a temptation to guide students towards desired outcomes. Furthermore, the coders’ familiarity with the study's design and objectives might have influenced their interpretation of the data, potentially leading to biased coding results, where responses could be overfitted to meet expected findings. Typically, avoiding these issues would involve having an independent third party conduct the interviews and conduct the coding. However, such an investment is usually made once there is a substantial evidence base from exploratory studies. Given that this dissertation represents the initial exploration of contrasting cases for SRC, these limitations are not unexpected at this stage of research.

Ecological Validity Ecological validity, as defined by Andrade [4] concerns whether study findings can be generalized to real-life settings. Given our focus on designing SRC assignments and conducting both studies in this context, we believe the study adequately addresses ecological validity, ensuring the findings are applicable to actual educational settings.

Reliability Due to time constraint, we chose not to pursue inter-coder reliability directly but instead opted for multiple coding rounds with collaborative discussions and refinements of the codebook and rubric. Another potential threat to reliability could arise from inconsistent application of the codebook and quality rubric. There is a risk that coders' biases could influence their judgments over time; for example, if coders initially encounter well-structured responses to a particular category, they might subconsciously deem that category as inherently relevant. This predisposition could later lead them to classify subsequent mentions of the same category as relevant, even if those responses failed to make a sufficient argument on how it connects to the central topic.

6.2 Reflections

6.2.1 Factors Influencing Study Outcomes

Our study yielded different results between CS111 and CS32, mostly in terms of the categories referenced and the reliance on the provided cases. CS111 students tended to bring up more new categories than those pre-planted in the cases, relying heavily on their previous knowledge and personal experiences. Conversely, CS32 students tended to reference the cases more, and brought up more pre-planted categories. Several factors may account for these differences:

Topic Relevance The topics of the studies themselves—content moderation for CS111 and CAPTCHA for CS32—differ in their immediacy and relevance to students' everyday experiences.

of students' background knowledge, which included quantitative data from selected source checkboxes and deeper insights obtained from qualitative interviews. Indeed, content moderation is a highly relevant and frequently discussed topic in the context of social media and misinformation, particularly during the COVID-19 pandemic. In contrast, CAPTCHA is less prominently covered in news and media and generally less considered by users. The choice of topic can greatly influence how students engage with the cases.

Student Populations CS111 is an introductory class, whereas CS32 is an intermediate level class. It is possible that students in the intermediate class, with more experience in handling SRC topics, were more attuned to the details in the case studies and could thus engage more deeply with the content. Future iterations of the study could explore how different levels of students respond to contrasting cases, to determine if contrasting cases as an intervention is more effective for certain student population.

Timing within the Semester The timing of the study also varied: CS111 participated at the semester's start, without any relevant coursework to connect with the study, whereas CS32 had just completed a project related to accessibility and screen readers. This prior engagement may have prepared CS32 students to engage more thoroughly with the cases, particularly regarding accessibility and universal design. This observation suggests that the timing of such studies in relation to other course content could significantly influence student outcomes. Future research should explore the most effective moments within a curriculum to introduce the contrasting cases study for maximum impact.

6.2.2 Recommendations for the Next Study

In addition to these considerations, we also gained several insights on improving study design and execution:

Topic Selection Opting for a topic that is not currently "hot" may reduce the reliance on prior knowledge, allowing students to engage more fully with the study materials rather than drawing primarily on external information. However, it is also important to choose a topic that interests students, one that they found compelling and motivating, to facilitate effective learning.

Task Design It may be beneficial to provide more specific guidelines for the highlighting task, as we discovered that students had varied interpretations of what constitutes as "interesting or notable". This variability led to inconsistencies, with some students potentially overlooking critical parts of the case due to the open-ended nature of the task. Tightening the criteria could help push students to read the cases more carefully. For the task of listing factors to consider, it may be beneficial to add more instructions to encourage students to utilize details from the cases.

Coding Process Introducing multiple coding rounds with varied data ordering and the inclusion of additional coders could help reduce anchoring bias and improve the reliability of coding outcomes. For instance, implementing an extra round of coding where the order of data is altered, and involving a new coder to apply the codebook and quality rubric, could serve as a robust check. By comparing whether consistent codes emerge across different rounds and coders, we can more effectively mitigate the influence of anchoring bias, where coders might disproportionately rely on the initial information they encounter. This method ensures a more objective and reliable coding process.

6.3 Future Research Directions

As detailed in section 3.3, our application of contrasting cases diverges from the conventional approach of using syntax-based cases that highlight the structural differences (such as the clowns example shown in fig. 3.2). We opted instead for a narrative in textual format, which introduces a layer of abstraction. By refraining from directly presenting instances of the design principles (such as the Santa Clara Principles [1] or the 7 Principles of Universal Design [33]), we introduce an extra layer of abstraction. This approach aims to steer the focus away from mere reading comprehension. These variations of abstractions suggest that we may be exploring a distinct application of contrasting cases, which warrants further investigation, particularly in how they frame contrasting cases in terms of perceptual and discovery learning.

Another dimension of contrasting cases not employed in this study is the use of negative examples. As illustrated in fig. 6.1, Schwartz et al. [55] have used this to teach the definition of polygons, emphasizing the importance of including “not polygons”. They used this example to argue that “learning what a thing is also depends on learning what it is not”.

In our context, however, the definition of a negative example is less clear. It might be the absence of a pre-planted factor in a case, or a failure to consider a design principle. Yet, these scenarios differ from typical negative examples. What, then, would a negative example look like when teaching SRC? Does such an example exist, and could its inclusion enhance learning? These questions should guide the direction of future research.

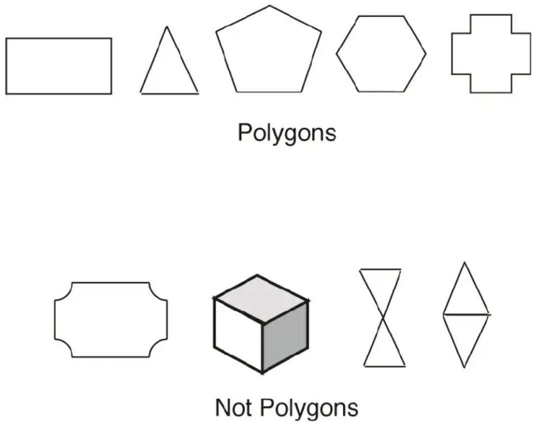


Figure 6.1: Contrasting Cases with Negative Examples[55]

6.4 Conclusion

This dissertation focused on exploring pedagogic approaches to teach socially responsible computing (SRC). Throughout this work, a series of design explorations were developed and analyzed, aimed at enhancing student engagement with SRC concepts beyond superficial interactions. A notable contribution of this study is the experimentation with contrasting cases in the setting of SRC, grounded in the established effectiveness of contrasting cases for helping students identify key features and discern subtle differences. While our study results indicate that contrasting cases are a promising method for presenting and helping students identify and analyze SRC topics, they also highlight the need for further investigation, as the evidence from the current study was not conclusively strong.

