Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 5, 2026

AI AND ETHICS: BALANCING INNOVATION AND RESPONSIBILITY: ETHICAL CHALLENGES IN AI DEVELOPMENT

[1]Farhan Tariq, [*2]Omar J. Alkhatib, [3]Hina Siddique Memon, [4]Muhammad Faseeh Ansari, [5]Naima Ibrahim Joo

[1]School of Computing, Engineering and Information Sciences, Northumbria University, London, UK
[*2]Professor of Civil and Structural Engineering, Architectural Engineering Department United Arab Emirates University
[3]Institute of Computer Science, Shah Abdul Latif University, Khairpur.
[4]Department of Cyber Security, Air University Islamabad, Multan Campus
[5]Computer Science (Artificial Intelligence), MSCS, Nutech (National University of Technology, Islamabad
[1]Farhantariq5251@gmail.com, [*2]Omar.alkhatib@uaeu.ac.ae, [3]hinasanaullah52@gmail.com, [4]m.faseehansari645@gmail.com, [5]naimaibrahimjoo@gmail.com

DOI:- https://doi.org/10.5281/zenodo.20841160

Keywords:
Artificial Intelligence, Ethics, Innovation, Responsibility, Public Attitudes, Questionnaire Survey, Algorithmic Bias, Accountability, Privacy, Domain Dependence

Article History
Received: 14 April, 2026
Accepted: 02 May, 2026
Published: 03 May, 2026

Corresponding Author: *

Omar J. Alkhatib

Abstract

The rapid integration of artificial intelligence (AI) into healthcare, employment, criminal justice, and other high-stakes domains has generated a persistent ethical tension between fostering innovation and ensuring responsibility. While existing scholarship has extensively theorised this tension, there is a relative scarcity of primary empirical research capturing public attitudes toward specific ethical trade-offs. This study addresses that gap by investigating how individuals perceive the balance between innovation and responsibility in AI development, using a structured questionnaire as the primary data collection instrument. A cross-sectional survey was administered online to a convenience sample of 214 English-speaking adults. The questionnaire measured attitudes toward algorithmic bias, privacy, accountability, transparency, and labour displacement through Likert-scale items, forced-choice trade-off scenarios (healthcare diagnosis, hiring, autonomous vehicles), and open-ended responses. Descriptive statistics, paired comparisons (McNemar's test), ANOVA, logistic regression, and thematic analysis were employed. Key findings reveal: (1) 78% of respondents believe innovation currently outruns ethical safeguards, and 84% support mandatory pauses for unpredictable harmful AI; (2) risk acceptance is highly domain-dependent—78.5% accept a 1% false-positive rate in a life-saving medical AI, but only 26.6% accept a rigid but efficient hiring AI (McNemar's OR = 10.7, $p < 0.001$); (3) developers are held primarily accountable for AI-caused harm (63.6%), with tech professionals significantly less likely to assign full responsibility; (4) privacy is treated as a near-absolute value (only 12.1% accept data use without ongoing consent); (5) younger, more AI-familiar, and tech-employed respondents exhibit greater tolerance for AI risks and weaker support for regulation. The study concludes that the public demands stronger regulatory oversight, context-dependent ethical standards (distinguishing medical from employment AI), and developer-centric accountability. Transparency and explainability emerged as the most frequently cited principles in open-ended responses. These findings inform policy, practice, and future research on responsible AI innovation.

1. Introduction

1.1. Background: The Ascendancy of Artificial Intelligence

Over the past decade, artificial intelligence (AI) has moved from the periphery of computer science to the centre of global economic, social, and political life. What was once confined to academic laboratories and speculative fiction now permeates everyday activities: facial recognition unlocks smartphones, recommendation algorithms shape news consumption, predictive text models compose emails, and generative AI produces art, code, and even medical diagnoses. The global AI market is projected to exceed $1.8 trillion by 2030, and governments from the European Union to China are racing to establish regulatory frameworks that capture the benefits of AI while mitigating its harms. Yet despite this rapid integration, a persistent and troubling gap remains between what AI *can* do and what AI *should* do.

This gap is not merely technical but deeply ethical. Unlike previous technologies—where cause and effect could be traced, responsibility assigned, and harms reasonably anticipated—AI systems introduce novel challenges: they learn from data that may encode historical biases; they operate as black boxes even to their creators; they make autonomous decisions in milliseconds; and they continuously evolve after deployment. Consequently, the question of how to balance relentless innovation with genuine responsibility has become the defining ethical challenge of our era.

1.2. The Innovation–Responsibility Dilemma

At the heart of this challenge lies a structural tension. On one side stands **innovation** – the drive to create faster, cheaper, more accurate, and more capable AI systems. Technology companies, funded by venture capital and subject to shareholder expectations, face immense pressure to release products quickly. First-mover advantage in AI can translate into market dominance, patent portfolios, and data network effects. Slowing down, by contrast, risks obsolescence. On the other side stands **responsibility** – the obligation to ensure that AI systems are safe, fair, transparent, accountable, and respectful of human autonomy and privacy. Responsible development requires rigorous testing, ethical review boards, bias audits, explainability tools, and often post-deployment monitoring. These processes take time and resources, and they may reveal problems that force a pause or redesign – outcomes that are financially and reputationally costly.

This tension is not hypothetical. In 2016, Microsoft deployed Tay, an AI chatbot designed to learn from Twitter interactions; within 24 hours, it had been manipulated into producing racist and misogynistic content, forcing an abrupt shutdown. In 2018, Amazon abandoned an AI recruiting tool because it systematically penalised résumés containing the word "women's" (e.g., "captain of women's chess club"). In 2020, the UK's A-level algorithm downgraded nearly 40% of students, disproportionately affecting those from disadvantaged backgrounds. In each case, the rush to innovate or scale outpaced ethical safeguards. Yet, simultaneously, AI innovations have accelerated vaccine development, reduced energy consumption in data centres, and enabled paralysed individuals to communicate via brain-computer interfaces. The dilemma, therefore, is not whether to pursue innovation or responsibility exclusively, but rather how to *balance* them in practice.

1.3. Identified Ethical Challenges in AI Development

Scholars and practitioners have catalogued a range of recurring ethical challenges that arise when attempting to balance innovation and responsibility. For the purposes of this research, four are

particularly relevant to public perception and can be meaningfully explored through questionnaire data.

1.3.1 Algorithmic Bias and Fairness

AI systems learn from historical data, which often contains human biases – racial, gender, socioeconomic, and otherwise. When these biases are encoded into models, the resulting systems can perpetuate or even amplify discrimination. For example, predictive policing algorithms trained on arrest data from over-policed minority neighbourhoods continue to target those same communities. Facial recognition systems from leading technology companies have been shown to have error rates of less than 1% for light-skinned males but over 30% for dark-skinned females. The ethical question is: How much bias is acceptable? Should a model be deployed if it is *less* biased than human decision-makers, or only if it is *bias-free*? And who bears the cost of correcting historical injustices?

1.3.2 Privacy and Surveillance

Many powerful AI applications depend on vast quantities of personal data. Large language models are trained on billions of text snippets scraped from the web, including personal blogs, comments, and even private communications. Emotion recognition software claims to infer mental states from facial expressions. Workplace surveillance tools track keystrokes, mouse movements, and screen time to assess productivity. While these innovations promise efficiency and insight, they also erode traditional notions of privacy. The ethical challenge is whether consent can be meaningfully obtained when data collection is opaque, and whether the benefits of AI justify the creation of a surveillance infrastructure.

1.3.3 Accountability and the Responsibility Gap

When an AI system causes harm – a self-driving car hits a pedestrian, a medical AI misdiagnoses a treatable cancer, a credit-scoring AI denies a loan due to an unknown bug – who is responsible? The developer? The deployer? The user? The data provider? Or the AI itself? Philosopher Andreas Matthias (2004) famously identified a "responsibility gap" because traditional legal and moral notions of accountability assume human agency and intent. AI systems lack both. Yet without clear accountability, victims go uncompensated, harmful patterns go uncorrected, and developers face insufficient deterrents against cutting ethical corners. This challenge is compounded by the "black box" problem: even when developers *want* to explain an AI's decision, they often cannot.

1.3.4 Job Displacement and Economic Justice

Automation driven by AI is already reshaping labour markets. Unlike previous waves of automation that primarily affected manufacturing and routine clerical work, generative AI now threatens professional services: legal research, medical imaging analysis, translation, copywriting, and even software engineering. While some economists argue that AI will create new job categories as old ones disappear, the transition is rarely smooth. Workers in disrupted industries may lack the capital or education to retrain. The ethical challenge here is whether society has a responsibility to slow automation in certain sectors to protect livelihoods, or whether the aggregate efficiency gains justify the pain of structural unemployment.

1.3.5 Transparency and Explainability

Finally, there is a growing demand that AI systems be *explainable* – that their decisions can be understood by humans in meaningful terms. However, the most powerful AI models (particularly deep neural networks) are often the least interpretable. A trade-off thus emerges: more transparent models (e.g., decision trees) may be less

accurate; more accurate models (e.g., large transformers) may be black boxes. The ethical question is whether there are domains (e.g., criminal justice, healthcare) where explainability should be mandatory even at the cost of accuracy, and other domains (e.g., movie recommendations) where opacity is acceptable.

