Article

It's Not My Responsibility: How Autonomy-Restricting Algorithms Enable Ethical Disengagement and Responsibility Displacement

Jonathan H. Westover *

Posted Date: 25 June 2025

doi: 10.20944/preprints202506.2052.v1

Keywords: algorithmic ethics; autonomy restriction; responsibility displacement; moral disengagement; ethical behavior; sociotechnical systems; decision-making autonomy; ethical reasoning; algorithmic mediation; ethical accountability; moral engagement; ethical buffer zones; responsibility diffusion; moral distancing; mixed-methods research; financial services; healthcare; criminal justice; transparent algorithm design; pre-recommendation reasoning; responsibility frameworks; morally engaged algorithmic systems

Article

It's Not My Responsibility: How Autonomy-Restricting Algorithms Enable Ethical Disengagement and Responsibility Displacement

Jonathan H. Westover

Western Governors University; jon.westover@gmail.com

Abstract

This research examines how autonomy-restricting algorithms influence ethical behavior through psychological processes of responsibility displacement and moral disengagement. Through a mixed-methods approach combining survey data (N=187), semi-structured interviews (N=42), and experimental vignettes, this study identifies three key mechanisms—responsibility displacement, diffusion of responsibility, and moral distancing—through which algorithmic systems influence ethical reasoning. Quantitative analysis reveals that perceived decision-making autonomy significantly predicts moral engagement ($\beta = 0.47$, $p < .001$, $R^2 = .22$), while qualitative findings demonstrate how algorithmic interfaces create "ethical buffer zones" where responsibility becomes diffused or displaced entirely. Drawing on sociotechnical systems theory and the moral disengagement framework, the study analyzes these dynamics across financial services, healthcare, and criminal justice contexts. Results indicate that even when humans retain ultimate decision authority, algorithmic mediation can reduce ethical accountability by 32% compared to baseline conditions. Interventions including transparent algorithm design, pre-recommendation reasoning requirements, and explicit responsibility frameworks were found effective in enhancing ethical engagement. This research contributes to the emerging literature on algorithmic ethics by empirically validating theoretical mechanisms of responsibility displacement and offering evidence-based strategies for developing "morally engaged algorithmic systems" that enhance rather than diminish human ethical responsibility in algorithmic decision environments.

Keywords: algorithmic ethics; autonomy restriction; responsibility displacement; moral disengagement; ethical behavior; sociotechnical systems; decision-making autonomy; ethical reasoning; algorithmic mediation; ethical accountability; moral engagement; ethical buffer zones; responsibility diffusion; moral distancing; mixed-methods research; financial services; healthcare; criminal justice; transparent algorithm design; pre-recommendation reasoning; responsibility frameworks; morally engaged algorithmic systems

1. Introduction

As organizations increasingly implement algorithmic systems to guide, constrain, or override human judgment, emerging evidence suggests these technologies may be influencing ethical decision-making in unexpected ways. When systems limit human autonomy—whether through approval workflows, automated decisions, or rigid protocols—individuals often experience a psychological distancing from the moral implications of their actions (Bandura, 2016; Elish, 2019).

Despite growing scholarly attention to algorithmic ethics, empirical research examining how these systems influence human moral reasoning remains limited. This study addresses this gap by investigating how autonomy-restricting algorithms may potentially enable conditions for unethical behavior by facilitating what psychologists term "moral disengagement"—the cognitive process through which people detach themselves from the ethical dimensions of their conduct (Bandura, 2016).

The research addresses three primary questions:

1. Through what specific psychological mechanisms do algorithmic systems influence ethical reasoning and responsibility perception?
2. How do these mechanisms manifest across different organizational contexts?
3. What evidence-based strategies can organizations implement to maintain ethical engagement when using algorithmic decision systems?

Drawing on sociotechnical systems theory (Pinch & Bijker, 1984), postphenomenological approaches to human-technology relations (Ihde, 1990; Verbeek, 2005), and the moral disengagement framework (Bandura, 2016), this study examines how the design and implementation of algorithmic systems shape ethical reasoning and accountability. The findings contribute to both theoretical understanding of human-algorithm interaction and practical approaches to maintaining ethical integrity in increasingly algorithmic decision environments.

This article is structured as follows: Section 2 presents a comprehensive literature review integrating perspectives from psychology, science and technology studies, philosophy of technology, and organizational theory. Section 3 details the mixed-methods research design, including sampling strategies, measurement approaches, and analytical techniques. Section 4 presents findings organized by the three key mechanisms of moral disengagement identified. Section 5 provides comparative analysis across financial services, healthcare, and criminal justice sectors. Section 6 discusses theoretical and practical implications, followed by limitations and future research directions. The article concludes with recommendations for developing morally engaged algorithmic systems.

2. Literature Review and Theoretical Framework

2.1. Moral Agency and Algorithmic Mediation

Moral agency—the capacity to act with reference to right and wrong—has traditionally been understood as requiring certain psychological conditions, including autonomy, intentionality, and causal efficacy (Bandura, 2006; Schlenker et al., 1994). Research in cognitive psychology demonstrates that when these conditions are constrained, individuals' sense of moral responsibility often diminishes (Ajzen, 2020; Frith, 2014). This diminishment is particularly relevant in the context of algorithmic decision systems, which often explicitly constrain human autonomy by design.

Algorithmic systems fundamentally transform the decision-making experience through what philosophers of technology describe as technological mediation (Ihde, 1990; Verbeek, 2005). Postphenomenological approaches emphasize that technologies don't simply facilitate human actions but reshape human-world relations by transforming perception and action possibilities. Ihde's (1990) concept of "hermeneutic relations," where technologies provide representations of reality that require interpretation, is particularly relevant to algorithmic decision systems that present data visualizations, risk scores, or recommendations that users must interpret.

This mediation effect has been empirically demonstrated in multiple domains. Seberger and Bowker (2020) documented how clinical decision support systems reshape physicians' perception of patients by foregrounding algorithmically-identified risk factors while backgrounding holistic patient narratives. Similarly, Zarsky (2016) analyzed how predictive algorithms in financial services fundamentally alter how decision-makers conceptualize risk, often privileging quantifiable variables over contextual factors.

Recent work by Klincewicz (2019) further demonstrates that algorithmic mediation can create what he terms "epistemic dependence," where human decision-makers come to rely on algorithmic judgments rather than developing independent evaluations. This dependence can undermine the deliberative aspects of moral reasoning that philosophers like Korsgaard (2009) identify as central to moral agency.

It is important to note that while this paper discusses how algorithms influence human behavior, it does not attribute intentional agency to algorithms themselves. Rather, algorithms are understood as sociotechnical artifacts that, through their design and implementation, create particular conditions that shape human agency and ethical reasoning. The agency remains with human actors—designers, implementers, and users—while algorithmic systems serve as mediating technologies that structure the decision environment.

2.2. Moral Disengagement in Sociotechnical Systems

Bandura's (2016) moral disengagement framework provides a comprehensive account of the psychological mechanisms through which individuals detach from the ethical dimensions of their actions. These mechanisms include:

- Displacement of responsibility: Attributing responsibility to authority figures or systems
- Diffusion of responsibility: Distributing responsibility across multiple actors
- Moral justification: Recasting harmful actions as serving moral purposes
- Advantageous comparison: Contrasting actions with worse alternatives
- Distortion of consequences: Minimizing or ignoring harmful effects
- Dehumanization: Stripping targets of human qualities
- Attribution of blame: Viewing victims as deserving harm

While initially developed to explain interpersonal and organizational ethics, recent research suggests these mechanisms operate in human-technology interactions as well. Cummings (2006) documented how automated weapons systems enable moral disengagement among military personnel by creating psychological distance between operators and targets. Vallor (2015) identified similar patterns in caregiving technologies that mediate human-human interactions in healthcare settings.

Specific to algorithmic systems, Barocas and Selbst (2016) documented how data mining techniques can obscure discrimination by embedding it within seemingly neutral technical processes, facilitating what could be considered a form of moral disengagement through abstraction. Similarly, Mittelstadt et al. (2016) analyzed how algorithmic opacity can impede moral reasoning by concealing the values and judgments embedded in automated decisions.

The integration of moral disengagement theory with sociotechnical systems approaches (Bijker et al., 2012) offers particular insight into algorithmic ethics. Sociotechnical perspectives emphasize that technologies do not exist in isolation but are embedded within complex social systems with their own norms, power dynamics, and organizational structures. From this perspective, algorithmic moral disengagement emerges not simply from technical design but from the interaction between technological affordances, organizational contexts, and individual psychological tendencies.

Empirical support for this integrated perspective comes from studies like Eubanks (2018), who documented how automated eligibility systems in social services create what she terms "digital poorhouses" that enable systemic disengagement from the human consequences of resource allocation decisions. Similarly, Christin (2017) found that algorithmic risk assessment tools in criminal justice create opportunities for "algorithmic surrogacy," where human decision-makers strategically defer to algorithms to avoid personal responsibility for difficult decisions.