In the broader context, the objective of this dissertation is to influence the design of teaching SRC by broadening the range of pedagogic approaches under consideration, thereby better preparing students to be more mindful and skillful in recognizing and analyzing issues related to socially responsible computing.

Appendix A

Summer 2023 Codebook

This codebook outlines the coding scheme utilized for the qualitative analysis of the first iteration of the contrasting cases study in summer 2023. Codes are organized hierarchically, beginning with top-level codes enclosed in brackets, such as [Refer], accompanied by brief summaries denoting their respective meanings, such as (mention details in the case study). Subsequently, sub-categories and corresponding summaries are provided beneath each top-level code.

- [Refer] (mention details in the case study)
    - TikTok (here are the factors that make detecting misinformation hard)
        - Rich media format
        - Community language and slangs
        - Logical fallacies
        - Misspelling
    - Twitter (here's what I learned about how misinformation spreads on Twitter)
        - People with fewer followers are spreading misinformation
        - Emotionally engaging
        - True and false information spreads at the same rate
        - Bots were not very responsible for spreading fake news
- [Strategy] (talk about the tradeoff and priotization issues of implementing the policy)
    - What to prioritize
        - Popular posts and set a threshold to determine popularity, spreading speed, and/or impact
        - New posts
        - Categorize/label the posts, but shouldn't rely on machine learning to do auto categorization
    - How to do content moderation
        - Get a diverse group of people to check content
        - moderation based on user reporting might be unfair
        - users should have a say in the policy, not based on follower count but based on consensus
        - user authentication
        - User must provide source
        - algorithm threshold to determine the spreading speed, then human comes in
        - users report things, then an algorithm comes in to check

- [Compare] (mention the differences between two cases)
  - Type of information and content are different
    - more new information on tiktok
    - Videos on tiktok vs. text-based tweets
      - AI is more helpful with the text-based content, but may not pick up the context in a video
      - the video format on tiktok may help someone look more official than twitter
      - Less effort to propagate misinformation on Twitter, because you can simply copy paste a link, but on tiktok you need to make a video
    - Long tweets vs. entire video, which both make it hard for audiences to go through the whole post
    - More political information on twitter vs. Covid misinformation on tiktok, they cause different kinds of harm
    - Twitter allows adult content
  - Platforms' functionality
    - User interactions
      - comments and tagging work differently. On tiktok, as soon as someone turns of the comments, people can't tag others in the comments.
      - The reply/response model is different: on twitter you can retweet, but on tiktok you need to stitch videos to add on your own reaction. The original post and reply post would get different popularity
      - Tiktok is more confined in terms of interacting with audience
    - Recommendation algorithm
      - more likely to see similar stuff right next to each other on tiktok than on twitter
      - Tweets get to the followers' timeline vs. Tik Tok video recommendations are more based on algorithm
    - Spreading model
      - info spreads slower on twitter. On tiktok everyone is taking in a ton of content all the time.
      - Users can tag others in the comment on tiktok, so we need to monitor the comments as well to stop the spreading
      - Focus on the source/original vs. focus on the spreading. Tiktok you can stitch the videos together and the focus will still be on the source.
      - text retweets vs. video share

            - different goals for creators, getting likes vs. getting retweets
    - Users and Stakeholders
        - parents on tiktok and partisanship on twitter
        - harm on tiktok is worse because kids use it
        - Scientific/academic community on Twitter
    - Status quo
        - Currently there’s pre-posting moderation on tiktok, but not on twitter
        - there is more research about Twitter than TikTok because TikTok is newer
        - tiktok is more recent and rapidly changing. Legislation may not be keeping up
- [Automated tool] (details about automated tools)
- [Human] (details about human intervention)
- [Background knowledge] (students refer to their own background knowledge)
- [New ideas] (things that they didn’t know before)
- [Question] (clarification questions about the instructions I gave)
- [Apply](take the idea from one case and apply it to another case)

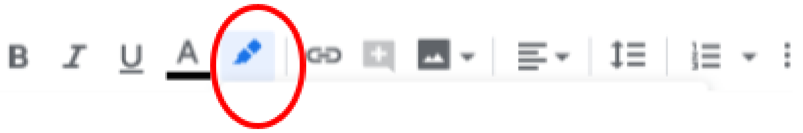
Appendix B

Assignment Handout

Task 1: Please read the case studies below, and use the highlighting tool of Google Docs to mark any parts that look interesting or notable to you.

Please set a timer of 12 minutes for this task, and read the cases again if you finish early. DO NOT proceed to task 2 until the time is up. Use any color you wish and highlight directly on the text (not with comments). If you are physically unable to mark highlights or text, reply to Yanyan (yanyan_ren@brown.edu) and we'll find another way to gather your thoughts.

Below is an image of the highlighting button. If you need more help using the highlighting tool, please refer to this tutorial or reach out to Yanyan.



If you encounter any terms that you don't know, please consult the FAQ provided below the cases on the next page. If there's a confusing term not in the FAQ, please email Yanyan to let her know.

Case 1: Alex is a content moderator at Company AAA, a US-based social media platform. The company's automated content-checking tools removed a post and, following company policy, flagged it for Alex to review. The post in question was written by a grandmother who affectionately called her grandchild with dark skin a "black diamond". Alex assumes (but can't be entirely sure) that the automated tool flagged the post as being racist, but notices that the grandmother is from a country which doesn't share the American concept of race. Alex wishes to reinstate the post, but their American boss prefers to err on the side of caution and upholds the deletion decision.	**Case 2:** Logan is on the content moderation team at Company BBB, a large social media platform. Their job is to help maintain the machine-learning tool (model) that handles content-moderation. The tool flags inappropriate posts (with a mark that is visible to all users of the platform), downranks them, and collects them in a database. Logan's team regularly adjusts the model based on the database. A user contacts customer service to find out why their posts appear with a "potentially offensive" mark. The tool's internal exploration feature reports that the posts use language that suggests potential violence to women. The user is allowed to edit the post in order to potentially have the mark removed.

FAQ

1. **Q: What does content moderation mean?**
   A: Content moderation is the process of reviewing and monitoring user-generated content on online platforms to ensure that it meets certain standards and guidelines. In other words, when a user submits content to a website, that content will undergo a screening process (known as the moderation process) to ensure that the content upholds the website's regulations and is not illegal, inappropriate, harassing, etc.

2. **Q: Is content moderation the same thing as blocking?**
   A: Not necessarily. Content moderation focuses on identifying concerning content. What to do after identification (e.g. blocking) is a separate question.

3. **Q: What does "automated tools" mean?**
   A: Use of a machine learning algorithm or AI, for example, to identify content of concern.

4. **Q: What does "downrank" mean?**
   A: The content will be demoted by the algorithm. It's not removed, but becomes less visible, like appearing on a second page of search-engine results.