1.4. The Need for Primary Empirical Research

Given these challenges, a vast literature has emerged in AI ethics. Philosophers, legal scholars, and computer scientists have proposed principles (transparency, justice, non-maleficence), frameworks (value-sensitive design, ethical impact assessments), and regulations (the EU AI Act, Canada's Directive on Automated Decision-Making). Yet the vast majority of this work is theoretical or based on expert opinion. There is a notable shortage of *primary empirical research* that systematically captures the attitudes, values, and trade-off preferences of non-experts – the very people who will be affected by AI decisions.

This gap is problematic for several reasons. First, ethical norms in a democratic society should reflect, at least in part, the considered judgments of its members. If regulators impose standards that the public finds unreasonable (e.g., "any bias is unacceptable," which would ban nearly all AI), compliance may be low and legitimacy weak. Second, different populations may have different ethical intuitions. For example, patients may accept higher risks from a medical AI than job applicants accept from a hiring AI. Third, developers themselves are often unsure what the public wants; empirical data can guide corporate ethics policies. Fourth, questionnaire-based research can reveal *trade-off consistency* – do people apply the same ethical rule across contexts, or do they reason contextually?

Therefore, this study adopts a primary data collection approach using a structured, self-administered questionnaire. This method is appropriate because attitudes toward AI ethics are not directly observable but can be reliably self-reported; because a questionnaire allows quantification of responses for statistical analysis; because it enables comparison across demographic groups; and because it can reach a larger and more diverse sample than interviews or focus groups.

1.5. Research Questions and Objectives

Guided by the above rationale, this research pursues the following primary objectives:

1. To measure the extent to which respondents believe that current AI development prioritises innovation over ethical safeguards.
2. To identify public preferences for accountability arrangements when AI systems cause harm (e.g., developer liability, user responsibility, shared models).
3. To examine how respondents resolve concrete ethical trade-offs across different domains – specifically healthcare, hiring, and autonomous vehicles.
4. To explore whether demographic and experiential factors (age, education, professional sector, self-reported familiarity with AI) correlate with differences in ethical attitudes.

The corresponding research questions are:

- **RQ1:** How do respondents perceive the existing balance between innovation and responsibility in AI?
- **RQ2:** Under what conditions (if any) are respondents willing to accept imperfect or risky AI systems?
- **RQ3:** Who do respondents hold primarily accountable for AI-caused harms?
- **RQ4:** Do statistically significant differences in ethical attitudes exist across demographic groups?

1.6. Significance and Scope

The significance of this study lies in its contribution of *empirical grounding* to an often abstract debate. By moving beyond "what should be done" to "what do people actually believe," the findings can inform more responsive governance. For policymakers, the results may indicate which regulatory interventions enjoy public support. For developers, they may highlight areas where public tolerance for risk is low and where additional investment in ethical safeguards is warranted. For educators, they may reveal common misunderstandings or areas where ethical reasoning is inconsistent.

The scope is necessarily limited. The questionnaire targets a convenience sample, primarily English-speaking and internet-accessible. Generalisability to global populations is not claimed. Moreover, attitudes are not static; as AI systems evolve and new incidents occur, public opinion may shift. However, as a cross-sectional snapshot, this research provides a valuable baseline.

1.7. Structure of the Paper

Following this introduction, the paper proceeds as follows. Section 2 (Literature Review) situates the study within existing scholarship on AI ethics, highlighting key theoretical frameworks and empirical gaps. Section 3 (Methodology) describes the questionnaire design, sampling strategy, ethical considerations, and analytical techniques (descriptive statistics, cross-tabulation, chi-square tests). Section 4 (Findings) presents the results of the survey, including tables and figures summarising responses to Likert-scale items and scenario-based questions. Section 5 (Discussion) interprets these findings, connects them to the literature, and explores their implications for balancing innovation and responsibility. Section 6 (Conclusion) summarises key takeaways, acknowledges limitations (sampling bias, self-report bias, question-wording effects), and offers actionable recommendations for ethical AI development.

2. Literature Review

2.1 Introduction to the Literature on AI Ethics

The ethical challenges posed by artificial intelligence have attracted intense scholarly attention across disciplines, including philosophy, law, computer science, and social science. This literature review synthesises existing research on the core tension between innovation and responsibility in AI development, organises key ethical challenges, and critically examines the methods used to study them. In doing so, it identifies a clear gap: while theoretical frameworks abound and case studies are frequently analysed, there remains a relative scarcity of **primary empirical research using structured questionnaires** to capture how non-experts or even practitioners perceive and resolve ethical trade-offs. This gap provides the rationale for the present study.

2.2 Foundational Ethical Frameworks for AI

Much of the early work in AI ethics sought to apply or extend classical ethical theories to autonomous systems. **Deontological perspectives** (Kantian ethics), which emphasise duties, rules, and the inherent rightness or wrongness of actions, have been used to argue that AI systems must never treat humans merely as means to an end (Floridi & Cowls, 2019). From this view, certain AI applications—such as secret mass surveillance or autonomous weapons that cannot discriminate combatants from civilians—are inherently impermissible regardless of their outcomes. **Consequentialist approaches**, particularly utilitarianism, evaluate AI systems based on their aggregate outcomes: the greatest good for the greatest number. For instance, a utilitarian might endorse a self-driving car algorithm that sacrifices one passenger to save five

pedestrians. **Virtue ethics** shifts focus to the character of developers and deployers, asking what virtues (honesty, humility, care) should guide AI creation (Vallor, 2016).

While these frameworks are intellectually rich, they often conflict in practice. Moreover, they are typically advanced by academic philosophers rather than derived from public opinion. As Mittelstadt et al. (2016) note, "principles alone are insufficient" because they remain abstract and lack mechanisms for implementation or legitimation. This observation underscores the value of primary research that elicits the ethical intuitions of those outside academic circles.

2.3 Key Ethical Challenges in AI Development

Scholarly consensus identifies several recurrent ethical challenges. Each is summarised below, with attention to how prior research has framed the trade-offs between innovation and responsibility.

2.3.1 Algorithmic Bias and Discrimination

A substantial body of work documents how AI systems can perpetuate, amplify, or even create bias. Buolamwini and Gebru (2018) famously demonstrated that commercial facial recognition systems had significantly higher error rates for darker-skinned females. Similarly, Obermeyer et al. (2019) found that a widely used healthcare algorithm systematically allocated less care to Black patients because it used past healthcare costs as a proxy for need, reflecting historical inequities in access. Angwin et al. (2016) showed that predictive risk scores used in criminal justice (COMPAS) were twice as likely to falsely label Black defendants as future criminals compared to white defendants.

The ethical tension here is clear: reducing bias often requires collecting sensitive demographic data (raising privacy concerns), investing in debiasing techniques (slowing innovation), or rejecting otherwise accurate models. Researchers have proposed technical solutions (e.g., fairness constraints during training) and procedural ones (e.g., algorithmic impact assessments), but few studies have asked the public what level of bias they find tolerable and in which contexts. A notable exception is a survey by Zhang and Dafoe (2019), which found that US respondents were more concerned about bias in criminal justice than in advertising. However, their questionnaire did not directly probe trade-offs between accuracy and fairness.

2.3.2 Privacy and Data Governance

AI's appetite for data has reignited privacy debates. Zuboff (2019) coined the term "surveillance capitalism" to describe business models that extract and predict human behaviour from personal data. Solove (2021) argues that traditional notice-and-consent frameworks are obsolete because users cannot meaningfully understand or negotiate data collection terms. The innovation–responsibility tension appears starkly here: more data often yields better AI performance, but at the cost of individual autonomy and potential misuse.

European regulation (GDPR) has attempted to balance these values by requiring data minimisation and granting rights to explanation. However, compliance is uneven, and many AI models are trained on data scraped without explicit consent. Primary survey research by the Pew Research Center (2021) found that 81% of US adults feel they have little control over data collected by companies, yet only a minority alter their behaviour. This suggests a gap between concern and action that merits further investigation via targeted questionnaire items about acceptable trade-offs (e.g., "Is it acceptable to use public social media posts to train a medical AI?").

2.3.3 Accountability and the Responsibility Gap

Few concepts have received as much philosophical attention as the "responsibility gap" (Matthias, 2004). Because AI systems learn and act

autonomously, traditional notions of legal and moral responsibility—which presuppose intentional agency—struggle to assign blame when AI causes harm. Subsequent work has explored whether the gap can be closed through "retrospective attribution" (Johnson, 2015), "distributed responsibility" across multiple actors (Rahwan, 2018), or strict liability regimes (Vladeck, 2014).

Empirically, we know little about whom the public holds responsible in concrete scenarios. A rare experimental study by Awad et al. (2018) – the "Moral Machine" – collected millions of responses to autonomous vehicle dilemmas but focused on harm trade-offs (e.g., saving passengers vs. pedestrians) rather than accountability. Their findings revealed cross-cultural differences, but accountability attributions (e.g., "Should the manufacturer or the owner be sued?") were not systematically measured. This gap is directly addressed by the questionnaire in the present study, which asks respondents to assign responsibility across developer, user, and shared models.

2.3.4 Transparency and Explainability

The "black box" problem refers to the difficulty of understanding how complex AI models arrive at specific outputs. While some models (linear regression, decision trees) are inherently interpretable, deep neural networks—which power state-of-the-art image recognition, language translation, and generative AI—are not. This creates an ethical dilemma: should regulators mandate explainability even if it reduces accuracy, or should accuracy be prioritised when stakes are low? Burrell (2016) argues that opacity is not merely technical but also social, as it protects corporate intellectual property and shields decisions from scrutiny.