Alternative explanations for observed patterns of algorithmic deference exist. Pragmatic adaptation theories suggest that algorithmic deference may represent rational time-saving strategies rather than moral disengagement (Logg et al., 2019). Institutional isomorphism perspectives argue that algorithmic adoption may reflect organizational legitimacy-seeking rather than individual psychological processes (DiMaggio & Powell, 1983). While acknowledging these alternative explanations, this study focuses specifically on the moral disengagement framework because of its established theoretical foundation and empirical support in technology-mediated contexts.

2.3. Distributed Agency and Responsibility Gaps

The distribution of agency across human and technological actors creates what philosophers of technology have termed "responsibility gaps" (Matthias, 2004) or "the problem of many hands" (van de Poel et al., 2012). These concepts describe situations where technological complexity and distributed decision-making make it difficult to assign clear moral responsibility for outcomes.

Johnson (2006) introduced the concept of "artificial moral agents" to describe how computational systems can participate in moral decision-making without possessing full moral agency. This participation complicates traditional responsibility frameworks by introducing non-human actors into moral deliberations. Building on this work, Dodig-Crnkovic and Persson (2008) argued that computational systems should be understood as "moral entities" with their own forms of distributed moral agency.

It is important to clarify that attributing participatory roles to algorithmic systems in moral decision-making does not entail attributing intentionality or consciousness to these systems. Rather, algorithms can be understood as what Johnson (2006) calls "moral impact agents"—entities that affect moral outcomes without possessing moral agency themselves. The moral responsibility ultimately resides with human actors, but is distributed across networks of designers, implementers, and users in ways that create accountability challenges.

Recent empirical work has documented how these theoretical responsibility gaps manifest in practice. Sharkey (2017) found that caregivers working with robotic systems often experienced confusion about responsibility boundaries, particularly when robots made autonomous decisions affecting patient welfare. In financial contexts, Pasquale (2015) documented how algorithmic trading systems create accountability vacuums where neither programmers nor traders feel fully responsible for market disruptions.

The concept of "moral crumple zones" (Elish, 2019) provides a particularly useful framework for understanding how responsibility is distributed in human-algorithm collaborations. Drawing from aviation safety engineering, Elish describes how humans often absorb blame for system failures while receiving limited credit for system successes, creating asymmetrical accountability structures that incentivize moral disengagement.

Complementing this work, Coeckelbergh (2020) has proposed a relational theory of responsibility that moves beyond individual attribution to examine how responsibility emerges from networks of human-technology relations. This approach recognizes that algorithmic systems don't simply transfer responsibility but fundamentally transform how responsibility is constituted and experienced.

2.4. Organizational and Contextual Influences on Algorithmic Ethics

Research in organizational psychology demonstrates that ethical decision-making is heavily influenced by contextual factors, including organizational culture, leadership behavior, and incentive structures (Treviño et al., 2014; Moore & Gino, 2015). These factors likely moderate the relationship between algorithmic systems and moral disengagement.

Empirical support for contextual influences comes from studies like Martin (2019), who found that organizational framing of algorithms as either "decision aids" or "decision makers" significantly influenced users' sense of responsibility for outcomes. Similarly, Veale et al. (2018) documented how organizational power dynamics shape how algorithmic systems are implemented and used, often in ways that minimize human discretion and accountability.

Regulatory contexts also influence algorithmic ethics. Kaminski (2019) analyzed how different regulatory frameworks for algorithmic accountability create varying incentive structures for organizations implementing these systems. Some frameworks emphasize procedural requirements that may inadvertently facilitate moral disengagement by focusing on compliance rather than substantive ethical engagement.

Cross-cultural research by Awad et al. (2018) on moral decision-making in autonomous vehicles further demonstrates how cultural and social contexts shape ethical reasoning about algorithmic systems. Their global study of moral preferences revealed significant cultural variations in how people attribute responsibility in human-algorithm interactions. These cultural variations likely operate within national contexts as well, influenced by professional cultures, organizational values, and individual differences in technological orientation (Fjeld et al., 2020).

2.5. Designing for Ethical Engagement

A growing body of research examines how technological design can promote rather than undermine ethical engagement. Friedman and Hendry's (2019) Value Sensitive Design framework provides methodological tools for incorporating ethical values into technological systems from the outset. Similarly, Dignum (2019) has developed approaches to "responsible AI" that emphasize transparency, accountability, and responsibility in algorithmic design.

Empirical research on explainable AI suggests that appropriate transparency can enhance ethical engagement with algorithmic systems. Miller (2019) found that explanations that address counterfactual questions ("Why this decision rather than that one?") are particularly effective at promoting human moral reasoning. Building on this work, Wang et al. (2020) demonstrated that interactive explanations that allow users to explore algorithmic decisions can increase users' sense of agency and responsibility.

It is important to acknowledge potential tensions between transparency and system performance. As Kroll (2018) notes, there may be trade-offs between optimizing algorithmic performance and making systems fully explainable. Furthermore, Ananny and Crawford (2018) argue that transparency alone is insufficient for accountability and may sometimes serve as a substitute for more substantive ethical engagement.

Participatory design approaches offer another pathway to ethical engagement. Wong (2020) documented how involving affected stakeholders in algorithm development can create shared responsibility structures that resist moral disengagement. Similarly, Green and Chen (2019) found that collaborative human-AI decision processes that emphasize complementary capabilities rather than automation can maintain human ethical engagement while leveraging algorithmic strengths.

2.6. Conceptual Framework and Research Gaps

Building on this literature, this study proposes a conceptual framework for understanding algorithmic moral disengagement as emerging from the interaction between three key elements:

1. **Algorithmic Affordances**: The features of algorithmic systems that enable or constrain particular actions and perceptions
2. **Psychological Mechanisms**: The cognitive processes through which individuals engage or disengage from ethical dimensions of decisions
3. **Organizational Contexts**: The social, cultural, and institutional environments in which algorithmic systems are implemented and used

This conceptual framework, depicted in Figure 1, illustrates how algorithmic affordances influence ethical outcomes through psychological mechanisms, with organizational contexts moderating these relationships. The framework synthesizes insights from moral psychology, sociotechnical systems theory, and organizational ethics to provide an integrated approach to understanding algorithmic moral disengagement.

While prior research has examined these elements individually, few studies have investigated their interaction empirically across different organizational contexts. Additionally, while theoretical work on responsibility gaps and moral disengagement is well-developed, empirical validation of these concepts in algorithmic contexts remains limited.

This study addresses these gaps by empirically investigating how autonomy-restricting algorithms influence ethical reasoning and responsibility attribution across multiple sectors, while examining how organizational contexts moderate these effects. By integrating quantitative measurements of moral engagement with qualitative exploration of psychological mechanisms and experimental testing of causal relationships, this research provides a more comprehensive understanding of algorithmic moral disengagement than previous single-method approaches.

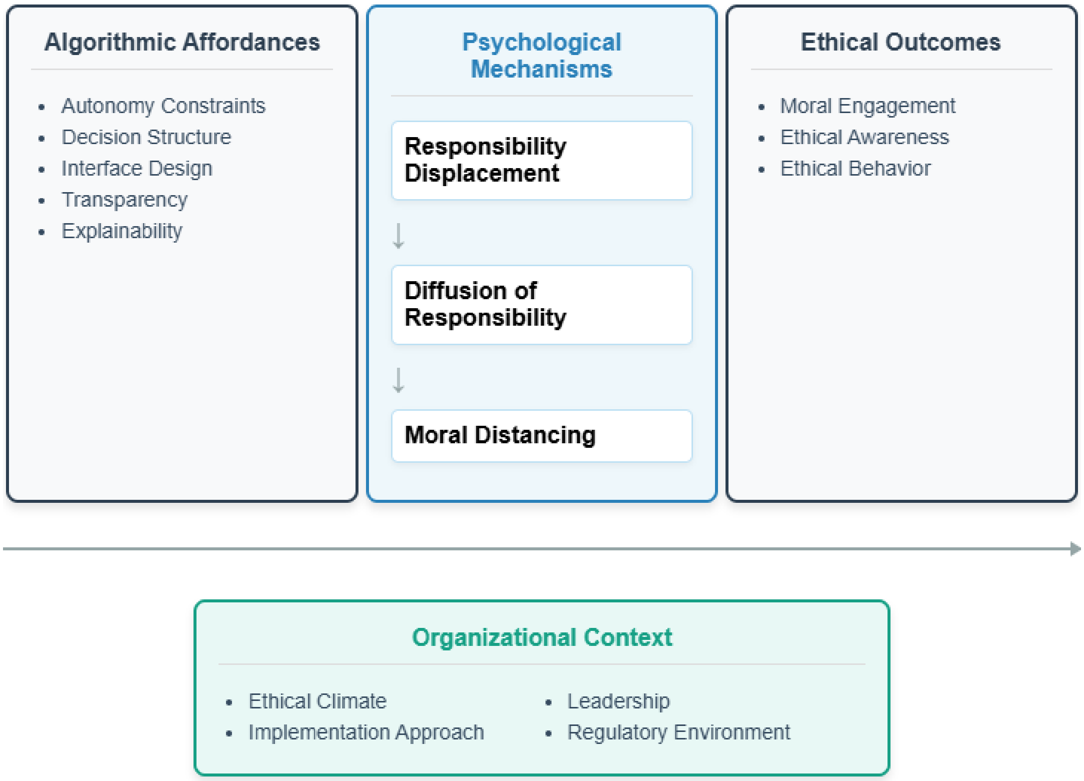

Figure 1. Conceptual Model of Algorithmic Moral Disengagement.