Task 2: Imagine you are a software engineer designing content moderation tools for a social media startup CCC. What are some factors you need to consider?

Put your answers the table below, and feel free to leave some rows blank or add more rows. For each row, please enter the factor that needs to be considered, a sentence or two about what's interesting about this factor, and click on the corresponding box in the right column to indicate what inspired you to think of the factor.

General expectations and example response text:
Please substantiate your answers with details from the cases or other examples as you can think of. For example, if you put "bias" under the "Factor to consider" column, for the "what's interesting about this factor" column, avoid providing simple answers like "the data could be biased". Instead, provide thorough answers with details such as "Bias could come from datasets' limited geographic representation. The dataset mentioned in the case only contains data from English-speaking countries."

Submit your response: Please fill out this form when you are done with both tasks.

Factor to consider	**What's interesting about this factor?**	**Source**
		□ **Cases** (that you just read) □ **Previous knowledge** (from a class, a website, a book, an internship, etc.) □ **Personal experience** (as a user on some social media platform)
		□ **Cases** (that you just read) □ **Previous knowledge** (from a class, a website, a book, an

		internship, etc.) □ **Personal experience** (as a user on some social media platform)
		□ **Cases** (that you just read) □ **Previous knowledge** (from a class, a website, a book, an internship, etc.) □ **Personal experience** (as a user on some social media platform)
		□ **Cases** (that you just read) □ **Previous knowledge** (from a class, a website, a book, an internship, etc.) □ **Personal experience** (as a user on some social media platform)
		□ **Cases** (that you just read) □ **Previous knowledge** (from a class, a website, a book, an internship, etc.) □ **Personal experience** (as a user on some social media platform)
		□ **Cases** (that you just read) □ **Previous knowledge** (from a class, a website, a book, an internship, etc.) □ **Personal experience** (as a user

		on some social media platform)
		□ **Cases** (that you just read) □ **Previous knowledge** (from a class, a website, a book, an internship, etc.) □ **Personal experience** (as a user on some social media platform)
		□ **Cases** (that you just read) □ **Previous knowledge** (from a class, a website, a book, an internship, etc.) □ **Personal experience** (as a user on some social media platform)
		□ **Cases** (that you just read) □ **Previous knowledge** (from a class, a website, a book, an internship, etc.) □ **Personal experience** (as a user on some social media platform)
		□ **Cases** (that you just read) □ **Previous knowledge** (from a class, a website, a book, an internship, etc.) □ **Personal experience** (as a user on some social media platform)

Appendix C

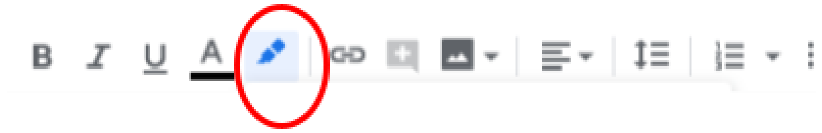
Assignment Handout Used in CS32

Task 1: Please read the case studies below, and use the highlighting tool of Google Docs to mark any parts that look interesting or notable to you.

Please note that the implementation details are intentionally unspecified, as they are not the primary focus. Please concentrate on the SRC-related aspects instead.

Please set a timer of 12 minutes for this task, and read the cases again if you finish early. DO NOT proceed to task 2 until the time is up. Use any color you wish and highlight directly on the text (not with comments). If you are physically unable to mark highlights or text, reply to Yanyan (yanyan_ren@brown.edu) and we'll find another way to gather your thoughts.

Below is an image of the highlighting button. If you need more help using the highlighting tool, please refer to this tutorial or reach out to Yanyan.

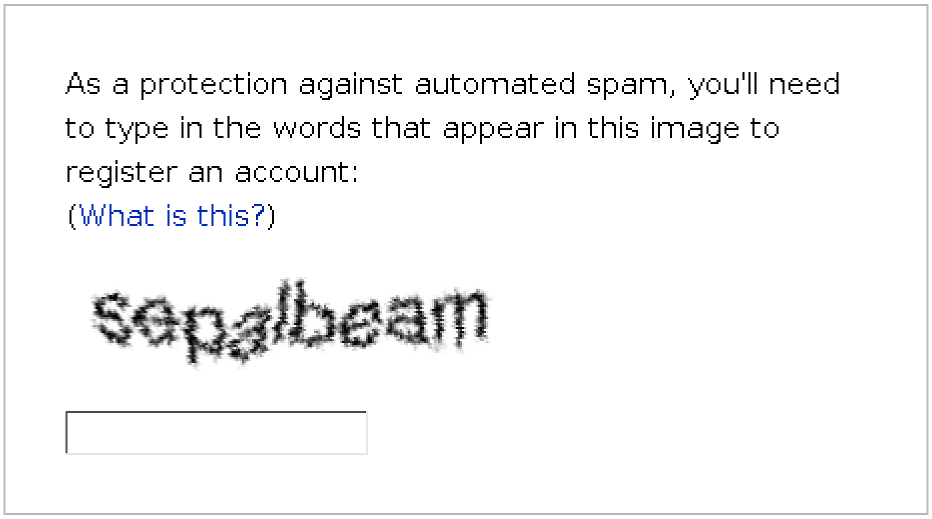


Before you get started, if you need clarification on the terms "CAPTCHA", "bot traffic" or "brute force attacks", please read this FAQ. If there's a confusing term not mentioned in the FAQ, please email Yanyan to let her know.

Case 1: Jordan is an engineer working at AAA, an e-commerce platform. As the platform grows, the team notices increasing bot traffic, and has asked Jordan to propose a CAPTCHA system that would be easy to use. Jordan's system uses simple questions (e.g. "what's 1+six"), and the problems are only available in English. The questions are fully compatible with screen readers and can be navigated using a keyboard. Users are prompted to answer the questions within two minutes, but can press a "give me two more minutes" button to get extra time. Users are allowed up to three wrong attempts. There's also an alternative, non-interactive option where the system uses machine learning to detect human users, but this option requires users to provide their email address as a unique identifier within the system.	**Case 2:** Charlie is an engineer at BBB, an online ticketing platform. The current CAPTCHA system has been effective preventing bots from bulk buying tickets, but isn't particularly user-friendly. Charlie proposes using a third-party tool in which users who are already logged into their Google account in the same browser will be verified as human without filling in the CAPTCHA. Under the hood, cookies track users' online activities and share that data with Google. Those not logged into Google, will be asked to type out a single word shown in an image. An audio alternative is also available. Both the images and audios are from database built from old English books. The system also allows users to request a new CAPTCHA challenge if they find the current one too difficult. To prevent brute force attacks by bots, the proposed tool requires users to finish the task within a minute.

FAQ

1. **Q: What does CAPTCHA mean?**
   A: It is a contrived acronym for "Completely Automated Public Turing test to tell Computers and Humans Apart". It is a type of challenge-response test used in computing to determine whether the user is human in order to deter bot attacks and spam. (source: Wikipedia) Here's an example of a CAPTCHA:



As a protection against automated spam, you'll need to type in the words that appear in this image to register an account:
(What is this?)