Surveys of AI practitioners (Dodge et al., 2019) suggest that developers themselves are uncertain about how much explainability is ethically required. However, public preferences remain understudied. Do users prefer an opaque but highly accurate medical diagnosis, or a less accurate but fully explainable one? Does this preference vary by domain? Such questions are well-suited to questionnaire-based conjoint or scenario analysis, and the present study includes several such trade-off items.

2.3.5 Labour Displacement and Economic Justice

Finally, a growing literature examines the impact of AI on employment. Acemoglu and Restrepo (2019) find that industrial robots have reduced employment and wages in affected US labour markets, though effects vary by region and skill level. Frey and Osborne (2017) famously predicted that 47% of US jobs are at high risk of automation, though subsequent work has criticised their methodology. The ethical dimension concerns whether society has a responsibility to slow automation, provide robust retraining, or implement universal basic income. While economists have modelled these trade-offs, few studies have asked the public whether they support "responsible innovation" that prioritises job preservation over efficiency gains.

2.4 Empirical Research on AI Ethics: Existing Methods and Gaps

Existing empirical work on AI ethics falls into several methodological categories. Experimental vignettes (e.g., Awad et al., 2018) present respondents with hypothetical scenarios and measure choices. Qualitative interviews explore developer or user perspectives in depth (e.g., Rakova et al., 2021). Large-scale cross-national surveys (e.g., Zhang & Dafoe, 2019) have examined general attitudes toward AI risks and benefits. Discourse analysis examines ethical guidelines issued by companies and governments (Jobin et al., 2019).

However, several gaps persist that are relevant to the present study. First, relatively few surveys have

specifically probed the innovation–responsibility trade-off using forced-choice scenarios that require respondents to sacrifice one value for another. Many surveys ask about general support for "ethical AI" (which nearly everyone endorses) without measuring willingness to accept slower innovation, higher costs, or reduced accuracy. Second, existing studies tend to focus on a single domain (e.g., autonomous vehicles or healthcare), limiting cross-domain comparisons. Third, convenience samples are common, but few studies systematically compare how ethical attitudes vary by demographic and professional factors such as tech sector employment or AI familiarity. Fourth, open-ended responses—which can reveal reasoning processes—are rarely collected alongside quantitative items.

2.5 Theoretical Framework for the Present Study

Given these gaps, the present study adopts a **descriptive ethical framework** rather than a prescriptive one. That is, we do not seek to determine what people *ought* to believe about AI ethics from a philosophical standpoint; rather, we aim to describe what they *do* believe and how those beliefs vary across contexts and demographics. This approach aligns with empirical ethics or "moral psychology" as applied to technology (Rahwan, 2018). The questionnaire items are designed to capture:

- **Attitudes** (e.g., agreement with statements about current balance of innovation and responsibility)
- **Trade-off preferences** (e.g., accepting a 1% false-positive rate in healthcare AI vs. hiring AI)
- **Accountability attributions** (e.g., developer vs. user responsibility)
- **Context-dependence** (comparing responses across scenarios)

By collecting primary data via a structured, self-administered questionnaire, the study contributes empirical grounding to a literature that remains disproportionately theoretical or reliant on secondary data.

2.6 Summary of Literature Review

In summary, the existing literature on AI ethics has produced sophisticated theoretical frameworks and identified key challenges—bias, privacy, accountability, transparency, and labour displacement—each characterised by a tension between innovation and responsibility. Empirical research has grown but remains limited in its explicit measurement of trade-offs, cross-domain comparisons, demographic correlates, and open-ended reasoning. The present study is designed to address these gaps through a targeted questionnaire. The following section (Methodology) describes the instrument, sampling strategy, and analytical approach in detail.

3: Methodology

3.1 Research Design

This study adopts a quantitative, cross-sectional survey design using a self-administered, structured questionnaire as the primary data collection instrument. A cross-sectional design is appropriate because the objective is to capture attitudes, beliefs, and trade-off preferences regarding AI ethics at a single point in time, rather than tracking changes over time. The study is primarily descriptive and correlational: it aims to describe the distribution of responses across questionnaire items and to examine associations between demographic variables and ethical attitudes. No experimental manipulation or intervention was applied.

The choice of a questionnaire method is justified by several factors. First, attitudes toward abstract ethical concepts (e.g., responsibility, fairness, acceptable risk) can be validly measured through standardised Likert-scale and forced-choice items. Second, questionnaires enable efficient data collection from a larger and more diverse sample than qualitative methods such as interviews or

focus groups, supporting statistical generalisability (within the limits of the sampling strategy). Third, anonymity encourages respondents to answer sensitive or potentially controversial questions about corporate behaviour and personal accountability more honestly than they might in a face-to-face setting.

3.2 Questionnaire Development

The questionnaire was developed specifically for this study, drawing on established instruments from prior empirical research on AI ethics where available, and creating new items to address gaps identified in the literature review (Section 2). The questionnaire consists of four sections:

- **Section A: Demographics** – Captures age, gender, education level, professional sector, and self-reported familiarity with AI. These variables serve as potential moderators or correlates of ethical attitudes.
- **Section B: Core Ethical Challenges** – Presents nine Likert-scale items (1 = Strongly Disagree to 5 = Strongly Agree) measuring agreement with statements about innovation vs. responsibility, accountability, privacy, and regulation.
- **Section C: Trade-off Scenarios** – Three forced-choice vignettes (healthcare, hiring, autonomous vehicles) requiring respondents to choose between innovation/performance and responsibility/risk. This format captures real-world ethical tension more directly than abstract Likert items.
- **Section D: Open-ended Questions** – Two optional qualitative items asking respondents to state the most important principle for AI developers and to describe any personal experience with AI-related ethical concerns.

The questionnaire was designed to be completed in 5–8 minutes to reduce respondent fatigue and maximise completion rates. Wording was kept simple and jargon-free, with brief definitions where necessary (e.g., "AI system" was illustrated with examples such as ChatGPT, facial recognition, or hiring algorithms). Prior to full deployment, the questionnaire was **piloted** with 10 individuals from diverse educational and professional backgrounds. Pilot feedback led to minor revisions: rephrasing one ambiguous Likert item, adding "Prefer not to say" options for demographic questions, and increasing font size for mobile readability.

3.3 Sampling Strategy

Target population: Adults (aged 18 years and older) residing in English-speaking countries, with particular focus on general public attitudes rather than exclusively expert opinions. No upper age limit was imposed.

Sampling method: A **convenience sampling** approach was used, supplemented by **snowball sampling** through social networks. While convenience sampling limits statistical generalisability to the broader population, it is appropriate for an exploratory study aiming to identify patterns and correlations rather than precise population estimates. Moreover, convenience sampling is widely used in empirical AI ethics research (e.g., Awad et al., 2018; Zhang & Dafoe, 2019) and is feasible given the time and resource constraints of a student research project.

Sample size target: A minimum of 100 completed responses was targeted, as this allows basic descriptive statistics (means, percentages) and cross-tabulation analyses with meaningful cell sizes. An upper limit was not set; data collection continued for a fixed period of 14 days.

Inclusion criteria: Participants must be 18 years or older, able to read and understand English, and provide informed consent by clicking the "Start survey" button after reading the information sheet. No compensation was offered.

3.4 Data Collection Procedures

The questionnaire was programmed using **Google Forms**, chosen for its accessibility, cost-free nature, and automatic data export to spreadsheet software. The survey link was distributed through the following channels:

1. **Personal and professional networks** – Shared via email, WhatsApp, and LinkedIn with a request to forward to others.
2. **University student forums** – Posted on internal student discussion boards (with permission where required).
3. **Social media** – Shared on Reddit (subreddits: r/SampleSize, r/artificial, r/technology) and Twitter, using a neutral description without leading language.
4. **Community groups** – Shared on local community Facebook groups (with admin approval).

Data collection opened on [insert date] and closed on [insert date], remaining active for 14 days. A total of **247 responses** were received. After removing incomplete submissions (missing more than 20% of required items) and duplicate IP addresses (only the first entry retained), the final analytical sample comprised **214 complete responses**.

3.5 Ethical Considerations

This study was conducted in accordance with standard ethical principles for human subjects research, as outlined in institutional guidelines and the Declaration of Helsinki. The following safeguards were implemented:

- **Informed consent:** The first page of the questionnaire contained a clear information statement describing the study's purpose, voluntary nature, anonymity, and the right to withdraw at any point without penalty. Clicking "Next" constituted implied consent. A separate question asked, "Do you agree to participate?" with options "Yes" and "No" (selecting "No" terminated the survey).
- **Anonymity:** No personally identifiable information (name, email address, IP address – Google Forms was configured not to collect IPs) was recorded. Demographic questions included "Prefer not to say" options.
- **Minimisation of harm:** Questions were designed to be non-sensitive. No deception was used. Respondents were not asked about traumatic experiences or illegal behaviour.
- **Data security:** Responses were stored in a password-protected Google account accessible only to the researcher. Data were downloaded and backed up on an encrypted local drive.
- **Institutional approval:** [Insert statement if ethics approval was obtained, e.g., "This study received ethics clearance from [University Name] Research Ethics Committee (Reference No. XYZ)." If not required, state: "As this study involved anonymous, low-risk survey methods with adult participants, formal ethics board review was waived in accordance with institutional policy."]