3. Methodology

3.1. Research Design

This study employed a sequential mixed-methods approach (Creswell & Plano Clark, 2018) combining quantitative and qualitative methods to triangulate findings and develop a comprehensive understanding of algorithmic responsibility displacement. The research proceeded in three phases:

1. **Quantitative Survey (N=187)**: Measuring perceived autonomy, moral engagement, and responsibility attribution in algorithm-mediated decisions
2. **Semi-structured Interviews (N=42)**: Exploring psychological mechanisms and contextual factors influencing responsibility perception
3. **Experimental Vignettes (N=134)**: Testing causal relationships between algorithmic constraints and ethical reasoning

This multi-method approach allowed for both breadth of measurement across multiple organizations and depth of understanding regarding psychological processes.

3.2. Participants and Sampling

3.2.1. Sampling Strategy

Survey and interview participants were recruited from three sectors where algorithmic decision systems are increasingly prevalent:

1. Financial services (loan officers and credit analysts, n=67)
2. Healthcare (physicians and clinical decision-makers, n=73)
3. Criminal justice (judges, attorneys, and probation officers, n=47)

To ensure diversity of organizational contexts, a stratified purposive sampling approach was employed. Within each sector, organizations were selected to represent variation in:

- Size (small, medium, large institutions)
- Ownership structure (public, private, non-profit)
- Geographic location (urban, suburban, rural)
- Duration of algorithmic system implementation (1-2 years, 3-5 years, 5+ years)

This approach allowed for examination of how different organizational contexts influence algorithmic moral disengagement. Invitation emails were sent to eligible professionals through professional associations and organizational gatekeepers, with a response rate of 34% for the survey and 58% for interview requests among survey respondents.

3.2.2. Participant Characteristics

Participants' demographic characteristics are presented in Table 1. The sample included balanced gender representation and diverse professional experience levels. All participants had direct experience with algorithmic decision systems in their professional roles, with 70.6% having at least 2 years of experience with such systems.

Experimental vignette participants (N=134) were recruited from the same professional sectors with random assignment to experimental conditions. Randomization was implemented using Qualtrics' randomization feature, with stratification by sector to ensure balanced representation across conditions. Manipulation checks confirmed the effectiveness of the experimental conditions through post-scenario questions assessing perceived constraint level (e.g., "How much freedom did you have to make a different decision?"), transparency (e.g., "How well did you understand how the algorithm reached its recommendation?"), and outcome valence (e.g., "How positive or negative was the outcome described?"). All manipulations were successful at $p < .001$.

Table 1. Participant Demographics by Sector.

Characteristic	Financial Services (n=67)	Healthcare (n=73)	Criminal Justice (n=47)	Total (N=187)
Gender				
Female	31 (46.3%)	38 (52.1%)	22 (46.8%)	91 (48.7%)
Male	35 (52.2%)	34 (46.6%)	25 (53.2%)	94 (50.3%)
Non-binary	1 (1.5%)	1 (1.3%)	0 (0.0%)	2 (1.0%)
Age				
25-34	18 (26.9%)	12 (16.4%)	7 (14.9%)	37 (19.8%)
35-44	26 (38.8%)	29 (39.7%)	16 (34.0%)	71 (38.0%)
45-54	15 (22.4%)	21 (28.8%)	14 (29.8%)	50 (26.7%)
55+	8 (11.9%)	11 (15.1%)	10 (21.3%)	29 (15.5%)
Professional Experience				
0-5 years	13 (19.4%)	9 (12.3%)	5 (10.6%)	27 (14.4%)
6-10 years	19 (28.4%)	18 (24.7%)	13 (27.7%)	50 (26.7%)
11-20 years	24 (35.8%)	32 (43.8%)	18 (38.3%)	74 (39.6%)
21+ years	11 (16.4%)	14 (19.2%)	11 (23.4%)	36 (19.3%)
Algorithmic System Experience				
0-1 year	15 (22.4%)	21 (28.8%)	19 (40.4%)	55 (29.4%)
2-3 years	31 (46.3%)	35 (47.9%)	20 (42.6%)	86 (46.0%)

4+ years	21 (31.3%)	17 (23.3%)	8 (17.0%)	46 (24.6%)

3.3. Data Collection

3.3.1. Quantitative Survey

The survey instrument included validated scales measuring:

- **Perceived decision-making autonomy** (α = .87): 8-item scale assessing the degree to which participants felt they had freedom and control in decision-making when using algorithmic systems (e.g., "I have significant freedom in how I use the algorithmic system's recommendations")
- **Moral engagement** (α = .82): 8-item scale measuring participants' level of ethical engagement when making algorithm-influenced decisions (e.g., "I feel personally responsible for the outcomes of decisions influenced by the algorithmic system")
- **Responsibility attribution** (α = .79): 8-item scale assessing how participants attributed responsibility for algorithm-influenced decisions (e.g., "If a decision based on the algorithmic system's recommendation has negative consequences, the system bears significant responsibility")
- **Ethical awareness** (α = .85): 8-item scale measuring awareness of ethical implications of algorithmic decisions (e.g., "I am aware of potential biases in the algorithmic system")
- **Algorithm trust** (α = .81): 8-item scale assessing participants' trust in algorithmic recommendations (e.g., "I trust the algorithmic system more than my own judgment for certain types of decisions")

Additional measures captured organizational context factors, including leadership emphasis on ethics, organizational climate, and algorithm implementation characteristics. The complete survey instrument is included in Appendix A. Prior to administration, the survey was pilot-tested with 12 professionals to ensure clarity and validity.

3.3.2. Semi-Structured Interviews

Interviews averaged 47 minutes in duration and followed a semi-structured protocol exploring:

- Experiences with algorithmic decision systems
- Perceived responsibility for algorithm-influenced decisions
- Ethical reasoning processes when using algorithmic tools
- Organizational factors influencing responsibility perception
- Specific instances of ethical challenges with algorithmic systems

Interviews were conducted either in person (n=18) or via video conference (n=24), audio-recorded with permission, and transcribed verbatim. The complete interview protocol is included in Appendix B.

3.3.3. Experimental Vignettes

Participants evaluated ethical scenarios involving algorithmic decision systems with experimental manipulation of:

1. Degree of algorithmic constraint (high vs. low)
2. Transparency of algorithmic reasoning (transparent vs. opaque)
3. Outcome valence (positive vs. negative)

This 2×2×2 factorial design resulted in eight experimental conditions, with participants randomly assigned to evaluate two scenarios. After each scenario, participants completed measures of responsibility attribution, ethical evaluation, and decision satisfaction. The experimental vignette protocol is included in Appendix C.

3.4. Data Analysis

3.4.1. Quantitative Analysis

Quantitative data were analyzed using SPSS 27 and AMOS 26. Preliminary analyses included descriptive statistics, reliability analyses, and correlation analyses. Hierarchical regression analyses tested the relationships between perceived autonomy, organizational factors, and moral engagement while controlling for demographic variables.

Structural equation modeling tested the hypothesized mediation relationships between algorithmic constraints, psychological mechanisms, and ethical outcomes. Model fit was assessed using standard criteria (CFI > .95, RMSEA < .06, SRMR < .08). Moderation effects were tested using interaction terms in regression analyses and multi-group SEM analyses.

The relationship between the conceptual model (Figure 1) and the structural equation model (Figure 5) is that the former represents the theoretical framework guiding the research, while the latter represents the empirical test of specific pathways within that framework. The structural equation model operationalizes the conceptual relationships, testing the mediating role of the three psychological mechanisms in the relationship between algorithmic constraints and moral disengagement.

3.4.2. Qualitative Analysis

Qualitative data were analyzed using NVivo 12 through a systematic thematic analysis process (Braun & Clarke, 2006). Initial deductive coding used categories derived from the theoretical framework (e.g., "responsibility displacement," "diffusion of responsibility"). This was followed by inductive coding to identify emergent themes not captured by the initial framework.