2. **Q: What is bot traffic?**
   A: Non-human internet activity generated by automated software, such as scraping, spamming, or viewing the website. Bot traffic could be malicious, potentially leading to a DDoS (Distributed Denial of Service) attack, where the bot traffic overwhelms the website, causing it to slow down or crash.

3. **Q: What does "brute force attacks by bots" mean?**
   A: The bots repeatedly try different combinations to solve the challenge.

Task 2: Imagine you are a software engineer designing the user-facing part of a CAPTCHA system for an online voting platform CCC. What are some factors you need to consider?

Put your answers in the table below, and feel free to leave some rows blank or add more rows. For each row, please enter the factor that needs to be considered, a sentence or two about what's interesting about this factor, and click on the corresponding box in the right column to indicate how you thought of the factor.

General expectations and example response text:
Please substantiate your answers with details from the cases or other examples you can think of. For example, if you put "bias" under the "Factor to consider" column, for the "what's interesting about this factor" column, avoid providing simple answers like "the data could be biased". Instead, provide **thorough answers with details** such as "Bias could come from datasets' limited geographic representation. The dataset mentioned in the case only contains data from English-speaking countries."

Submit your response: Please fill out this form when you are done with both tasks.

Factor to consider	**What's interesting about this factor?**	**Source**
		□ **Cases** (that you just read) □ **Previous knowledge** (from a class, a website, a book, an internship, etc.) □ **Personal experience** (as a user)
		□ **Cases** (that you just read)

		□ **Previous knowledge** (from a class, a website, a book, an internship, etc.) □ **Personal experience** (as a user)
		□ **Cases** (that you just read) □ **Previous knowledge** (from a class, a website, a book, an internship, etc.) □ **Personal experience** (as a user)
		□ **Cases** (that you just read) □ **Previous knowledge** (from a class, a website, a book, an internship, etc.) □ **Personal experience** (as a user)
		□ **Cases** (that you just read) □ **Previous knowledge** (from a class, a website, a book, an internship, etc.) □ **Personal experience** (as a user)
		□ **Cases** (that you just read) □ **Previous knowledge** (from a class, a website, a book, an internship, etc.)

		□ **Personal experience** (as a user)
		□ **Cases** (that you just read) □ **Previous knowledge** (from a class, a website, a book, an internship, etc.) □ **Personal experience** (as a user)
		□ **Cases** (that you just read) □ **Previous knowledge** (from a class, a website, a book, an internship, etc.) □ **Personal experience** (as a user)
		□ **Cases** (that you just read) □ **Previous knowledge** (from a class, a website, a book, an internship, etc.) □ **Personal experience** (as a user)
		□ **Cases** (that you just read) □ **Previous knowledge** (from a class, a website, a book, an internship, etc.) □ **Personal experience** (as a user)

Appendix D

Data Usage FAQ

1. **What is the purpose of this study?**
    a. We are studying how to design exercises that are effective in helping students learning socially-responsible computing (how to avoid harm and create benefits when computing interacts with society).
    b. You are being asked to participate in the "written part" of the study because you are taking a course with socially-responsible computing (SRC) content.
    c. You might be invited to participate in a subsequent "interview part" if your written response suggests that we might get additional insight from a conversation about how you worked on the written portion.

2. **Are there any risks associated with the study?**
    a. There is minimal risk involved.
    b. The personal information gathered in the written portion consists of your answers to questions about whether you had prior knowledge or personal experience with aspects of the case studies within the written study. We do not ask for details of these in the written portion.
    c. While completing the assignment is mandatory, your consent to use coursework for research analysis is optional. It does not affect how your professor uses your work for grading and course completion.
    d. You do not have to be in this study if you do not want to be. Even if you decide to be in this study, you can change your mind and stop at any time. Participation or the choice not to participate will not impact your grade or participation in class.

3. **Are there any benefits associated with the study?**
    a. You may not directly benefit from being in this research study.
    b. The main benefit to you is any satisfaction you might get from contributing to a research study that aims to improve education for future groups of students.

4. **What happens to my data?**

a. Your written submission will be uploaded to a folder within Brown's Google Drive. Access to that folder will be limited to authorized research team members (all of whom are faculty or students at Brown).

b. Your work will be associated with a unique anonymous participant ID, ensuring the anonymity of your data during our analysis. Your email will be securely stored separately, to be used exclusively for study-related communication and scheduling purposes. We will only look at the emails associated with specific IDs in order to potentially invite you to the interview portion or to remove your data should you later decide to withdraw from the study.

c. In any sort of report or paper that we make public, only anonymized direct quotes and aggregated data will be included, and no identifying information will be disclosed.

5. **Who should I contact if I have questions about the study?**

   If you have any questions about your participation in this study or wish to withdraw later, you can email Yanyan Ren at yanyan_ren@brown.edu. You can also contact Yanyan's faculty advisor Kathi Fisler at kathryn_fisler@brown.edu.

Appendix E

Interview Script Used for CS111

Highlighted notes serve as internal reminders, guiding our analysis of transcripts rather than being direct inquiries posed to participants.

General Expectation: (read out loud to participants at the beginning)

This is a talk-aloud study. It's important that you verbalize your thoughts in real time as you work through the questions. And it's important to express all your thoughts about the activity, even if you find them irrelevant or unimportant.

Before we begin, I want to emphasize that your comfort and autonomy during this interview are our top priority. If there's any prompt you prefer not to answer, simply let me know that you'd like to skip it, and we'll move on to the next question without further discussion on the skipped item. Do you have any questions before we start?

I am turning on the screen recording and autotranscript now. As I explained in the email, the screen recording is intended to capture the researcher's screen share, and the auto transcription will be used to facilitate data analysis. You do not need to turn on your camera or share your screen.

Task 1

- Let's take a look at your response to Task 1. Could you walk me through why you chose to highlight these parts? What's interesting about them?

Did the participant notice things that we didn't expect? Did they interpret some details in an interesting way?

- (optional) I noticed that you used different colors here. What do the colors represent?

Do the colors mean that they notice the contrast between cases?

- Were there other parts that caught your attention but you did not highlight? Why? What other comments or questions do you have about the cases?

For students who got two cases

- What similarities do you notice between the two cases?
- What differences do you notice between the two cases?

What did students notice? Are they seeing things the way we expected them to? Are they bringing up other similarities or differences that we didn't think of?

Task 2

- Let's take a look at your responses to Task 2.
- Can we go through the factors one by one? Could you give me some examples for each of them, and say more about why the factor is interesting or important?

I probably said this to every student to get them to talk more about their reasoning and connect it to content moderation. How did their response here differ from their written response? What additional details did they mention?

- For the customer support factor, you mentioned that user feedback is important. What should the feedback look like? Can you give me an example?
- I thought it's interesting that you put quotes around "right solution" and "perfect state". Could you expand on that?
- For the company's policy factor, I noticed that you checked "cases" as a source. How does this show up in the cases?
- Is being sensitive and overfitting a good thing here?
- Could you talk more about the role of automated tools here?