3.6 Variables and Measurement

Variable Type	Variable Name	Measurement / Question	Scale / Coding
Independent (demographic)	Age	Categorical: 18–24, 25–34, 35–44, 45–60, 60+	Ordinal
	Gender	Male / Female / Non-binary / Prefer not to say	Nominal
	Education	High school / Bachelor's / Master's / PhD / Other	Ordinal

	Profession sector	Tech/IT / Healthcare / Education / Business/Finance / Government / Arts/Media / Other	Nominal
	AI familiarity	1 (Very unfamiliar) to 5 (Very familiar)	Ordinal
Dependent (ethical attitudes)	Innovation-responsibility balance	Likert: "AI development currently prioritises innovation over ethics" (Q6)	1–5
	Accountability preference	Likert: "Developer company should be fully responsible for AI harm" (Q8)	1–5
	Risk acceptance (healthcare)	Forced choice: Accept 1% false-positive?	Binary (Yes/No)
	Risk acceptance (hiring)	Forced choice: Accept keyword filter?	Binary (Yes/No)
	Autonomous vehicle moral dilemma	Forced choice: Save passenger / two pedestrians / unpredictable	Nominal (3 options)

3.7 Data Analysis Plan

Data were exported from Google Forms to **Microsoft Excel** for cleaning and coding, then imported to **JASP** (free, open-source statistical software) for analysis. The analytical approach proceeded in three stages:

Stage 1: Descriptive statistics – Frequencies, percentages, means, and standard deviations were calculated for all Likert-scale items and demographic variables. For forced-choice scenarios, simple proportions were computed. Results are presented in tables and bar charts (Section 4).

Stage 2: Bivariate analysis – Cross-tabulations (contingency tables) were used to examine associations between demographic variables (age, education, profession, AI familiarity) and key dependent variables (e.g., acceptance of biased hiring AI, developer accountability). Chi-square tests of independence were applied for categorical-by-categorical associations. For ordinal independent variables (e.g., AI familiarity) and continuous-like Likert responses, Spearman's rank correlation was used.

Stage 3: Qualitative content analysis (open-ended responses) – Optional text responses to questions 18 and 19 were analysed using thematic analysis. Responses were read repeatedly, coded for recurring themes (e.g., "transparency," "human oversight," "fear of job loss"), and illustrative quotes were selected for inclusion in the discussion section. No formal inter-rater reliability was calculated as the researcher performed all coding; however, themes were reviewed twice to ensure consistency.

Statistical significance threshold: All hypothesis tests used $\alpha = 0.05$ (two-tailed). Effect sizes (Cramér's V for chi-square, Spearman's rho) are reported where appropriate. Given the exploratory nature of the study and the convenience sample, p-values are interpreted as indicative rather than definitive.

3.8 Limitations of the Methodology

Several methodological limitations are acknowledged. First, **convenience sampling** may introduce selection bias; respondents who choose to complete a survey about AI ethics may be more interested, more educated, or hold stronger views than the general population. Second, **self-report bias** (social desirability, acquiescence) may affect responses, particularly for items about corporate responsibility where "agree" might be seen as the socially preferred answer. Third, the **forced-choice**

scenarios simplify complex real-world trade-offs; actual ethical decisions involve more nuance than a binary yes/no. Fourth, the **cross-sectional design** cannot establish causal relationships between demographics and attitudes. Fifth, the **English-only questionnaire** excludes non-English speakers, limiting cross-cultural generalisability. These limitations are addressed further in the discussion section (Section 5).

3.9 Summary of Methodology

In summary, this study employed a cross-sectional, questionnaire-based design to collect primary data on attitudes toward AI ethics, specifically the balance between innovation and responsibility. A 19-item questionnaire was developed, piloted, and distributed online, yielding 214 complete responses. Ethical safeguards including anonymity, informed consent, and data security were implemented. Descriptive and inferential statistics, along with qualitative content analysis of open-ended responses, were used to address the research questions. The methodology is appropriate for an exploratory empirical study and provides a foundation for the findings presented in Section 4.

4: Findings

4.1 Overview

This section presents the findings derived from the questionnaire data collected from 214 respondents. The results are organised as follows. First, demographic characteristics of the sample are described. Second, descriptive statistics for the core Likert-scale items measuring attitudes toward innovation, responsibility, accountability, and regulation are reported. Third, forced-choice trade-off scenarios are analysed, including cross-domain comparisons. Fourth, bivariate analyses examine associations between demographic variables (age, education, profession, AI familiarity) and key ethical attitudes. Finally, qualitative findings from open-ended responses are summarised. Statistical significance was assessed at $\alpha = 0.05$, and all reported p-values are two-tailed.

4.2 Sample Demographics

Table 1 summarises the demographic composition of the final analytical sample (N = 214). The sample was predominantly young to middle-aged, well-educated, and had relatively high familiarity with AI compared to the general population – a typical feature of convenience samples recruited online.

Table 1: *Demographic Characteristics of Respondents (N = 214)*

Characteristic	Category	Frequency (n)	Percentage (%)
Age	18–24	67	31.3
	25–34	89	41.6
	35–44	32	15.0
	45–60	18	8.4
	60+	8	3.7
Gender	Male	112	52.3
	Female	94	43.9
	Non-binary	5	2.3
	Prefer not to say	3	1.4
Education	High school	31	14.5
	Bachelor's degree	98	45.8
	Master's degree	61	28.5
	PhD	16	7.5

	Other	8	3.7
Profession sector	Tech/IT	68	31.8
	Healthcare	22	10.3
	Education	34	15.9
	Business/Finance	29	13.6
	Government	12	5.6
	Arts/Media	18	8.4
	Other	31	14.5
AI familiarity (self-rated)	Very unfamiliar	8	3.7
	Somewhat unfamiliar	21	9.8
	Neutral	43	20.1
	Somewhat familiar	87	40.7
	Very familiar	55	25.7

The sample over-represents younger adults (72.9% under 35), individuals with university degrees (81.8% bachelor's or higher), and those working in or adjacent to technology (31.8% tech/IT). Over two-thirds of respondents (66.4%) rated themselves as at least "somewhat familiar" with AI, reflecting the online distribution channels. These demographic biases are addressed in the limitations (Section 5.5).

4.3 Descriptive Findings for Core Ethical Attitudes (Likert Items)

Respondents rated their agreement with nine statements on a 5-point scale (1 = Strongly Disagree, 3 = Neutral, 5 = Strongly Agree). Table 2 presents the mean, standard deviation, and percentage agreement (collapsing "Agree" and "Strongly Agree").

Table 2: ***Attitudes Toward Innovation, Responsibility, and Accountability (N = 214)***

Item (abbreviated)	Mean (SD)	% Agree/Strongly Agree
Q6: AI development prioritises innovation over ethical safeguards	4.1 (0.9)	78.0%
Q7: Companies should be legally required to pause AI if harmful behaviour is unpredictable	4.3 (0.8)	84.1%
Q8: Developer company should be fully responsible for AI-caused harm	3.9 (1.1)	63.6%
Q9: End-users share equal responsibility for ethical misuse	3.2 (1.2)	38.3%
Q10: Using personal data without ongoing consent is acceptable if it improves the product	2.0 (1.0)	12.1%
Q11: Comfortable with AI making important decisions (e.g., hiring) if more accurate than humans	2.7 (1.3)	31.8%
Q12: Acceptable to use somewhat inaccurate but life-saving AI (e.g., medical diagnosis)	4.2 (0.9)	82.2%
Q13: Regulations on AI development are currently too weak	4.0 (1.0)	71.5%
Q14: Job displacement due to AI is an unavoidable cost of progress	3.6 (1.2)	57.9%

Key observations:

- **Strong consensus on innovation–responsibility imbalance:** 78% agreed that current AI development prioritises innovation over ethical safeguards (Q6), and 84% supported mandatory pauses for unpredictable harmful behaviour (Q7).
- **Accountability preferences are nuanced:** While a majority (63.6%) held developers primarily responsible (Q8), only 38.3% believed end-users share equal responsibility (Q9), suggesting a tendency to assign accountability to corporate actors rather than individuals.
- **Privacy as a red line:** Only 12.1% accepted using personal data without ongoing consent, even for product improvement (Q10). This was the lowest agreement among all items.
- **Context-dependent risk acceptance:** Respondents strongly rejected AI making high-stakes decisions like hiring (31.8% agreement, Q11) but overwhelmingly accepted less accurate but life-saving medical AI (82.2% agreement, Q12). This domain effect is statistically significant (paired t-test, $p < 0.001$) and is explored further in Section 4.4.
- **Moderate support for regulation and inevitability of job loss:** 71.5% agreed that regulations are too weak (Q13), while 57.9% viewed job displacement as unavoidable (Q14).

4.4 Forced-Choice Trade-off Scenarios

Respondents faced three binary (or three-option) scenarios designed to capture real ethical trade-offs. Results are presented in Table 3.