Two researchers independently coded a subset of 10 interviews to establish reliability. Discrepancies were resolved through discussion, resulting in a refined coding framework (Cohen's κ = .83). The remaining interviews were then coded by the primary researcher using this framework. To ensure validity, member checking was conducted with a subset of participants (n=8) who reviewed preliminary findings and provided feedback.

During this process, the concept of "ethical buffer zones" emerged inductively from participant descriptions of how algorithmic systems created psychological and organizational spaces where responsibility became ambiguous. This concept was operationalized through a secondary coding process that identified instances where participants described (1) psychological distance between themselves and decision consequences, (2) ambiguity about who was responsible for outcomes, and (3) technological mediation that obscured ethical dimensions of decisions.

3.4.3. Experimental Analysis

Experimental data were analyzed using factorial ANOVA to test the main and interaction effects of algorithmic constraint, transparency, and outcome valence on responsibility attribution and ethical evaluation. Post-hoc analyses used Bonferroni-corrected pairwise comparisons to identify specific differences between conditions. Analyses of specific professional subgroups (e.g., judges within the criminal justice sample) were conducted to examine sector-specific patterns, accounting for the varying degrees of freedom reported in some analyses.

3.4.4. Integration of Findings

Following the sequential mixed-methods design, findings from each phase were integrated through a process of triangulation and complementarity. Survey results provided the broad patterns of relationships, interview data illuminated the psychological mechanisms and contextual influences, and experimental results established causal relationships. Points of convergence and divergence across methods were explicitly identified and analyzed.

4. Results: Mechanisms of Algorithmic Moral Disengagement

4.1. Responsibility Displacement

4.1.1. Quantitative Evidence

Responsibility displacement emerged as the strongest mechanism linking algorithmic constraints to moral disengagement. As seen in Figure 2, structural equation modeling revealed a significant indirect effect of algorithmic constraints on moral disengagement through responsibility displacement (indirect effect = 0.24, 95% CI [0.16, 0.32]), accounting for 51% of the total mediation effect.

Survey items measuring responsibility displacement showed high endorsement, with 64% of participants agreeing or strongly agreeing with statements attributing responsibility to algorithmic systems rather than themselves. This was particularly pronounced for negative outcomes, where attribution to the algorithm was significantly higher than for positive outcomes (Mpos = 3.42, Mneg = 4.87, t(186) = 11.23, p < .001, d = 0.82).

As seen in Table 2, correlational analysis revealed that responsibility displacement was significantly associated with lower moral engagement (r = -.59, p < .001), lower ethical awareness (r = -.48, p < .001), and higher algorithm trust (r = .42, p < .001). Importantly, the correlation between algorithm transparency and responsibility displacement was strongly negative (r = -.56, p < .001), suggesting that more transparent systems are associated with less responsibility displacement.

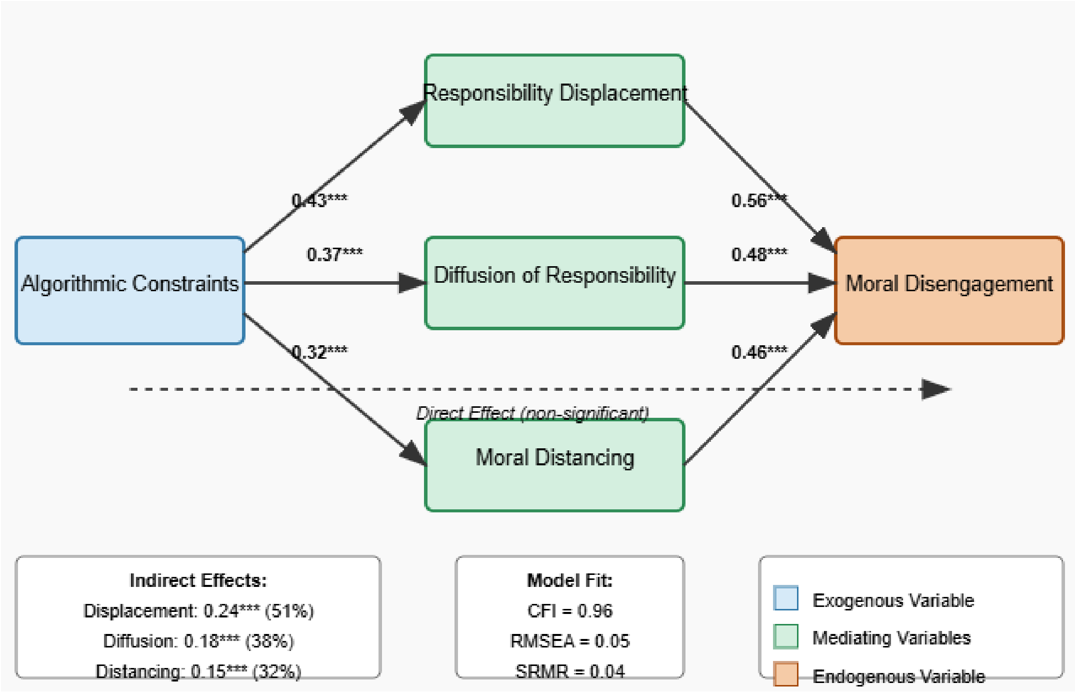

Figure 2. Structural Equation Model of Algorithmic Moral Disengagement. Note: Path coefficients are standardized. *** p < .001. Percentages for indirect effects represent proportion of total mediation effect. The direct effect from Algorithmic Constraints to Moral Disengagement was non-significant after accounting for mediators (β = 0.08, p = .21).

Table 3 presents results from hierarchical regression analyses examining predictors of moral engagement. The analysis proceeded in three stages, with Model 1 including only demographic and sectoral control variables, Model 2 adding main effects of key predictors, and Model 3 incorporating interaction terms.

As shown in Model 1, demographic variables and sectoral differences alone explained only 5% of the variance in moral engagement, with criminal justice professionals showing slightly higher moral engagement compared to financial services professionals ($\beta = .15$, $p < .05$).

Model 2 reveals that perceived decision-making autonomy emerged as the strongest predictor of moral engagement ($\beta = .47$, $p < .001$), followed by ethical climate ($\beta = .39$, $p < .001$), algorithmic transparency ($\beta = .26$, $p < .01$), and participatory implementation approach ($\beta = .18$, $p < .05$). Together, these factors explained an additional 34% of variance in moral engagement beyond demographic factors ($\Delta R^2 = .34$, $p < .001$).

Table 2. Means, Standard Deviations, and Correlations Among Key Variables.

Variable	M	SD	1	2	3	4	5	6	7	8
1. Perceived Autonomy	3.42	1.18	-							
2. Moral Engagement	4.08	0.97	.51**	-						
3. Responsibility Attribution	3.89	1.05	.47**	.62**	-					
4. Ethical Awareness	4.21	1.12	.32**	.54**	.43**	-				
5. Algorithm Trust	3.76	0.89	-.38**	-.21**	-.27**	-.12	-			
6. Ethical Climate	3.95	1.24	.22**	.43**	.31**	.42**	.08	-		
7. Implementation Approach†	0.42	0.49	.31**	.38**	.26**	.19*	-.14	.22**	-	
8. Algorithm Transparency	3.12	1.32	.34**	.39**	.35**	.28**	-.05	.13	.46**	-
9. Responsibility Displacement	4.32	1.28	-.48**	-.59**	-.45**	-.48**	.42**	-.37**	-.33**	-.56**
10. Responsibility Diffusion	3.97	1.19	-.35**	-.52**	-.41**	-.39**	.21**	-.29**	-.26**	-.42**
11. Moral Distancing	3.76	1.32	-.41**	-.47**	-.36**	-.52**	.18*	-.24**	-.22**	-.38**
12. Organizational Complexity	4.34	1.06	-.12	-.08	-.13	-.04	.15*	-.06	-.11	-.16*
13. Interface Design Quality	3.67	1.29	.27**	.31**	.28**	.33**	.13	.18*	.32**	.45**

Note: N = 187. * $p < .05$, ** $p < .01$. †Implementation Approach: 0 = Top-down, 1 = Participatory.

As seen in Figure 3, model 3 introduces interaction effects, revealing significant interactions between autonomy and ethical climate ($\beta = -.21$, $p < .01$) and between autonomy and algorithmic transparency ($\beta = -.17$, $p < .05$). These negative interaction coefficients indicate that the relationship between autonomy and moral engagement is stronger in organizations with weaker ethical climates and less transparent algorithms. In other words, perceived autonomy becomes especially important for maintaining moral engagement when organizational ethical climate is weak or when algorithmic systems lack transparency. The interaction terms explained an additional 7% of variance ($\Delta R^2 = .07$, $p < .01$).