The questions I asked in task 2 are not the same for everyone. But basically we want to look for

1. keywords that students brought up and their understanding of it (e.g. maybe someone has a very deep understanding on the consequences of overfitting).
2. clarifications on their written responses (e.g. they wrote a very breif answer, now they are actually covering more categories or they meant something else)
3. Any mentions about the cases

- I noticed that you checked a lot of boxes of personal experience and previous knowledge. Could you talk more about them? How much did you know about content moderation before doing this exercise?

Get a sense of the level and kinds of background knowledge students have before doing the exercise.

For students who got one case

- Show them **the second case**
- What's interesting about this case?
- Now we look at the two cases side by side,
  - What similarities do you notice between the two cases?
  - What differences do you notice between the two cases?
- Now if we look at your responses to task 2, would you change anything?

Basically we want to see if students identify more factors or come up with better reasonings after seeing the second case.

(optional) Explicitly giving them the factors

- What do you think of the use of automated tools in the cases? Was that an interesting factor when considering content moderation practices? Why or why not?

- What did you notice about the cultural context in the cases? How does it impact the company's content moderation practices?
- What action would you take for content that violates the rules? What do you think of the action taken in the given case?
- Have you heard of the term "transparency"? How does that apply to the case here? Would user awareness be a related factor here?
- How did the content moderator interact with the tool? Did they get any explanation on why the tool made the decision this way?

Does explicitly giving them the factors help students articulate its connection to content moderation?

Summary

- Is there any additional information or insights you would like to add regarding the topic of content moderation?
- After spending 30 minutes talking about the case studies and content moderation today, if we take another look at the list of factors and reasons you have here, would you like to add or change anything?

Appendix F

Interview Script Used for CS32

Highlighted notes serve as internal reminders, guiding our analysis of transcripts rather than being direct inquiries posed to participants.

General Expectation: (read out loud to participants at the beginning)

This is a talk-aloud study. It’s important that you verbalize your thoughts in real time as you work through the questions. And it’s important to express all your thoughts about the activity, even if you find them irrelevant or unimportant.

Before we begin, I want to emphasize that your comfort and autonomy during this interview are our top priority. If there’s any prompt you prefer not to answer, simply let me know that you’d like to skip it, and we’ll move on to the next question without further discussion on the skipped item. Do you have any questions before we start?

I am turning on the screen recording and autotranscript now. As I explained in the email, the screen recording is intended to capture the researcher’s screen share, and the auto transcription will be used to facilitate data analysis. You do not need to turn on your camera or share your screen.

Task 1

- Let’s take a look at your response to Task 1. Could you walk me through why you chose to highlight these parts? What’s interesting about them?

Did the participant notice things that we didn't expect? Did they interpret some details in an interesting way?

- (optional) I noticed that you used different colors here. What do the colors represent?

Do the colors mean that they notice the contrast between cases?

- Were there other parts that caught your attention but you did not highlight? Why? What other comments or questions do you have about the cases?

For students who got two cases

- What similarities do you notice between the two cases?
- What differences do you notice between the two cases?

What did students notice? Are they seeing things the way we expected them to? Are they bringing up other similarities or differences that we didn’t think of?

Task 2

- Let's take a look at your responses to Task 2.
- Can we go through the factors one by one? Could you give me some examples for each of them, and say more about why the factor is interesting or important?

I probably said this to every student to get them to talk more about their reasoning and connect it to content moderation. How did their response here differ from their written response? What additional details did they mention?

- Do you see the bias factor showing up in the case? Where do you see it?
- Could you expand on the accessibility factor? What other examples can you think of? Are they handled well in the case?
- For the time spent factor, I thought it's interesting that you used word "balance". Could you expand on that? Are you seeing any tradeoffs here?

The questions I asked in task 2 are not the same for everyone. But basically we want to look for

1. keywords that students brought up and their understanding of it (e.g. maybe someone has a very deep understanding on the consequences of overfitting).
2. clarifications on their written responses (e.g. they wrote a very breif answer, now they are actually covering more categories or they meant something else)
3. Any mentions about the cases

- I noticed that you checked a lot of boxes of personal experience and previous knowledge. Could you talk more about them? How much did you know about captcha before doing this exercise?

Get a sense of the level and kinds of background knowledge students have before doing the exercise.

For students who got one case

- Show them **the second case**
- What's interesting about this case?
- Now we look at the two cases side by side,
  - What similarities do you notice between the two cases?
  - What differences do you notice between the two cases?
- Now if we look at your responses to task 2, would you change anything?

Basically we want to see if students identify more factors or come up with better reasonings after seeing the second case.

(optional) Explicitly giving them the factors

- What is your understanding of universal design? How does this connect to the case studies you just read?
- A group of researchers from North Carolina State University have developed a set of principles for universal design. I'll read some of the principles from there. For each

principle, how does it relate to the cases you just read? Do you see it the factor you listed? Anything you’d like to add or change to your list of factors?

- Equitable use: Provisions for privacy, security, and safety should be equally available to all users.
- Flexibility in use: Provide adaptability to the user's pace.
- Simple and intuitive use: Accommodate a wide range of literacy and language skills.
- Perceptible information: Use different modes (pictorial, verbal, tactile) for the redundant presentation of essential information.
- Tolerance for error: Provide fail-safe features.

Does explicitly giving them the factors help students articulate its connection to content moderation?

Summary

- Is there any additional information or insights you would like to add regarding the topic of captcha?
- After spending 30 minutes talking about the case studies today, if we take another look at the list of factors and reasons you have here, would you like to add or change anything?

Appendix G

Codebook for CS111

Pre-Planted Categories

- **Culture context**
    - [lng] language differences. The same word can have different meanings in different languages.
    - [mng] difference in meanings of words/phrases. A single word can have multiple meanings.
    - [ccx] anything about cultural context, cultural differences, groupings of languages without focusing on the linguistic side
- **Automated tool**
    - [lim] limitation of automated tools
    - [qdt] quality of data, biased caused by bad quality of data, reliability of data
    - [mtn] maintenance/updating of models
    - [cau] being cautious, being okay with false positive, want to have a sensitive model. Or on the flip side, argue about the harm of having too many false positives or being overly sensitive.
    - [pro] programmer bias
    - [algo] algorithmic bias
    - [acc] accuracy
    - [spe] speed of the model
    - [res] resources going into the model (money, time, training cycle, etc.)
    - [auto] use of automated tools, balancing human vs. automated tool, says something about automated tools (AI/ML/model/etc.)
- **User awareness**
    - [uawa] user awareness of action taken against the post that violates content moderation policies
    - [corr] correction, user's ability to edit their post after it's been moderated
- **Downranking vs. removal**
    - [dvr] downranking vs. removal, or anything that talks about the action taken to moderate content

- **Explanation to moderators**
    - [etm] whether and how the automated tool provides an explanation to moderators. Or mentions transparency, and talks about how the content moderation policies are explained to users.