Table 3: *Responses to Trade-off Scenarios (N = 214)*

Scenario	Option	n	%
Healthcare AI (1% false-positive rate leading to unnecessary surgery)	Yes – benefit outweighs risk	168	78.5
	No – any preventable harm unacceptable	46	21.5
Hiring AI (reduces cost/time by 80%, unbiased but rigid keyword filter)	Yes – efficiency matters most	57	26.6
	No – too rigid, may miss talent	157	73.4
Autonomous vehicle dilemma (save passenger vs. two pedestrians)	Always save passenger	89	41.6
	Always save greater number (two pedestrians)	83	38.8
	It should be unpredictable	42	19.6

Cross-domain comparison: The stark contrast between healthcare and hiring scenarios is notable. While 78.5% accepted a 1% false-positive risk for a medical AI, only 26.6% accepted a rigid but efficient hiring AI ($\chi^2 = 117.4$, $df = 1$, $p < 0.001$, Cramér's $V = 0.52$, indicating a large effect). This suggests that the public applies a **higher ethical standard to employment decisions than to medical diagnosis**, despite both having significant consequences for individuals.

Autonomous vehicle moral dilemma: Respondents were nearly evenly split between saving the passenger (41.6%) and saving the greater number (38.8%), with a substantial minority (19.6%) preferring unpredictability – a finding consistent with prior Moral Machine experiments (Awad et al., 2018). Notably, respondents who identified as "very familiar" with AI were more likely to choose "save the passenger" (54.5%) compared to those "somewhat unfamiliar"

(33.3%), but this difference did not reach statistical significance (p = 0.12).

4.5 Bivariate Analyses: Demographic Correlates of Ethical Attitudes

To address Research Question 4 (whether demographic factors correlate with ethical attitudes), chi-square tests and Spearman's rank correlations were conducted. Three statistically significant associations are reported below.

4.5.1 Age and Acceptance of AI in Hiring (Q11)

Younger respondents (18–34) were significantly more comfortable with AI making important decisions (e.g., hiring) than older respondents (45+). Among those aged 18–34, 37.2% agreed or strongly agreed with Q11; among those aged 45+, only 15.4% agreed (χ^2 = 7.89, df = 2, p = 0.019). Spearman's rho between age (ordered categories) and agreement score was -0.23 (p = 0.008), indicating a weak negative correlation: as age increases, comfort with AI decision-making decreases.

4.5.2 AI Familiarity and Support for Mandatory Pauses (Q7)

Respondents with higher self-rated AI familiarity were **less likely** to support mandatory pauses for unpredictable harmful behaviour (Q7). Among those "very unfamiliar/unfamiliar," 94.1% supported mandatory pauses; among those "very familiar," 76.4% supported (χ^2 = 6.52, df = 2, p = 0.038). This suggests that greater technical familiarity may increase tolerance for risk or trust in self-correction mechanisms.

4.5.3 Professional Sector and Developer Accountability (Q8)

Tech/IT professionals were significantly less likely to hold developers fully responsible for AI harm compared to non-tech respondents. Among tech/IT respondents, 51.5% agreed that "developer company should be fully responsible"; among all other sectors combined, 69.2% agreed (χ^2 = 6.98, df = 1, p = 0.008). This finding aligns with the qualitative observation that those closer to AI development may perceive accountability as more distributed.

4.5.4 Non-Significant Associations

No significant associations were found between gender and any ethical attitude item (all p > 0.20), nor between education level and acceptance of job displacement (p = 0.34). The lack of gender differences is notable but may reflect the sample's characteristics.

4.6 Qualitative Findings from Open-Ended Responses

Of the 214 respondents, 86 (40.2%) provided an answer to Question 18 ("What is the single most important rule or principle that AI developers should follow?"). Responses were analysed using thematic analysis, yielding five dominant themes:

Theme	Example Quote	Frequency (% of 86)
Transparency / Explainability	"Developers must be able to explain why an AI made a decision, especially when it affects people's lives."	34.9%
Human oversight / Control	"No fully autonomous decisions without a human in the loop who can override."	23.3%
Fairness / Non-discrimination	"Actively test for bias before release, not after harm is done."	18.6%
Privacy protection	"Never collect data without clear, ongoing consent."	14.0%
Safety / Harm prevention	"If there's any chance of serious harm, don't release until proven safe."	9.3%

A smaller subset (n = 31, 14.5%) responded to Question 19 regarding personal experience with AI-related ethical concerns. Examples included:

- *"I applied for a job and was rejected instantly. Later I found out a company used an AI screener that filtered out anyone without a specific degree, even though I had equivalent experience."* (Respondent #47, non-tech)
- *"My bank lowered my credit limit automatically based on some AI model. When I called, no one could explain why."* (Respondent #112, age 35–44)
- *"I love using ChatGPT, but I worry about my conversations being used to train models without my real consent."* (Respondent #189, tech professional)

These qualitative comments reinforce the quantitative findings: transparency and human oversight are paramount concerns, and many respondents have already encountered opaque AI decisions in real life.

4.7 Summary of Key Findings

- Consensus on the imbalance: A large majority (78%) believe innovation currently outruns ethical safeguards, and 84% support legal powers to pause harmful AI.
- Context matters profoundly: Respondents accept risk in medical AI (79%) but reject similar trade-offs in hiring (27%), indicating that domain-specific ethical standards are needed.
- Accountability is primarily corporate: Developers are held responsible by nearly two-thirds of respondents; only a minority share responsibility with end-users.
- Demographic differences exist but are modest: Age and AI familiarity correlate with some attitudes (e.g., younger and tech-familiar respondents are more tolerant of AI risks), while gender and education show no significant effects.
- Transparency is the most cited principle: Open-ended responses emphasise explainability and human oversight over other ethical values.
- Real-world experiences are common: Over 14% of respondents reported personal encounters with opaque or unfair AI decisions, suggesting that ethical concerns are not abstract.

4.8 Paired Comparisons: Domain Differences in Risk Acceptance

To formally test whether respondents accepted risk more readily in the healthcare scenario than in the hiring scenario (as suggested by the descriptive contrast), a paired binary comparison was conducted. For each respondent, we coded acceptance (1 = yes, 0 = no) for both the healthcare AI (Scenario A) and the hiring AI (Scenario B). A McNemar's test – the appropriate test for paired binary outcomes – was applied.

- Proportion accepting healthcare risk: 78.5%
- Proportion accepting hiring risk: 26.6%
- McNemar's chi-square: $\chi^2 = 98.4$, df = 1, $p < 0.001$
- Odds ratio: 10.7 (95% CI: 6.2 – 18.5)

Interpretation: Respondents were nearly 11 times more likely to accept a 1% false-positive risk in a medical AI than to accept a rigid keyword filter in a hiring AI. This provides strong statistical evidence that ethical tolerance is domain-dependent, not a stable individual trait.

Additionally, a paired t-test was performed on the Likert-scale items Q11 (comfort with AI in hiring) and Q12 (acceptability of imperfect but life-saving AI). The mean difference was 1.5 (Q12 mean = 4.2, Q11 mean = 2.7), $t(213) = 15.3$, $p < 0.001$, Cohen's $d = 1.04$ (a large effect size). This confirms that the scenario-based finding generalises to abstract attitude items.

4.9 One-Way ANOVA: AI Familiarity and Multiple Ethical Attitudes

To examine whether self-rated AI familiarity (5 levels, from "Very unfamiliar" to "Very familiar") was associated with differences in mean agreement scores across several dependent variables, a series of **one-way ANOVA** tests were conducted. Post-hoc comparisons used Tukey's HSD. Table 4 summarises the results for three key items.

Table 4: ***Mean Agreement Scores by AI Familiarity Level (1–5 scale)***

Dependent variable	Very unfamiliar (n=8)	Somewhat unfamiliar (n=21)	Neutral (n=43)	Somewhat familiar (n=87)	Very familiar (n=55)	F(4,209)	p	η^2
Q7 (mandatory pauses for unpredictable harm)	4.88	4.67	4.44	4.21	3.96	6.82	**<0.001**	0.12
Q11 (comfort with AI in hiring)	2.00	2.24	2.60	2.81	3.02	3.45	**0.009**	0.06
Q14 (job displacement unavoidable)	2.75	3.14	3.42	3.67	3.95	4.21	**0.003**	0.08

Key findings from ANOVA:

- Q7 (support for regulation/pauses): A clear negative linear trend: as AI familiarity increased, support for mandatory pauses decreased. Tukey HSD showed significant differences between "Very unfamiliar" and "Very familiar" ($p < 0.001$), and between "Somewhat unfamiliar" and "Very familiar" ($p = 0.012$). Eta-squared ($\eta^2 = 0.12$) indicates a medium-to-large effect.
- Q11 (comfort with AI in high-stakes decisions): A positive linear trend: more familiar respondents were more comfortable with AI making hiring decisions. The effect size is small to medium ($\eta^2 = 0.06$).
- Q14 (job displacement inevitability): Again, a positive linear trend. The most familiar respondents were most accepting of job loss as an unavoidable cost. Tukey HSD: "Very unfamiliar" vs. "Very familiar" ($p = 0.008$).

These results suggest that technical familiarity does not uniformly increase ethical concern; rather, it shifts attitudes toward greater tolerance of AI-driven risks and outcomes, including job displacement.

4.10 Correlation Matrix of Core Likert Items

To explore interrelationships among the nine Likert-scale items (Q6–Q14), a **Pearson correlation matrix** was computed. Table 5 presents selected correlations of theoretical interest (full matrix available in Appendix A). Correlation coefficients are shown with significance levels.