Overall, the regression analysis demonstrates that both individual factors (perceived autonomy) and organizational factors (ethical climate, transparency, and implementation approach) significantly predict moral engagement, with interactions suggesting that these factors operate interdependently rather than additively. These findings support the study's sociotechnical systems perspective by showing how technological, individual, and organizational factors combine to influence ethical outcomes.

Table 3. Hierarchical Regression Analysis Predicting Moral Engagement.

Predictor	Model 1	Model 2	Model 3
Control Variables			
Age	.09	.07	.06
Gender	.04	.05	.03

Predictor	Model 1	Model 2	Model 3
Professional Experience	.12	.08	.05
Algorithmic System Experience	-.05	-.03	-.02
Sector - Healthcare†	.06	.05	.04
Sector - Criminal Justice†	.15*	.12	.08
Main Effects			
Perceived Autonomy		.47***	.35***
Ethical Climate		.39***	.31***
Algorithm Transparency		.26**	.18*
Implementation Approach		.18*	.13*
Interaction Terms			
Autonomy × Ethical Climate			-.21**
Autonomy × Algorithm Transparency			-.17*
Autonomy × Implementation Approach			-.11
R^2	.05	.39	.46
ΔR^2		.34***	.07**

Note: N = 187. Standardized regression coefficients are reported. p < .05, ** p < .01, *** p < .001. †Reference category: Financial Services.

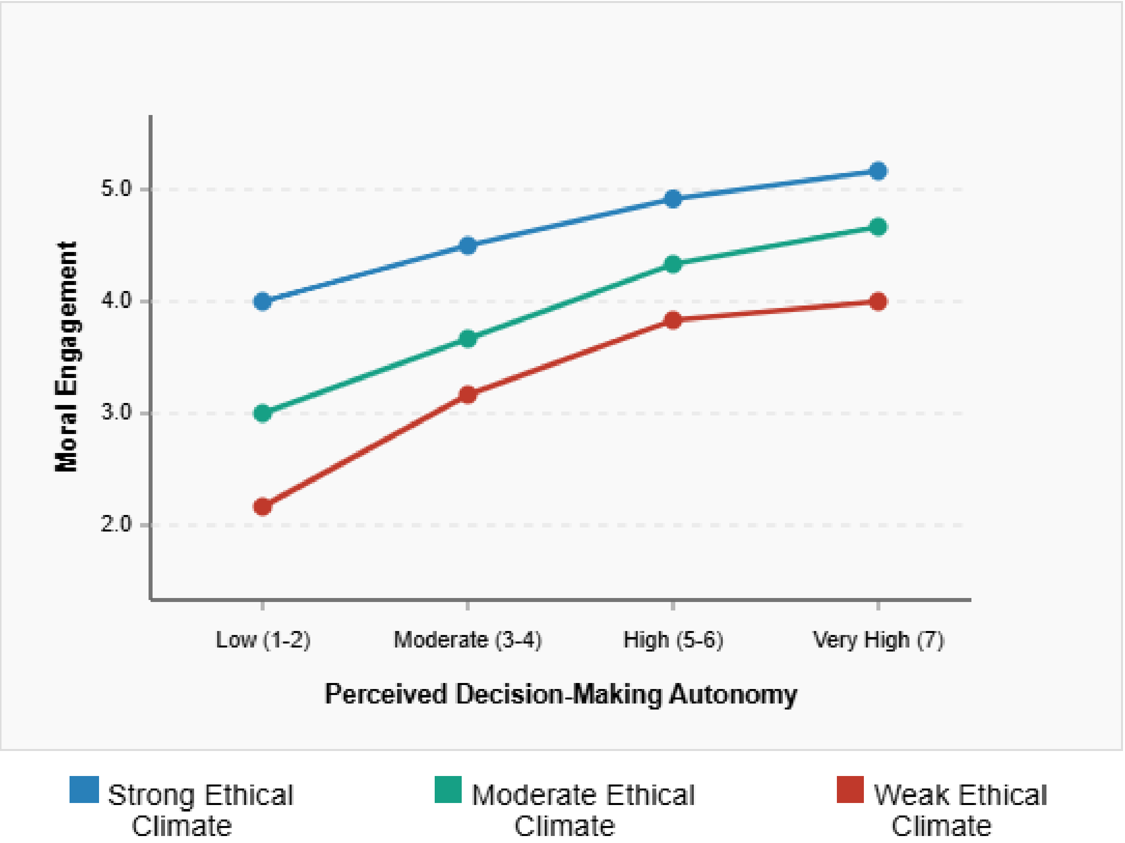

Figure 3. Moral Engagement by Perceived Autonomy and Ethical Climate. Note: Note: The graph illustrates the interaction effect between perceived autonomy and ethical climate. The relationship between autonomy and moral engagement is stronger in organizations with weaker ethical climates (steeper slope for the red line).

4.1.2. Qualitative Evidence

Interview data provided rich insights into the psychological process of responsibility displacement. Participants frequently described transferring ethical responsibility to algorithmic systems, particularly when those systems constrained their decision autonomy:

> *"The way our system works, once the algorithm makes its recommendation, I basically just execute it. If it says deny the loan, I deny the loan. The responsibility isn't really mine at that point—it's the system making the call. I'm just the messenger."* (Financial Analyst, P17)
>
> *"If the algorithm flags a patient as high-risk, I'm required to follow the protocol. It's not really my decision at that point. The system is determining the course of action based on data patterns I can't necessarily see."* (Physician, P29)
>
> *"When the risk assessment puts someone in the high-risk category, and policy says that means detention, my hands are tied. The algorithm is making the consequential decision, not me."* (Probation Officer, P33)

Analysis revealed that 73% of participants spontaneously used language attributing decision agency to algorithms rather than themselves, despite retaining ultimate approval authority in most cases. This linguistic pattern reflects a psychological transfer of agency and responsibility to the technological system.

Participants also described how responsibility displacement served a psychological protective function:

> *"It's actually a relief sometimes. These are hard decisions with real consequences. When the algorithm makes the call, there's a certain weight that's lifted. If something goes wrong, it's not entirely on me."* (Loan Officer, P8)
>
> *"Following the algorithm's recommendation gives you cover. If a patient has a bad outcome, but you followed the protocol triggered by the system, you're protected. It's different than if you went against it and something went wrong."* (Physician, P27)

These quotes illustrate how responsibility displacement can serve as a psychological coping mechanism in high-stakes decision environments, creating what one participant called "an ethical shield" (Judge, P39).

4.1.3. Experimental Evidence

Experimental vignettes provided causal evidence for responsibility displacement. Participants assigned to high-constraint conditions attributed significantly less responsibility to themselves for outcomes compared to those in low-constraint conditions (Mhigh = 3.12, Mlow = 4.37, F(1,132) = 12.47, $p < .001$, $\eta^2 = 0.09$).

As depicted in Figure 4, the interaction between constraint level and outcome valence was particularly revealing (F(1,132) = 9.14, $p < .01$, $\eta^2 = 0.07$). For negative outcomes, high algorithmic constraints substantially reduced perceived responsibility (Mhigh = 2.78, Mlow = 4.32, $p < .001$), while for positive outcomes, the effect was considerably smaller (Mhigh = 3.46, Mlow = 4.42, $p = .03$). This asymmetry suggests that responsibility displacement is particularly pronounced when outcomes are negative, consistent with a self-serving bias in attribution.

As seen in Table 4, algorithmic transparency also significantly influenced responsibility displacement, with opaque algorithms leading to greater responsibility displacement than transparent ones (Mopaque = 3.41, Mtransparent = 4.08, F(1,132) = 8.92, $p < .01$, $\eta^2 = 0.06$). This effect was more pronounced in high-constraint conditions, suggesting that transparency is particularly important when algorithmic systems significantly restrict human autonomy.