New Categories

- **Language**
    - [evl] evolution of language, how the same word's meaning can change over time
    - [typ] typos, misspelled words (unintentional)
    - [eva] evasion of content moderation policies, using tactics like coded words, and intentional misspell
    - [ctx] context needed to understand the true meaning of the words. Searching and blocking keywords is not enough. Need to take into account emotion, humor, sarcasm, etc.
- **Kinds of content moderated**
    - [mis] misinformation, fake news, consequences/damage of misinformation
    - [cont] content, giving examples of what kind of content needs to be moderated
- **User information**
    - [age] age of the user, parental controls, consideration for kids or teens
    - [ux] user experience, user reaction, how content moderation impacts these
    - [feed] user feedback, incorporating user feedback into content moderation practices
    - [user] user's info, background, history of offenses, previous post, and using these information to do content moderation
- **Company and platform**
    - [cpl] company policy, company values
    - [env] environment and community of the social media platform
    - [rep] company's reputation

- **Human moderators**
  - [div] diversity of the human moderator team
  - [har] harm done to human moderators for reviewing bad content
- **Ungrouped**
  - [crct] whether a correct answer or ground truth exists, difficult to get the ‘right answer’
  - [fsp] freedom of speech, issues of censorship
  - [law] legal issues, regulations, country or state level policies
  - [priv] privacy, how user privacy is protected or violated
  - [ind] individual standards of what content should be moderated, what is considered “offensive”, “inappropriate”, or “immoral” (this is not linked to culture because it’s based on individuals)
  - [intr] interactions with the content moderated, what responses/comments the content provoke, how large the size of audience is for a given post
  - [intn] user’s intention (e.g. for educational purposes, malicious, political, etc.)
  - [bot] ways to deal with content generated by bots
  - [eth] whether the content is ethical, whether the moderation tool or human moderator is making ethical decisions.

Appendix H

Codebook for CS32

Pre-Planted Categories

- [equi] equitable use, provisions for privacy, security, and safety should be equally available to all users.
- [flex] flexibility in use, provide adaptability to the user's pace.
- [sim] simple and intuitive, accommodate a wide range of literacy and language skills.
    - [lan] Can’t use English as the only language
    - [det] deterrence from usage/user experience, not creating frustration or confusion
    - [bar] barrier to platform
    - [ccx] context (cultural, otherwise…); not everything is universally understood
    - [ins] instructions to user
    - [age] age considerations
- [per] perceptible information, use different modes (pictorial, verbal, tactile) for redundant presentation of essential information.
- [tol] tolerance for error, provide fail-safe features.

New Categories

Effectiveness

- [eff] effectiveness of CAPTCHA system. The system should prevent bots and recognize human users.
- [mtn] maintenance of system
- [ml] concerns about ML + AI specifically
- [scal] scalability of the system
- [inter] interpretability of how the system makes the decision

Technology

- [dev] device compatible, should work on all phones, computers, tablets, etc.
- [email] Google can’t be the only way to sign up, need to add other alternatives
- [brws] browser, similar to the email tag above, ad-blockers

Privacy

- [pri] privacy in general, or talking about consequences for not protecting data properly
- [trans] transparency, informing the user, getting consent

Voting

- [vote] brought up concerns/requirements that are related to the voting platform, but not related to CAPTCHA

Ungroup

- [law] legal considerations
- [time] time taken to complete captcha (separate from flex)
- [pres] pressure as a result of imposed time limits
- [cont] captcha content

Unrelated

- [unrelated] not related to CAPTCHA

Appendix I

Quality Rubric

Good:

Satisfies ONE of the “good reasoning criteria”:

- Use of an example generally is a sign of higher-quality reasoning
- If the student cites the case, they should expand on what the quote implies
- If a student raises a very broad question without providing more insights on how to solve it or why it’s a hard problem to solve, then it’s a “Getting there”. But if the student brings up several possibilities using the form of questions, then it’s a “Good”
- Discusses tensions in the factor/why its difficult/limitations/tradeoff, so the factor is not just important to consider, but there’s some complexity
- Doesn’t need to explain how to incorporate the factor in their own content moderation tool, as long as the reasoning explains why it’s important *and why it’s hard to consider.*
- Talks about the consequences of not taking the factor into account/the stakes of the factor/its importance

AND satisfies BOTH of the hard requirements:

- Makes a good argument for the tagged category. If it’s just mentioning a keyword, then don’t put the tag.
- Talks about how the factor relates to content moderation.

Getting there:

- Doesn’t satisfy any of the ‘good reasoning’ criteria listed above, or missing a hard requirement of ‘Good’
- Use of ‘buzzwords’/a component that is very related to content moderation, but doesn’t provide much explanation
- Only expands on the factor, but doesn’t tie it back to content moderation (almost missing a sentence like “therefore, this factor needs to be considered when designing a content moderation tool, to make it …)

Weak:

- Focuses on a component that’s not obviously a part of content moderation tool. And doesn’t provide any explanation on how it connects to content moderation.
- The writing lacks coherence, making the reader confused.
- Weak responses are more generic than getting there responses.

Examples:

All following responses are labeled as [ccx], cultural context, but the quality varies:

- **Good:** Cultural Bias - What is viewed as humor (or more generally normal behavior/language) to one cultural group could come across as offensive to others, so it is important for engineers building content moderation tools to account for these differences. In order to do this, geographical considerations definitely need to be made when creating the moderation tool (e.g reweighting certain trigger words according to how people in that area or people of a certain culture behave).
- **Getting there:** Culture - How is violent language determined based on culture? Consideration about how language and culture might translate content differently.
- **Weak:** Language barriers - People could have difficulty crossing a language barrier and properly understanding something.

Bibliography

[1] The santa clara principles on transparency and accountability in content moderation. `https://santaclaraprinciples.org/`.

[2] Diana Acosta-Navas. Big data systems (cs265) - 2019 spring, 2019. URL: `https://embeddedethics.seas.harvard.edu/classes/cs-265-2019-spring`.

[3] Louis Alfieri, Timothy J Nokes-Malach, and Christian D Schunn. Learning through case comparisons: A meta-analytic review. *Educational Psychologist*, 48(2):87–113, 2013.

[4] Chittaranjan Andrade. Internal, external, and ecological validity in research design, conduct, and evaluation. *Indian journal of psychological medicine*, 40(5):498–499, 2018.

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, May 23, 2016.

[6] Jo Bates, David Cameron, Alessandro Checco, Paul Clough, Frank Hopfgartner, Suvodeep Mazumdar, Laura Sbaffi, Peter Stordy, and Antonio de la Vega de León. Integrating fate/critical data studies into data science curricula: Where are we going and how do we get there? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 425–435, New York, NY, USA, 2020. Association for Computing Machinery. `doi:10.1145/3351095.3372832`.