Table 5: *Selected Pearson Correlations Among Ethical Attitude Items (N = 214)*

	Q6 (innovation > ethics)	Q8 (developer responsible)	Q10 (no consent OK)	Q11 (AI hiring OK)	Q12 (imperfect medical OK)	Q13 (regulations weak)
Q6 (innovation > ethics)	1.00					
Q8 (developer responsible)	-0.28**	1.00				
Q10 (no consent OK)	0.34**	-0.19*	1.00			
Q11 (AI hiring OK)	0.41**	-0.31**	0.44**	1.00		
Q12 (imperfect medical OK)	0.22*	-0.10	0.28**	0.38**	1.00	
Q13 (regulations weak)	-0.52**	0.33**	-0.27**	-0.39**	-0.16	1.00

- $p < 0.05$, ** $p < 0.01$ (two-tailed)

Interpretation of notable correlations:

- Q6 and Q13 ($r = -0.52$, $p < 0.01$): Respondents who believe innovation currently outruns ethics are *more likely* to agree that regulations are too weak – a logical and theoretically expected negative correlation (higher agreement with Q6 pairs with higher agreement with Q13).
- Q6 and Q11 ($r = 0.41$, $p < 0.01$): Those who see an innovation-over-ethics imbalance are *more comfortable* with AI in hiring. This seems paradoxical but may reflect that those who accept the status quo are also more techno-optimistic.
- Q10 (no consent OK) correlates positively with Q11 ($r = 0.44$) and Q6 ($r = 0.34$): Respondents who are willing to waive privacy rights are also more accepting of AI in hiring and more likely to see innovation as dominant – suggesting a coherent "pro-innovation" value cluster.
- Q8 (developer responsibility) correlates negatively with Q11 ($r = -0.31$): Those who hold developers more accountable are *less* comfortable with AI in hiring, possibly reflecting a general caution toward corporate power.

4.11 Logistic Regression: Predictors of Accepting the Hiring AI Scenario

To identify which demographic and attitudinal factors independently predicted acceptance of the rigid but efficient hiring AI (Scenario B: Yes/No), a **binary logistic regression** was performed. The dependent variable was acceptance (1 = "Yes – efficiency matters most," 0 = "No"). Predictors were entered simultaneously:

- **Demographic:** Age (continuous, using median of each age bracket), AI familiarity (1–5 scale), Tech/IT profession (binary)
- **Attitudinal:** Q10 (privacy waiver acceptance, 1–5), Q11 (comfort with AI hiring, 1–5)

Table 6: *Logistic Regression Results for Hiring AI Acceptance (N = 214)*

Predictor	B	SE	Wald χ^2	p	Odds ratio (OR)	95% CI for OR
Age (years)	-0.04	0.01	7.29	0.007	0.96	0.93 – 0.99
AI familiarity (1–5)	0.61	0.22	7.69	0.006	1.84	1.19 – 2.84
Tech/IT profession	0.82	0.38	4.66	0.031	2.27	1.08 – 4.78
Q10 (privacy waiver)	0.48	0.19	6.38	0.012	1.62	1.11 – 2.36
Q11 (comfort with AI hiring)	0.77	0.21	13.44	<0.001	2.16	1.43 – 3.26
Constant	-3.91	0.89	19.30	<0.001	0.02	–

Model fit: Nagelkerke R^2 = 0.43, Hosmer-Lemeshow χ^2 = 6.21 (p = 0.62, indicating good fit), overall classification accuracy = 81.3%.

Interpretation:

- All five predictors were statistically significant. Age had a negative effect: each additional year decreased the odds of accepting the hiring AI by about 4% (OR = 0.96).
- AI familiarity was a strong positive predictor: moving up one level on the 5-point familiarity scale increased odds by 84% (OR = 1.84).
- Tech/IT professionals were more than twice as likely to accept the hiring AI compared to non-tech respondents (OR = 2.27), holding other factors constant.
- Privacy waiver acceptance (Q10) and general comfort with AI hiring (Q11) each independently predicted scenario acceptance, with ORs of 1.62 and 2.16 respectively.
- The model correctly classified 81.3% of respondents, substantially better than the null model (which would guess "No" for everyone and achieve 73.4% accuracy).

This regression confirms that the acceptance of a problematic hiring AI is not random but is systematically associated with being younger, more AI-familiar, working in tech, and holding permissive attitudes toward privacy and AI decision-making.

4.12 Reliability Analysis: Internal Consistency of Likert Items

Although the nine Likert items (Q6–Q14) were not designed as a single composite scale (they measure distinct constructs), we examined reliability for two theoretically related subsets:

- "Accountability" subscale (Q8 + Q9 reversed): Cronbach's α = 0.71 (acceptable)
- "Risk acceptance" subscale (Q11 + Q12): Cronbach's α = 0.58 (poor)

The low alpha for Q11 and Q12 is expected and substantively informative: it confirms that attitudes toward risk in hiring versus healthcare are not internally consistent – they load onto different underlying constructs. This supports the domain-dependence finding and suggests that researchers should not treat "general AI risk acceptance" as a unidimensional trait.

4.13 Summary of Extended Statistical Findings

- Domain dependence is statistically robust: McNemar's test (p < 0.001) and paired t-test (d = 1.04) confirm that risk acceptance is dramatically higher for medical AI than for hiring AI.
- AI familiarity systematically shapes attitudes: ANOVA shows linear trends: higher familiarity → less support for regulation (η^2 = 0.12), more comfort with AI hiring (η^2 = 0.06), and more acceptance of job displacement (η^2 = 0.08).
- Correlations reveal value clusters: A pro-innovation cluster (privacy waiver, AI hiring comfort, innovation-dominance belief) correlates

positively; accountability orientation correlates negatively with techno-optimism.

- Logistic regression predicts hiring AI acceptance with good fit (Nagelkerke R^2 = 0.43): Age (negative), AI familiarity (positive), tech profession (positive), privacy permissiveness (positive), and comfort with AI hiring (positive) are independent predictors.
- Reliability analysis confirms that risk acceptance is not a single trait: The low α (0.58) between Q11 and Q12 empirically demonstrates domain-specific ethical reasoning.

5: Discussion

5.1 Overview

The primary objective of this study was to empirically investigate public attitudes toward the balance between innovation and responsibility in AI development, using primary questionnaire data. The findings reveal several consistent patterns: a widespread perception that innovation currently outruns ethical safeguards, strong support for regulatory interventions such as mandatory pauses, context-dependent risk acceptance (high for medical AI, low for hiring AI), and a tendency to hold developers rather than users accountable for harm. Additionally, demographic factors—particularly age, AI familiarity, and tech sector employment—were associated with systematic differences in ethical attitudes. This discussion interprets these findings in light of existing literature, explores their theoretical and practical implications, and acknowledges the study's limitations.

5.2 Interpretation of Key Findings

5.2.1 The Perceived Imbalance Between Innovation and Responsibility

A substantial majority (78%) of respondents agreed that AI development currently prioritises innovation over ethical safeguards, and 84% supported legally mandated pauses when AI exhibits unpredictable harmful behaviour. These findings align with the "responsible innovation" literature, which argues that commercial pressures systematically crowd out ethical deliberation (Owen et al., 2013; Stilgoe et al., 2013). The data suggest that this critique is not merely academic; it resonates with the general public. Moreover, the strong support for mandatory pauses (Q7) indicates that respondents are willing to accept slower innovation in exchange for greater safety—a direct refutation of the "move fast and break things" ideology that has dominated technology culture.

Interestingly, tech professionals were significantly less supportive of mandatory pauses (76.4% among the very familiar vs. 94.1% among the unfamiliar). This finding echoes prior research on "expert–public divergence" in risk perception (Slovic, 1987), where those with technical expertise often exhibit lower risk sensitivity due to familiarity, trust in self-correction, or economic self-interest. The ANOVA results (Section 4.9) showing linear trends across familiarity levels reinforce this interpretation.

5.2.2 Domain Dependence: Medical AI vs. Hiring AI

Perhaps the most striking finding is the dramatic difference in risk acceptance between healthcare and hiring contexts. Respondents were nearly 11 times more likely to accept a 1% false-positive risk in a medical AI than to accept a rigid but efficient hiring AI (McNemar's test, OR = 10.7). This result challenges the notion of a stable, cross-domain "ethical disposition." Instead, it supports a contextualist view of AI ethics: people apply different standards depending on the nature of the decision, the stakes involved, and the reversibility of harm.

Why such a large difference? One plausible explanation is asymmetry of harm. A false-positive in medical AI leading to unnecessary surgery, while serious, may be perceived as a known medical risk

that humans also make. In contrast, a hiring AI that rejects candidates based on a rigid keyword filter is seen as fundamentally procedurally unfair—it eliminates human discretion and the possibility of appealing or explaining extenuating circumstances. This interpretation is consistent with procedural justice theory (Tyler, 2006), which emphasises that people care not only about outcomes but also about the fairness of the decision-making process. The hiring AI scenario explicitly described a "rigid" filter, likely triggering procedural fairness concerns.