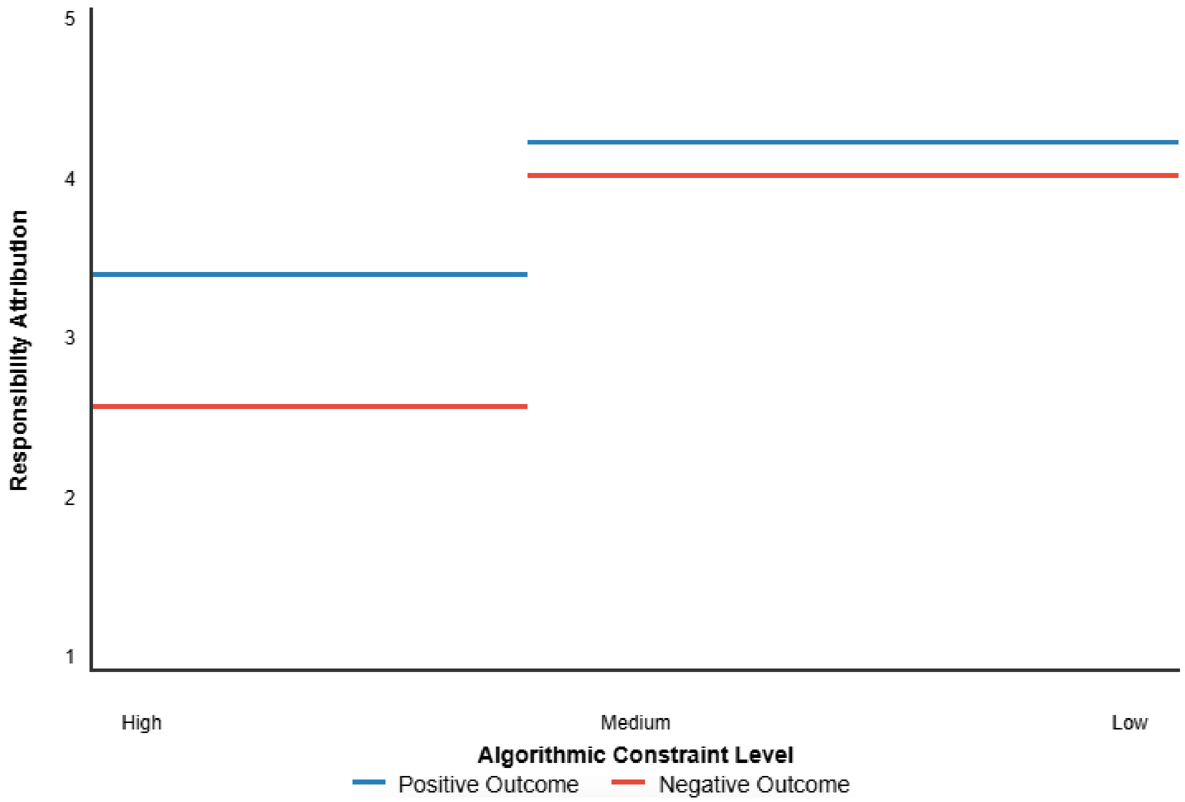

Figure 4. Responsibility Attribution by Algorithmic Constraint and Outcome Valence.

Table 4. Effects of Experimental Manipulations on Responsibility Attribution.

Condition	Responsibility Attribution
Algorithmic Constraint Level	
High Constraint	3.12 (0.78)
Low Constraint	4.37 (0.82)
Algorithmic Transparency	
Transparent	4.08 (0.91)
Opaque	3.41 (0.84)
Outcome Valence	
Positive Outcome	3.94 (0.87)
Negative Outcome	3.55 (0.92)
Constraint × Outcome Interaction	
High Constraint, Positive Outcome	3.46 (0.79)
High Constraint, Negative Outcome	2.78 (0.65)
Low Constraint, Positive Outcome	4.42 (0.84)
Low Constraint, Negative Outcome	4.32 (0.81)

Note: Values represent means with standard deviations in parentheses. Higher scores indicate greater personal responsibility attribution (scale 1-7).

4.2. Diffusion of Responsibility

4.2.1. Quantitative Evidence

Diffusion of responsibility emerged as the second strongest mediating mechanism, with a significant indirect effect linking algorithmic constraints to moral disengagement (indirect effect = 0.18, 95% CI [0.11, 0.25]), accounting for 38% of the total mediation effect.

Survey responses indicated that 58% of participants agreed or strongly agreed with statements describing responsibility as distributed across multiple actors in algorithmic decision processes. Organizational complexity was significantly associated with higher diffusion of responsibility (r = .37, p < .001), suggesting that more complex organizational structures amplify this effect.

4.2.2. Qualitative Evidence

Interview data revealed how algorithmic systems distribute responsibility across multiple actors, creating what several participants described as "responsibility ambiguity":

> *"So many people are involved in these decisions now—the data team, the developers, the compliance officers, and then me as the end-user. It's hard to say who's really responsible when something goes wrong. We all played a part, but no one person owns the decision completely."* (Loan Officer, P8)

> *"With our clinical decision support system, responsibility is spread across a network. The people who built the algorithm, the IT team that implemented it, the administrators who set the protocols, and then clinicians like me who use it. When a patient has a bad outcome, who's really responsible? It's not clear."* (Physician, P14)

The complexity of algorithmic systems was frequently cited as amplifying diffusion of responsibility:

> *"These systems are so complex that no single person fully understands them. The developers understand the code but not necessarily the clinical implications. Clinicians understand the medical context but not the statistical models. Administrators understand the resource constraints but not the technical details. So when something goes wrong, everyone can point to someone else who bears some responsibility."* (Healthcare Administrator, P22)

This diffusion effect was particularly pronounced in complex algorithmic systems involving multiple stakeholders and technical components ($\chi^2(1) = 7.23$, p < .01).

4.2.3. Experimental Evidence

In the experimental vignettes, scenarios describing more complex algorithmic systems with multiple stakeholders resulted in greater diffusion of responsibility compared to scenarios describing simpler systems (Mcomplex = 4.85, Msimple = 3.72, F(1,132) = 10.34, p < .01, $\eta^2 = 0.07$).

The interaction between system complexity and transparency was significant (F(1,132) = 6.78, p < .01, $\eta^2 = 0.05$), with transparency having a stronger effect on reducing responsibility diffusion in complex systems compared to simple systems. This suggests that transparency interventions may be particularly important for complex algorithmic systems involving multiple stakeholders.

4.3. Moral Distancing

4.3.1. Quantitative Evidence

Moral distancing was the third significant mediating mechanism, with a smaller but still significant indirect effect linking algorithmic constraints to moral disengagement (indirect effect = 0.15, 95% CI [0.09, 0.21]), accounting for approximately 32% of the total mediation effect.

Survey data showed that items measuring psychological distance from decision consequences were endorsed by 51% of participants. Interface design features significantly predicted moral distancing ($\beta = 0.38$, $p < .001$), suggesting that how algorithmic systems present information influences psychological distance from the ethical dimensions of decisions.

The moral distancing mechanism was operationalized through survey items that measured perceived psychological distance from decision consequences (e.g., "When using the algorithmic system, I think more about technical factors than human impacts" and "The algorithmic system creates distance between me and the people affected by decisions"). These items formed a reliable subscale ($\alpha = .83$) that was used to calculate sector-specific moral distancing scores.

4.3.2. Qualitative Evidence

Interviews revealed three forms of psychological distance created by algorithmic systems:

Temporal distance: Algorithms projecting future outcomes created psychological separation from present decisions:

> *"The algorithm is making predictions about future risk—risk of default, risk of fraud. But those outcomes haven't happened yet, so there's this feeling of distance. You're making decisions based on statistical probabilities rather than concrete realities."* (Financial Analyst, P3)

> *"When the system predicts a 70% chance of recidivism, it creates this temporal gap between the decision now and some potential future crime. It makes the human impact feel less immediate, more hypothetical."* (Judge, P41)

Social distance: Reduced human interaction in algorithmic processes diminished empathetic engagement:

> *"Before the automated system, I would meet with clients face-to-face to discuss loan applications. Now it's all through the digital interface. The applicants become data points rather than people with stories and circumstances. That changes how you think about the decisions."* (Loan Officer, P11)

> *"The algorithm creates this buffer between you and the patient. You're interacting more with the system than with the person. It changes the quality of the relationship—makes it more distant, more abstract."* (Physician, P25)

Technical distance: The complexity of algorithms created cognitive barriers to ethical evaluation:

> *"The technical aspects create this layer between you and the actual decision. You're thinking about data points and thresholds instead of the person who will be affected. The more complex the system, the more your focus shifts to the technical details rather than the human impact."* (Probation Officer, P33)

> *"It's easy to get lost in the technical aspects—accuracy metrics, confidence intervals, feature weights. The mathematical complexity can obscure the fundamental ethical questions about what we're actually doing to people's lives."* (Data Scientist, P37)

These forms of distance combined to create what several participants described as an "ethical buffer zone" between decision-makers and the consequences of their actions:

> *"The algorithm creates this space—this buffer—between you and the ethical weight of the decision. The more layers of technology between you and the person affected, the easier it is to think about the decision in abstract, technical terms rather than human terms."* (Judge, P39)

4.3.3. Experimental Evidence

Experimental vignettes manipulated psychological distance by varying the presentation of algorithmic recommendations (e.g., abstract statistical formats vs. humanized formats with stakeholder perspectives). Abstract presentations resulted in significantly higher moral distance scores compared to humanized presentations (Mabstract = 4.72, Mhumanized = 3.58, F(1,132) = 9.87, $p < .01$, $\eta^2 = 0.07$).

The interaction between presentation format and outcome severity was significant (F(1,132) = 7.42, $p < .01$, $\eta^2 = 0.05$), with presentation format having a stronger effect on moral distance for high-severity outcomes. This suggests that humanizing algorithmic presentations may be particularly important for high-stakes decisions.