[7] Tom L Beauchamp and James F Childress. *Principles of biomedical ethics 5th edn.* Oxford University Press, 2001.

[8] Volkert Beekman, HCM de Bakker, Heike Baranzke, Oyvind Baune, MK Deblonde, Ellen-Marie Forsberg, RPM de Graaff, Hans-Werner Ingensiep, Jesper Lassen, Ben Mepham, et al.

Ethical bio-technology assessment tools for agriculture and food production. Technical report, 01 2006.

[9] Elizabeth Bonawitz, Patrick Shafto, Hyowon Gweon, Noah D Goodman, Elizabeth Spelke, and Laura Schulz. The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, 120(3):322–330, 2011.

[10] Tristan G. Brown, Alexander Statman, and Celine Sui. Public Debate on Facial Recognition Technologies in China. *MIT Case Studies in Social and Ethical Responsibilities of Computing*, (Summer 2021), aug 10 2021. https://mit-serc.pubpub.org/pub/public-debate-on-facial-recognition-technologies-in-china.

[11] Brown University. Socially responsible computing at brown. `https://responsible.cs.brown.edu/`.

[12] Bureau of Internet Accessibility. How to make captcha accessible to everyone, 2021. URL: `https://www.boia.org/blog/how-to-make-captcha-accessible-to-everyone`.

[13] Gilbert Cockton. Value-centred HCI. In *Proceedings of the third Nordic conference on Human-computer interaction (NordiCHI '04)*, pages 149–160, New York, NY, USA, 2004. Association for Computing Machinery. `doi:10.1145/1028014.1028038`.

[14] Lena Cohen, Heila Precel, Harold Triedman, and Kathi Fisler. A new model for weaving responsible computing into courses across the CS curriculum. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, SIGCSE '21, page 858–864, New York, NY, USA, 2021. Association for Computing Machinery. `doi:10.1145/3408877.3432456`.

[15] danah boyd. Be careful what you code for. `https://points.datasociety.net/be-careful-what-you-code-for-c8e9f3f6f55e`, June 2016.

[16] Melissa Dark and Jelena Mirkovic. Evaluation theory and practice applied to cybersecurity education. *IEEE Security & Privacy*, 13(2):75–80, 2015. `doi:10.1109/MSP.2015.27`.

[17] Nicholas Diakopoulos, Sorelle Friedler, Marcelo Arenas, Solon Barocas, Michael Hay, Bill Howe, HV Jagadish, Kris Unsworth, Arnaud Sahuguet, Suresh Venkatasubramanian, Christo Wilson, Cong Yu, and Bendert Zevenbergen. Principles for accountable algorithms and a social impact statement for algorithms. `https://www.fatml.org/resources/principles-for-accountable-algorithms`, 2016.

[18] L. Duboc et al. Do we really know what we are building? raising awareness of potential sustainability effects of software systems in requirements engineering. In *Proceedings of the IEEE 27th International Requirements Engineering Conference (RE)*, pages 6–16, 2019. `doi:10.1109/RE.2019.00013`.

[19] Jerry S Fisher and Gabriel A Radvansky. Patterns of forgetting. *Journal of Memory and Language*, 102:130–141, 2018.

[20] James J Gibson and Eleanor J Gibson. Perceptual learning: Differentiation or enrichment? *Psychological review*, 62(1):32, 1955.

[21] Seth T. Hamman and Kenneth M. Hopkinson. Teaching adversarial thinking for cybersecurity. *Journal of The Colloquium for Information System Security Education*, September 2016.

[22] Harvard University. Embedded ethics. `https://embeddedethics.seas.harvard.edu/`.

[23] Shawn Hernan, Scott Lambert, Tomasz Ostwald, and Adam Shostack. Uncover security design flaws using the STRIDE approach. *MSDN Magazine*, November 2006.

[24] Amanda Holpuch and April Rubin. Remote scan of student's room before test violated his privacy, judge rules, Aug 2022. URL: `https://www.nytimes.com/2022/08/25/us/remote-testing-student-home-scan-privacy.html?mc_cid=7a0ac4a993&amp;mc_eid=6883e693fb`.

[25] Michael A. Jackson. *Principles of Program Design.* Academic Press, 1975.

[26] Thomas C. Jepsen. Just what is an ontology, anyway? *IT Professional*, 11(5):22–27, 2009. `doi:10.1109/MITP.2009.105`.

[27] David I Kaiser and Julie A Shah. MIT case studies in social and ethical responsibilities of computing. `https://mit-serc.pubpub.org/`.

[28] Paul Karoff. Harvard works to embed ethics in computer science curriculum. *The Harvard Gazette*, January 2019. URL: `https://news.harvard.edu/gazette/story/2019/01/harvard-works-to-embed-ethics-in-computer-science-curriculum/`.

[29] Magdalene Lampert. When the problem is not the question and the solution is not the answer: Mathematical knowing and teaching. *American educational research journal*, 27(1):29–63, 1990.

[30] Sage S. Lattman. What's going on with embedded ethics? *The Harvard Crimson*, April 2023. URL: `https://www.thecrimson.com/article/2023/4/20/embedded-ethics-CS/`.

[31] Yu Lin, Marcelline R Harris, Frank J Manion, Elizabeth Eisenhauer, Bin Zhao, Wei Shi, Alla Karnovsky, and Yongqun He. Development of a bfo-based informed consent ontology (ico). *Bioinformatics*, 1327:84–86, 2014.

[32] Xiaodong Lin-Siegler, David Shaenfield, and Anastasia D Elder. Contrasting case instruction can improve self-assessment of writing. *Educational Technology Research and Development*, 63:517–537, 2015.

[33] Ronald Mace. The 7 principles of universal design, 1997. URL: `https://universaldesign.ie/about-universal-design/the-7-principles`.

[34] C. Dianne Martin, Chuck Huff, Donald Gotterbarn, and Keith Miller. A framework for implementing and teaching the social and ethical impact of computing. *Education and Information Technologies*, 1(2):101–122, June 1996. `doi:10.1007/BF00168276`.

[35] Patrick McKenzie. Falsehoods programmers believe about names, 2010. `https://www.kalzumeus.com/2010/06/17/falsehoods-programmers-believe-about-names/`, last accessed 2021-03-25.

[36] National Academies of Sciences, Engineering, and Medicine. Fostering responsible computing research: Foundations and practices. 2022. `doi:10.17226/26507`.

[37] OECD privacy guidelines. `https://www.oecd.org/digital/ieconomy/privacy-guidelines.htm`.

[38] Harshvardhan J Pandit, Christophe Debruyne, Declan O'Sullivan, and Dave Lewis. Gconsent-a consent ontology based on the gdpr. In *The Semantic Web: 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2–6, 2019, Proceedings 16*, pages 270–282. Springer, 2019.

[39] B. Penzenstadler, A. Raturi, D. Richardson, and B. Tomlinson. Safety, security, now sustainability: The nonfunctional requirement for the 21st century. 31(3):40–47, 2014. `doi:10.1109/MS.2014.22`.