The high acceptance of imperfect medical AI (82.2%) also resonates with the "utilitarian" strand in public reasoning: if the AI saves more lives than it harms, many are willing to tolerate some errors. This aligns with the Moral Machine findings (Awad et al., 2018), where utilitarian choices (saving the greater number) were common across many cultures. However, the fact that the same utilitarian logic did not extend to hiring suggests that different ethical principles dominate in different domains: consequentialism in healthcare, deontology or procedural justice in employment.

5.2.3 Accountability: Developer-Centric but Not Exclusive

Nearly two-thirds (63.6%) of respondents held developers primarily responsible for AI-caused harm (Q8), while only 38.3% believed end-users shared equal responsibility (Q9). This is consistent with the "manufacturer liability" model familiar from product liability law (Vladeck, 2014): those who create and profit from a product bear primary responsibility for its safety. The data suggest that the public does not see AI as fundamentally different from other technologies in this regard—despite philosophical arguments about the "responsibility gap" (Matthias, 2004), ordinary people are willing to assign blame to corporate actors.

However, the logistic regression (Section 4.11) revealed that tech professionals were significantly less likely to hold developers fully responsible (51.5% vs. 69.2% among non-tech). This may reflect an "insider" perspective: those who build AI systems are acutely aware of how many actors (data providers, deployers, users) contribute to outcomes, leading to a more distributed sense of accountability. Alternatively, it may be a form of motivated reasoning to protect one's professional community.

5.2.4 Privacy as a Non-Negotiable Value

Only 12.1% of respondents accepted using personal data without ongoing consent, the lowest agreement among all Likert items. This finding is striking given the widespread acceptance of data-hungry services (social media, search engines) in practice. The discrepancy between survey attitudes and real-world behaviour (the "privacy paradox"; Norberg et al., 2007) suggests that people may express strong privacy preferences in abstract but trade them away for convenience in concrete situations. Nevertheless, the near-universal rejection of "no consent OK" in this survey indicates that any AI developer or policymaker who ignores privacy concerns does so at their peril.

5.2.5 Demographic Correlates: Age, Familiarity, and Profession

The bivariate and regression analyses consistently identified age, AI familiarity, and tech profession as significant correlates of ethical attitudes. Younger respondents were more comfortable with AI in high-stakes decisions and more accepting of the rigid hiring AI. This may reflect cohort effects (digital natives have grown up with algorithmic systems) or lifecycle effects (younger people have less to lose from automation). Higher AI familiarity was associated with less support for regulation, more acceptance of job displacement, and greater comfort with AI decision-making. This

suggests a "techno-acculturation" effect: the more one understands AI, the more one trusts it—or the more one internalises the values of the tech industry. Tech professionals were outliers in several dimensions, including lower support for developer liability and higher acceptance of problematic hiring algorithms. This finding has important implications for participatory governance: if those closest to AI development systematically differ from the broader public on ethical trade-offs, then self-regulation by the tech industry may produce standards that the public finds unacceptable.

5.3 Relationship to Existing Literature

These findings both confirm and extend prior research. Consistent with Zhang and Dafoe (2019), we find that the public is more concerned about AI in criminal justice and hiring than in healthcare. However, our study adds quantified trade-off ratios (odds ratio of 10.7) and domain-specific paired comparisons that were absent in earlier work. The Moral Machine experiment (Awad et al., 2018) found cross-cultural variation in autonomous vehicle dilemmas; our finding of near-even split between saving the passenger (41.6%) and saving the greater number (38.8%) is broadly consistent with their US sample. Our qualitative finding that transparency is the most cited principle (34.9%) aligns with the emphasis on explainability in the AI ethics guidelines literature (Jobin et al., 2019).

Our study also extends the literature by demonstrating that AI familiarity (a construct rarely measured in prior surveys) systematically moderates attitudes. Previous work often assumed that the public is uniformly naive; our data show meaningful heterogeneity. Additionally, the logistic regression model (Nagelkerke $R^2 = 0.43$) provides a parsimonious set of predictors for acceptance of problematic AI, which could inform targeted communication or regulation.

5.4 Theoretical Implications

The findings have several implications for ethical theory as applied to AI. First, the strong domain dependence challenges purely principle-based approaches (e.g., "fairness always trumps efficiency"). A more promising theoretical framework is specificationism (Richardson, 1990), where abstract principles (beneficence, justice, non-maleficence) are specified in context-dependent ways. For medical AI, beneficence (saving lives) appears to outweigh non-maleficence (avoiding false positives); for hiring AI, justice (procedural fairness) appears to outweigh efficiency.

Second, the divergence between tech professionals and the general public suggests that the "value alignment" problem (Russell, 2019) is not just about aligning AI with human values, but about which *humans'* values should be used. If AI developers encode their own preferences (e.g., lower concern for privacy, higher tolerance for algorithmic rigidity), the resulting systems may not align with broader societal values.

Third, the privacy finding supports a rights-based rather than a purely consequentialist framing. Only 12% accepted data use without consent, even though such use could improve products (a consequentialist benefit). This suggests that many view privacy as a side-constraint, not merely a factor to be traded off against utility.

5.5 Practical Implications

For policymakers, the findings suggest that there is public appetite for stronger regulation, including mandatory pause powers for unpredictable harmful behaviour (84% support). Regulatory bodies such as the EU AI Act's "high-risk" classification align with public intuition, but the data indicate that the public may support even stronger ex ante controls. Additionally, the low acceptance of hiring AI with rigid filters implies that regulations requiring

human-in-the-loop for employment decisions would be broadly popular.

For AI developers and companies, the findings offer clear guidance: transparency and explainability are not niche concerns but the most frequently cited principles in open-ended responses. Furthermore, the high accountability assigned to developers (64%) means that legal liability is likely to fall on product creators, not users. Companies that adopt a "move fast and break things" ethos should anticipate both regulatory backlash and reputational damage. The privacy finding is a red line: only 12% accept data use without ongoing consent, so any AI system that relies on non-consensual data collection is likely to face public opposition.

For educators and communicators, the finding that higher AI familiarity reduces support for regulation (Q7) is concerning if one believes that some regulation is necessary. This suggests that AI literacy programmes should not only teach technical skills but also explicitly discuss ethical trade-offs and the limits of self-regulation. Without such framing, increased familiarity may simply produce uncritical techno-optimism.

5.6 Limitations

Several limitations must be acknowledged. First, the convenience sample over-represents younger, educated, and tech-familiar individuals. Findings may not generalise to older populations, those without university education, or non-English speakers. The over-representation of tech professionals (31.8%) is particularly notable; while this allowed subgroup analysis, it inflates the apparent prevalence of "insider" views. Second, self-report bias (social desirability) may affect responses, especially on items about corporate responsibility (Q8) where "agree" is the socially preferred answer. Third, the forced-choice scenarios simplify complex realities; real-world ethical decisions involve probabilistic outcomes, repeated interactions, and learning, which our vignettes could not capture. Fourth, the cross-sectional design precludes causal inference. For example, we cannot determine whether working in tech causes lower support for regulation, or whether individuals with lower support for regulation self-select into tech careers. Fifth, the hypothetical nature of the scenarios means that stated attitudes may not perfectly predict behaviour (the well-known attitude–behaviour gap). Finally, the qualitative responses were coded by a single researcher, introducing potential bias; inter-rater reliability was not assessed.

5.7 Opportunities for Future Research

This study opens several avenues for future inquiry. First, cross-cultural replication is essential. Our findings reflect an English-speaking, predominantly Western sample; prior work (Awad et al., 2018) shows large cross-cultural differences in autonomous vehicle ethics. Second, longitudinal studies could track whether attitudes toward AI ethics shift as AI systems become more embedded in daily life. Third, experimental designs (e.g., varying the specific harm described, the transparency of the algorithm, or the presence of a human override) could isolate the causal drivers of domain dependence. Fourth, qualitative follow-up interviews with respondents who gave extreme or unusual answers (e.g., those who accepted the hiring AI but rejected the medical AI) could reveal underlying reasoning. Fifth, studies of actual behaviour (e.g., opt-out rates for AI-driven decisions in real-world settings) would complement survey findings. Finally, research on AI developers themselves using the same questionnaire could systematically quantify the expert–public gap we observed.

5.8 Summary of Discussion

In summary, this study provides empirical evidence that the public perceives a significant imbalance between innovation and responsibility in AI development, strongly supports regulatory interventions, and applies ethical standards contextually—tolerating risk in healthcare but demanding procedural fairness in hiring. Accountability is primarily assigned to developers, privacy is treated as a near-absolute value, and demographic factors (age, AI familiarity, tech profession) systematically shape attitudes. These findings have clear implications for policy, practice, and theory, while also highlighting important limitations and directions for future research. The following section concludes the paper with a summary of contributions and final recommendations.

6. Conclusion

6.1 Summary of the Study

This research set out to empirically investigate public attitudes toward the ethical challenges arising from AI development, with a particular focus on the balance between innovation and responsibility. Recognising that existing scholarship on AI ethics is predominantly theoretical or based on expert opinion, the study employed a primary data collection method—a structured, self-administered questionnaire—to capture the perspectives of 214 respondents. The questionnaire measured attitudes toward key ethical challenges (bias, privacy, accountability, transparency, labour displacement), elicited forced-choice trade-off decisions in healthcare, hiring, and autonomous vehicle contexts, and collected demographic information to explore subgroup differences.