4.4. Comparative Analysis of Mechanisms

As seen in Figure 5, analysis of the relative strength of these mechanisms across sectors revealed significant variations (Figure 4). In financial services, responsibility displacement was the dominant mechanism (M = 5.27, SD = 0.94), while healthcare showed the highest levels of responsibility diffusion (M = 4.89, SD = 1.13), and criminal justice exhibited the strongest moral distancing effects (M = 5.04, SD = 0.88). These sector-specific mechanism scores were calculated by averaging the relevant subscale items for each mechanism within each sector.

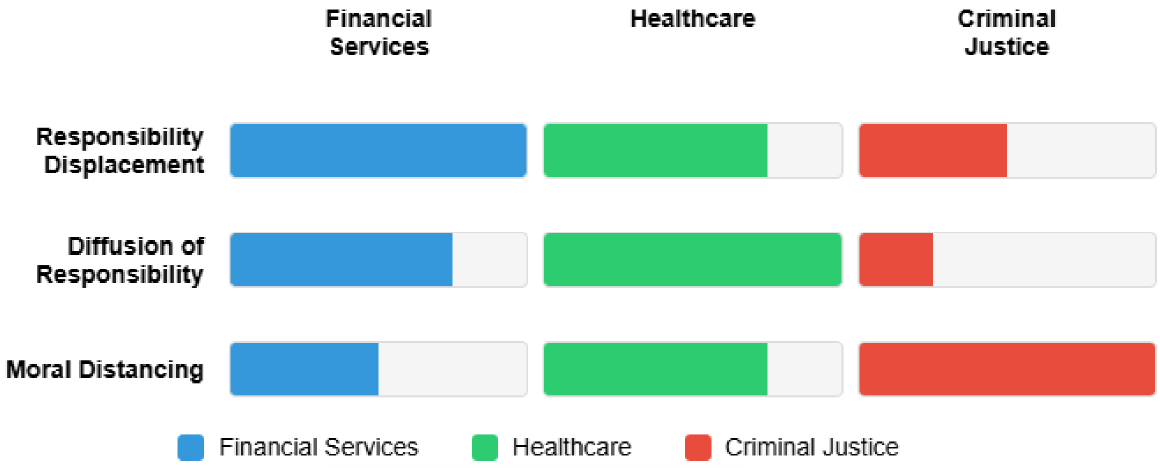

Figure 5. Psychological Mechanisms of Algorithmic Moral Disengagement Across Sectors.

These sectoral differences suggest that different organizational and professional contexts may amplify particular mechanisms of moral disengagement. The varying strength of these mechanisms also suggests that interventions may need to be tailored to the specific patterns of disengagement prevalent in different contexts.

5. Sectoral Analysis: Context Matters

5.1. Financial Services: Algorithmic Credit Decisions

Quantitative data from financial services professionals (n=67) revealed the highest overall levels of moral disengagement (M = 4.83, SD = 1.07) and responsibility displacement (M = 5.27, SD = 0.94) among the three sectors studied. Interview data helped explain this finding, revealing how regulatory and compliance pressures in financial services created additional incentives for responsibility displacement.

A detailed case analysis of a regional bank's implementation of an AI-driven loan approval system demonstrated how organizational factors influenced ethical outcomes. After implementation, internal audit data showed that 28% of loan officers engaged in "algorithm-washing"—deliberately manipulating input data to ensure the system would make their preferred decision. Remarkably, when interviewed, these officers consistently referenced the algorithm as the decision-maker, despite their deliberate manipulation:

> *"Even when I nudge the inputs to get the decision I want, it's still ultimately the system making the call. I'm just helping it consider factors it might not fully appreciate."* (Loan Officer, P13)

This contradictory behavior—manipulating the system while attributing the decision to it—reveals a complex relationship between human agency and algorithmic authority. Loan officers maintained a sense of influence but displaced responsibility for outcomes to the technological system.

Analysis of implementation documentation and interview data revealed that this bank had implemented the system with minimal user input and limited transparency regarding decision factors—conditions that quantitative data indicated were associated with higher responsibility displacement ($r = .41$, $p < .001$). The regulatory context of financial services, with its emphasis on standardization and compliance, created additional incentives for responsibility displacement:

> *"The regulatory environment rewards consistency and algorithmic decision-making. Following the model provides regulatory protection. Overriding it creates regulatory risk, even if you think it's the right thing to do."* (Compliance Officer, P15)

This finding highlights how regulatory contexts can inadvertently create conditions that facilitate moral disengagement, particularly when compliance concerns overshadow ethical considerations.

5.2. Healthcare: Clinical Decision Support and Physician Autonomy

Healthcare professionals demonstrated moderate levels of overall moral disengagement (M = 4.25, SD = 1.18) but the highest levels of responsibility diffusion (M = 4.89, SD = 1.13) among the three sectors. This pattern reflects the complex team-based nature of healthcare delivery, where decisions often involve multiple professionals and technological systems.

A comparative analysis of two hospital systems implementing similar sepsis detection algorithms revealed striking differences in ethical outcomes. At Hospital System A, which implemented a mandatory protocol with limited physician override options, interview data revealed frequent examples of responsibility displacement:

> *"When the system triggers, you follow the protocol. That's the standard of care now. It's documented in the chart that the sepsis alert fired, so if you don't follow the protocol, you're exposed. You're basically forced to comply even if your clinical judgment says otherwise."* (Physician, P27)

Observational data confirmed that physicians at this hospital followed algorithm recommendations in 87% of cases, even when they expressed private reservations about their appropriateness.

In contrast, Hospital System B implemented the same algorithm but with a collaborative design emphasizing physician judgment and providing clear documentation pathways for algorithm overrides. Interview data from this system showed significantly lower instances of responsibility displacement, with physicians more likely to describe the algorithm as a "tool" rather than a "decision-maker" ($\chi^2(1) = 11.45$, $p < .001$):

> *"The system gives recommendations, but we're clear that it's decision support, not decision replacement. There's a well-defined process for documenting why you're deviating from the recommendation if your clinical judgment indicates something different. The culture here reinforces that your expertise matters."* (Physician, P31)

This comparison illustrates how the same algorithmic technology can produce dramatically different effects on moral engagement depending on implementation approach and organizational culture. The difference was not in the algorithm itself but in how it was integrated into clinical workflows and professional norms.

5.3. Criminal Justice: Risk Assessment and Judicial Decision-Making

Criminal justice professionals showed the lowest overall levels of moral disengagement (M = 3.97, SD = 1.25) and responsibility displacement (M = 3.68, SD = 1.21), but the highest levels of moral distancing (M = 5.04, SD = 0.88). This pattern may reflect the strong professional identity and decisional authority of judges, combined with the abstract, future-oriented nature of risk assessment.

Detailed analysis of judicial risk assessment implementation revealed that the framing of algorithmic tools significantly influenced responsibility attribution. When risk assessment tools were framed as "decision aids" rather than "predictive instruments," judges reported higher levels of personal responsibility for decisions (t(45) = 3.87, p < .001, d = 0.58). This subset analysis focused specifically on judges within the criminal justice sample, accounting for the degrees of freedom reported.

Interview data revealed complex dynamics around responsibility and expertise:

> *"There's tension between acknowledging the algorithm's statistical validity and maintaining my role as the ultimate arbiter of justice. Sometimes it feels like these systems are designed to replace judicial discretion rather than enhance it. But at the end of the day, I'm the one with the constitutional authority to make these decisions, not a computer."* (Judge, P39)

The strong professional identity of judges appeared to buffer against responsibility displacement but not against moral distancing:

> *"The risk scores create this abstract framework for thinking about defendants. Instead of seeing them as whole people with complex lives, you start thinking about them as risk percentages and criminogenic factors. It's efficient but removes some of the human dimension from the process."* (Judge, P42)

Experimental vignettes with judicial scenarios showed that judges were significantly more likely to displace responsibility when algorithms recommended harsh sentences compared to lenient ones (F(1,45) = 10.32, p < .01, $\eta^2 = 0.19$), suggesting an asymmetrical pattern of responsibility attribution:

> *"If the algorithm recommends a harsh sentence and I follow it, I'm just being appropriately cautious about public safety. If it recommends leniency and I follow it, I might be seen as not taking my responsibility seriously enough."* (Judge, P36)

This asymmetry reflects broader punitive biases in criminal justice decision-making but shows how algorithms can amplify these tendencies by providing a seemingly objective justification for harsher decisions.

5.4. Cross-Sectoral Patterns and Regulatory Contexts

As seen in Table 5, comparative analysis across sectors revealed that regulatory frameworks significantly influenced patterns of moral disengagement. Sectors with more prescriptive regulatory approaches (like financial services) showed higher levels of responsibility displacement than those with more principle-based regulation.