[40] David N Perkins, Gavriel Salomon, et al. Transfer of learning. *International encyclopedia of education*, 2:6452–6457, 1992.

[41] William G Perry Jr. *Forms of Intellectual and Ethical Development in the College Years: A Scheme. Jossey-Bass Higher and Adult Education Series.* ERIC, 1999.

[42] David Pierson and Paresh Dave. Businesses accusing yelp of extortion lose another round in court, Sep 2014. URL: https://www.latimes.com/business/la-fi-yelp-ratings-20140905-story.html.

[43] Jesse Polhemus. Putting socially responsible computing at the heart of our undergraduate experience, Dec 2020. URL: https://cs.brown.edu/news/2020/12/15/putting-socially-responsible-computing-heart-our-undergraduate-experience/.

[44] Diandra Prioleau, Brianna Richardson, Emma Drobina, Rua Williams, Joshua Martin, and Juan E. Gilbert. How students in computing-related majors distinguish social implications of technology. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, page 1013–1019, New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3408877.3432360.

[45] Inioluwa Deborah Raji, Morgan Klaus Scheuerman, and Razvan Amironesei. You can't sit with us: Exclusionary pedagogy in ai ethics education. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 515–525, New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3442188.3445914.

[46] Yanyan Ren and Kathi Fisler. A social threat modeling framework to structure teaching about responsible computing. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, pages 402–408, 2023.

[47] Tony Rogers. Falsehoods programmers believe about names – with examples, 2018. https://shinesolutions.com/2018/01/08/falsehoods-programmers-believe-about-names-with-examples/, last accessed 2021-03-25.

[48] Shima Salehi, Martin Keil, Eric Kuo, and Carl E Wieman. How to structure an unstructured activity: Generating physics rules from simulation or contrasting cases. In *2015 Physics Education Research Conference Proceedings*, volume 291. American Association of Physics Teachers, 2015.

[49] Jeffrey Saltz, Michael Skirpan, Casey Fiesler, Micha Gorelick, Tom Yeh, Robert Heckman, Neil Dewar, and Nathan Beard. Integrating ethics within machine learning courses. *ACM Trans. Comput. Educ.*, 19(4), aug 2019. doi:10.1145/3341164.

[50] Kate Sanders, Jan Vahrenhold, and Robert McCartney. How do computing education researchers talk about threats and limitations? In *Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 1*, pages 381–396, 2023.

[51] Fred B. Schneider. Cybersecurity education in universities. *IEEE Security & Privacy*, 11(4):3–4, 2013. `doi:10.1109/MSP.2013.84`.

[52] Katharine Schwab. Google's new recaptcha has a dark side. *Fast Company*, June 2019. URL: `https://www.fastcompany.com/90369697/googles-new-recaptcha-has-a-dark-side`.

[53] Daniel Schwartz. Learning to perceive, telling too soon. Online video, Dec 2015. Available from: `https://www.youtube.com/watch?v=iWyQs5iVOrE`. URL: `https://www.youtube.com/watch?v=iWyQs5iVOrE`.

[54] Daniel L Schwartz, Catherine C Chase, Marily A Oppezzo, and Doris B Chin. Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of educational psychology*, 103(4):759, 2011.

[55] Daniel L Schwartz, Jessica M Tsang, and Kristen P Blair. *The ABCs of how we learn: 26 scientifically proven approaches, how they work, and when to use them.* WW Norton & Company, 2016.

[56] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 59–68, New York, NY, USA, 2019. Association for Computing Machinery. `doi:10.1145/3287560.3287598`.

[57] Hong Shen, Wesley H. Deng, Aditi Chattopadhyay, Zhiwei Steven Wu, Xu Wang, and Haiyi Zhu. Value cards: An educational toolkit for teaching social impacts of machine learning through deliberation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 850–861, New York, NY, USA, 2021. Association for Computing Machinery. `doi:10.1145/3442188.3445971`.

[58] Michael Skirpan, Nathan Beard, Srinjita Bhaduri, Casey Fiesler, and Tom Yeh. Ethics education in context: A case study of novel ethics activities for the cs classroom. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, SIGCSE '18, page 940–945, New York, NY, USA, 2018. Association for Computing Machinery. `doi:10.1145/3159450.3159573`.

[59] Heather Spradley. Economics and computation (cs136) - 2019 fall, 2019. URL: `https://embeddedethics.seas.harvard.edu/classes/cs-136-2019-fall`.

[60] Stanford University. Embedded ethics. `https://embeddedethics.stanford.edu/`.

[61] Jon R Star, Bethany Rittle-Johnson, Kelley Durkin, Kristie Newton, Courtney Pollack, Kathleen Lynch, and Claire Gogolen. The impact of a comparison curriculum in algebra i: A randomized experiment. *Society for Research on Educational Effectiveness*, 2013.

[62] Kurt VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational psychologist*, 46(4):197–221, 2011.

[63] Kurt VanLehn, Chandrani Banerjee, Fabio Milner, and Jon Wetzel. Teaching algebraic model construction: a tutoring system, lessons learned and an evaluation. *International Journal of Artificial Intelligence in Education*, 30:459–480, 2020.

[64] Nicholas Vincent, Hanlin Li, Nicole Tilly, Stevie Chancellor, and Brent Hecht. Data leverage: A framework for empowering the public in its relationship with technology companies. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 215–227, New York, NY, USA, 2021. Association for Computing Machinery. `doi:10.1145/3442188.3445885`.

[65] W3C. Inaccessibility of captcha: Alternatives to visual turing tests on the web, 2021. URL: `https://www.w3.org/TR/turingtest/`.

[66] Andy Z. Wang. Beyond embedding ethics: Understanding technology and society at harvard. *The Harvard Crimson*, October 2023. URL: `https://www.thecrimson.com/column/cogito-clicko-sum/article/2023/10/18/wang-embedded-ethics-computer-science/`.

[67] J. Whittle. Is your software valueless? *IEEE Software*, 36(3):112–115, 2019. `doi:10.1109/MS.2019.2897397`.

[68] J. Whittle, M. A. Ferrario, W. Simm, and W. Hussain. A case for human values in software engineering. *IEEE Software*, 38(1):106–113, 2021. `doi:10.1109/MS.2019.2956701`.

[69] David Wright. A framework for the ethical impact assessment of information technology. *Ethics and Inf. Technol.*, 13(3):199–226, sep 2011. `doi:10.1007/s10676-010-9242-6`.

[70] Nick Young and Shriram Krishnamurthi. *Early Post-Secondary Student Performance of Adversarial Thinking*, page 213–224. Association for Computing Machinery, New York, NY, USA, 2021. URL: `https://doi.org/10.1145/3446871.3469743`.

[71] Eric Yu, Paolo Giorgini, Neil Maiden, and John Mylopoulos, editors. *Social Modeling for Requirements Engineering.* MIT Press, 2011.