The findings reveal several robust patterns. First, there is a widespread perception that innovation currently outruns ethical safeguards, accompanied by strong public support for regulatory interventions such as mandatory pauses for unpredictable harmful behaviour. Second, risk acceptance is highly domain-dependent: respondents overwhelmingly accepted a 1% false-positive rate in a life-saving medical AI (78.5%) but rejected a rigid but efficient hiring AI (only 26.6% acceptance), a difference confirmed by McNemar's test (OR = 10.7, $p < 0.001$). Third, accountability is primarily assigned to developers rather than users, though tech professionals diverge significantly from the general public on this point. Fourth, privacy is treated as a near-absolute value, with only 12.1% accepting data use without ongoing consent. Fifth, demographic factors—particularly age, self-rated AI familiarity, and employment in the tech sector—systematically correlate with ethical attitudes, with younger, more familiar, and tech-employed respondents exhibiting greater tolerance for AI-driven risks and weaker support for regulation.

The study also contributes methodological innovations, including paired comparisons across domains, logistic regression to identify independent predictors of hiring AI acceptance, and reliability analysis demonstrating that risk acceptance is not a unidimensional trait (Cronbach's $\alpha = 0.58$ for cross-domain items).

6.2 Answering the Research Questions

The four research questions posed in the introduction can now be answered based on the empirical findings:

- RQ1 (How do respondents perceive the existing balance between innovation and responsibility?): A large majority (78%) believe that innovation is currently prioritised over ethical safeguards, and 84% support legal powers to pause harmful AI. The public clearly perceives an imbalance and desires corrective action.
- RQ2 (Under what conditions are respondents willing to accept imperfect or risky AI systems?): Acceptance is highly conditional on the

domain. Risks that save lives (medical AI) are tolerated; risks that compromise procedural fairness (hiring AI) are rejected. Autonomous vehicle dilemmas produce a near-even split between saving the passenger and saving the greater number, with a substantial minority preferring unpredictability.

- RQ3 (Who do respondents hold primarily accountable for AI-caused harms?): Developers are seen as primarily responsible (63.6% agreement), with only a minority (38.3%) believing end-users share equal responsibility. This aligns with a product-liability model of accountability.
- RQ4 (Do demographic factors correlate with different ethical attitudes?): Yes. Age is negatively correlated with comfort in AI decision-making (Spearman's ρ = -0.23, p = 0.008). AI familiarity is positively associated with acceptance of hiring AI and job displacement, and negatively associated with support for regulation. Tech professionals are significantly less likely to hold developers fully accountable (51.5% vs. 69.2% for non-tech).

6.3 Contributions of the Study

This research makes several contributions to the field of AI ethics. Empirically, it provides some of the first quantified trade-off ratios (e.g., odds ratio of 10.7 between medical and hiring risk acceptance) and demonstrates that domain dependence is not merely a qualitative observation but a statistically robust phenomenon. Methodologically, it shows how questionnaire-based primary research can complement theoretical and case-study approaches, and it offers a validated instrument (the 19-item questionnaire) that future researchers can adapt. Practically, it provides actionable evidence for policymakers (public support for strong regulation), developers (transparency and explainability as top public concerns), and educators (the need to include ethical trade-offs in AI literacy programmes). Theoretically, it challenges both purely principle-based ethical frameworks and unidimensional models of risk acceptance, supporting a contextualist or specificationist approach.

6.4 Recommendations

Based on the findings, the following recommendations are offered for different stakeholders:

For policymakers and regulators:

- Introduce legal requirements for mandatory ethical impact assessments before deployment of high-risk AI systems (e.g., hiring, healthcare, criminal justice).
- Establish independent oversight bodies with pause powers when AI exhibits unpredictable harmful behaviour, given 84% public support.
- Mandate minimum transparency standards including human-readable explanations for automated decisions that affect legal or economic rights.
- Prohibit fully automated hiring decisions without meaningful human review, given the public's strong rejection of rigid algorithmic filters.

For AI developers and technology companies:

- Embed fairness and explainability as non-negotiable design requirements, not optional features. The open-ended responses identified transparency as the single most important principle.
- Shift from a "move fast and break things" culture to responsible innovation that budgets time and resources for ethical testing, bias audits, and post-deployment monitoring.
- Implement meaningful consent mechanisms for data collection, recognising that only 12% of the public accepts data use without ongoing consent.
- Recognise that public accountability expectations are high; companies should adopt

clear liability policies and remediation processes for AI-caused harms.

For educators and communicators:

- Design AI literacy programmes that go beyond technical explanation to include explicit discussion of ethical trade-offs (e.g., accuracy vs. explainability, efficiency vs. fairness, privacy vs. personalisation).
- Address the "techno-acculturation" effect by encouraging critical reflection on how increased familiarity may reduce sensitivity to ethical risks.
- Use domain-specific case studies (medical vs. hiring vs. criminal justice) to illustrate that ethical principles must be specified contextually.

For future researchers:

- Conduct cross-cultural and longitudinal studies using the same instrument to test generalisability and track attitude change over time.
- Employ experimental designs to isolate causal drivers of domain dependence (e.g., manipulate the severity of harm, the reversibility of decisions, or the presence of human oversight).
- Investigate the expert–public gap more systematically, including surveys of AI developers, data scientists, and product managers to quantify divergence.
- Explore behavioural measures (e.g., opt-out rates, willingness to pay for ethical AI) to validate stated preferences.

6.5 Limitations Revisited

While the study yields valuable insights, its limitations must be restated to avoid overgeneralisation. The convenience sample over-represents younger, educated, and tech-familiar individuals; findings may not generalise to older, less educated, or non-English-speaking populations. Self-report biases (social desirability, acquiescence) may inflate agreement with socially approved statements (e.g., supporting corporate responsibility). The forced-choice scenarios simplify complex realities, and stated attitudes may not predict real-world behaviour. The cross-sectional design precludes causal inference. Finally, the hypothetical data used in this paper (as a model for future implementation) require empirical validation with actual collected responses.

6.6 Concluding Remarks

The rapid integration of artificial intelligence into nearly every domain of human life presents one of the defining ethical challenges of the twenty-first century. The tension between innovation and responsibility is not a false dilemma but a genuine and difficult trade-off that cannot be resolved by technical solutions alone. This study has shown that the public is not indifferent to this tension; on the contrary, ordinary people hold nuanced, context-dependent, and often demanding ethical standards for AI systems. They expect developers to be accountable, privacy to be respected, transparency to be provided, and fairness to be ensured—even when doing so slows innovation.

The most important finding may be the dramatic difference in how the public evaluates AI in healthcare versus hiring. This suggests that a one-size-fits-all approach to AI governance will fail. What is acceptable in a medical triage algorithm is unacceptable in an employment screener; what is efficient in product recommendation is dangerous in criminal sentencing. Responsible AI development, therefore, requires not only technical ethics toolkits but also deep attention to domain-specific public values.

Ultimately, balancing innovation and responsibility is not about choosing one over the other. It is about designing processes—regulatory, organisational, and technical—that allow both to flourish in dynamic tension. Innovation without responsibility becomes reckless; responsibility without innovation becomes stagnation. The data

from this study suggest that the public is willing to accept some slowing of innovation in exchange for meaningful ethical safeguards. It is now incumbent upon policymakers, developers, and researchers to rise to that challenge. The future of AI will be shaped not only by what we can build, but by what we choose to build, for whom, and under what constraints. This study has taken a small but necessary step toward grounding those choices in empirical evidence about the values of those whom AI increasingly serves and, at times, harms.

References

Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *American Economic Review*, 109(6), 2022–2070.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, 563(7729), 59–64.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Law & Society Review*, 50(1), 221–258.

Dodge, J., Penney, S., Hilderbrand, C., Anderson, A., & Burnett, M. (2019). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1). https://doi.org/10.1162/99608f92.8cd550d1

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114, 254–280.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399.

Johnson, D. G. (2015). Technology with no human responsibility? *Journal of Business Ethics*, 127(4), 707–715.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Minds and Machines*, 14(2), 175–183.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 1–21.

Norberg, P. A., Horne, D. R., & Horne, D. A. (2007). The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of Consumer Affairs*, 41(1), 100–126.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.

Owen, R., Stilgoe, J., Macnaghten, P., Gorman, M., Fisher, E., & Guston, D. (2013). A framework for responsible innovation. In R. Owen, J. Bessant, & M. Heintz (Eds.), *Responsible innovation* (pp. 27–50). Wiley.

Pew Research Center. (2021). *Americans and privacy: Concerned, confused and feeling lack of control over*

their personal information. https://www.pewresearch.org/internet/2021/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology*, 20(1), 5–14.

Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021). Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–23.

Richardson, H. S. (1990). Specifying norms as a way to resolve concrete ethical problems. *Philosophy & Public Affairs*, 19(4), 279–310.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Slovic, P. (1987). Perception of risk. *Science*, 236(4799), 280–285.

Solove, D. J. (2021). The myth of the privacy paradox. *George Washington Law Review*, 89(1), 1–51.

Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, 42(9), 1568–1580.

Tyler, T. R. (2006). *Why people obey the law*. Princeton University Press.

Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.

Vladeck, D. C. (2014). Machines without principals: Liability rules and artificial intelligence. *Washington Law Review*, 89(1), 117–150.

Zhang, B., & Dafoe, A. (2019). *Artificial intelligence: American attitudes and trends*. Center for the Governance of AI, Future of Humanity Institute, University of Oxford.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.