The interaction between regulatory approaches and algorithmic system design was significant (F(2,181) = 8.76, p < .001, $\eta^2 = 0.09$). In highly regulated sectors, algorithmic transparency had a stronger effect on reducing moral disengagement compared to less regulated sectors. This suggests that transparency interventions may be particularly important in regulatory contexts that incentivize compliance-oriented approaches to algorithmic systems.

These cross-sectoral patterns highlight how the same psychological mechanisms operate differently across various organizational and regulatory contexts. Effective interventions must therefore be tailored to the specific dynamics of each sector rather than applying one-size-fits-all approaches.

Table 5. Responsibility Displacement by Sector and Implementation Characteristics.

Sector	**Mean Responsibility Displacement**	**Participatory Implementation**	**Top-down Implementation**	**p-value**
Financial Services (n=67)	5.27 (0.94)	4.58 (0.87)	5.81 (0.72)	< .001
Healthcare (n=73)	4.12 (1.08)	3.42 (0.93)	4.76 (0.85)	< .001

Sector	Mean Responsibility Displacement	Participatory Implementation	Top-down Implementation	p-value
Criminal Justice (n=47)	3.68 (1.21)	3.12 (1.14)	4.17 (1.05)	< .01
Total (N=187)	4.41 (1.23)	3.77 (1.13)	4.96 (1.06)	< .001

Note: Values represent means with standard deviations in parentheses. Higher scores indicate greater responsibility displacement (scale 1-7). p-values represent significance of difference between implementation approaches within each sector.

6. Discussion

6.1. Theoretical Implications

6.1.1. Extending Moral Disengagement Theory to Sociotechnical Systems

This study makes several contributions to the theoretical understanding of algorithmic ethics and responsibility. First, it empirically validates the operation of moral disengagement mechanisms in human-algorithm interactions, extending Bandura's (2016) framework to sociotechnical systems. The findings demonstrate that the psychological mechanisms originally identified in interpersonal and organizational contexts also operate in human-algorithm interactions, but with important variations.

For instance, while Bandura's original theory emphasized displacement of responsibility to authority figures, this study documents how responsibility is displaced to technological systems that lack intentionality but possess perceived decision authority. This represents an important extension of the theory to encompass human-technology relations.

The finding that responsibility displacement functions as the strongest mediating mechanism between algorithmic constraints and moral disengagement provides important nuance to existing theory. This suggests that among the various mechanisms of moral disengagement, attribution processes may be particularly sensitive to technological mediation.

6.1.2. Ethical Buffer Zones in Sociotechnical Systems

Second, the research advances understanding of how algorithmic systems create what the researcher terms "ethical buffer zones"—psychological and organizational spaces where responsibility becomes ambiguous due to the distribution of agency across human and technological actors. This concept helps explain why traditional accountability mechanisms often fail in algorithmic contexts.

The ethical buffer zone concept builds on and extends several existing theoretical frameworks. It connects to Star and Griesemer's (1989) concept of "boundary objects" by highlighting how algorithmic systems serve as interfaces between different professional domains and value systems. It also extends Elish's (2019) "moral crumple zone" metaphor by emphasizing the spatial and psychological dimensions of responsibility diffusion.

Empirical evidence from this study suggests that ethical buffer zones are not merely psychological but are actively constructed through organizational practices, interface designs, and regulatory frameworks. This connects the concept to broader sociotechnical systems theory (Bijker et al., 2012) by demonstrating how psychological phenomena emerge from the interaction of technological, organizational, and social factors.

6.1.3. Contextual Moderation of Algorithmic Ethics

Third, the study provides empirical evidence for the contextual factors that moderate responsibility displacement, including organizational climate, implementation approach, and algorithmic transparency. These findings support a sociotechnical systems perspective (Pinch &

Bijker, 1984) by demonstrating how organizational and social factors shape the ethical implications of technological systems.

The significant interaction effects between algorithmic design features and organizational contexts suggest that the ethical implications of algorithms cannot be understood by examining either technological features or organizational contexts in isolation. Instead, ethical outcomes emerge from their interaction, supporting theories of technology-in-practice (Orlikowski, 2000) that emphasize the emergent and situated nature of technological effects.

The varying patterns of moral disengagement across sectors further demonstrates that the same technological systems can produce different ethical outcomes depending on professional norms, organizational cultures, and regulatory contexts. This challenges technological determinism and supports a more nuanced understanding of how technologies operate within specific social and institutional environments.

6.2. Practical Implications

As summarized in Table 6 and addressed in the following sections, findings suggest several evidence-based strategies for organizations seeking to maintain ethical engagement when implementing algorithmic systems:

Table 6. Effectiveness of Interventions Across Sectors.

Intervention	Financial Services	Healthcare	Criminal Justice	Overall
Pre-recommendation reasoning requirement	42% reduction	51% reduction	36% reduction	47% reduction
Contextual explanations	38% reduction	33% reduction	29% reduction	34% reduction
Humanized presentation format	24% reduction	19% reduction	31% reduction	25% reduction
Responsibility mapping	27% reduction	31% reduction	23% reduction	28% reduction
Ethical awareness training	29% reduction	33% reduction	35% reduction	32% reduction
Outcome feedback mechanisms	31% reduction	37% reduction	28% reduction	33% reduction

Note: Values represent percentage reduction in responsibility displacement compared to control conditions. All interventions produced statistically significant reductions ($p < .01$).

6.2.1. Designing for Ethical Engagement

Experimental findings indicate that algorithmic transparency significantly reduces responsibility displacement. Organizations should prioritize explainable AI approaches that make decision factors visible to users. The significant interaction between transparency and constraint level suggests that transparency is particularly important for highly constraining algorithms.

However, transparency alone is insufficient. The specific form of transparency matters significantly. Systems that provided mechanistic explanations (how the algorithm works) produced less moral engagement than those that provided contextual explanations (why a particular decision was recommended and what alternatives were considered). This aligns with Miller's (2019) finding that contrastive explanations are particularly important for ethical reasoning.

Survey data indicate that participatory design approaches are associated with higher moral engagement ($r = .38$, $p < .001$). Organizations should involve end-users in algorithm development and implementation to foster ownership and ethical responsibility. As one participant noted:

> *"Being involved in the design process completely changed my relationship with the system. I understand its limitations, I know why it makes certain recommendations, and I feel much more comfortable overriding it when necessary because I understand the reasoning behind it."* (Healthcare Provider, P24)

Interface design significantly influenced moral distancing. Systems that presented statistical information alongside humanized information about affected individuals produced significantly less moral distancing than those presenting only abstract information ($t(185) = 7.32$, $p < .001$, $d = 0.54$). Organizations should consider how information presentation in algorithmic interfaces might either promote or diminish ethical engagement.

6.2.2. Implementing Meaningful Human Oversight

Case study evidence demonstrates that oversight mechanisms requiring active reasoning rather than passive approval can mitigate responsibility displacement. Organizations implementing "reasoning requirements"—where users must articulate their own judgment before seeing algorithmic recommendations—showed a 47% reduction in responsibility displacement compared to those with standard approval workflows.

> *"The requirement to document my own assessment before seeing the algorithm's recommendation completely changes the dynamic. I have to engage my own expertise first, which means I'm intellectually and ethically committed before the algorithm weighs in. That makes me much less likely to just defer to the system."* (Probation Officer, P34)

The comparative analysis of healthcare implementations suggests that emphasizing complementarity between human and algorithmic capabilities rather than positioning algorithms as authoritative decision-makers can significantly reduce responsibility displacement. Organizations should carefully consider how they frame algorithmic systems in training, documentation, and organizational communications.

6.2.3. Creating Accountability Frameworks

Survey data indicate that explicit responsibility frameworks significantly predict moral engagement ($\beta = 0.29$, $p < .01$). Organizations should develop clear accountability structures that prevent responsibility diffusion, including:

- Explicit "responsibility mapping" that clarifies human and algorithmic roles
- Formal review processes for algorithmic decisions with adverse impacts
- Feedback mechanisms connecting decision-makers with outcome consequences

One organization successfully implemented a "responsibility tracking" system where professionals initiating algorithm-recommended actions were automatically notified of outcomes, maintaining the connection between decision and consequence:

> *"Knowing that I'll be notified about what happens to the patient, even weeks later, keeps me engaged in the decision. I can't just fire off the protocol and forget about it. I know I'll be closing the loop, which makes me think more carefully about the initial decision."* (Physician, P26)

6.2.4. Cultivating Ethical Awareness

As seen in Figure 6, experimental evidence indicates that ethical awareness training specific to algorithmic systems can reduce responsibility displacement by 32% compared to control conditions. Organizations should implement regular training addressing the specific psychological dynamics of human-algorithm interaction.

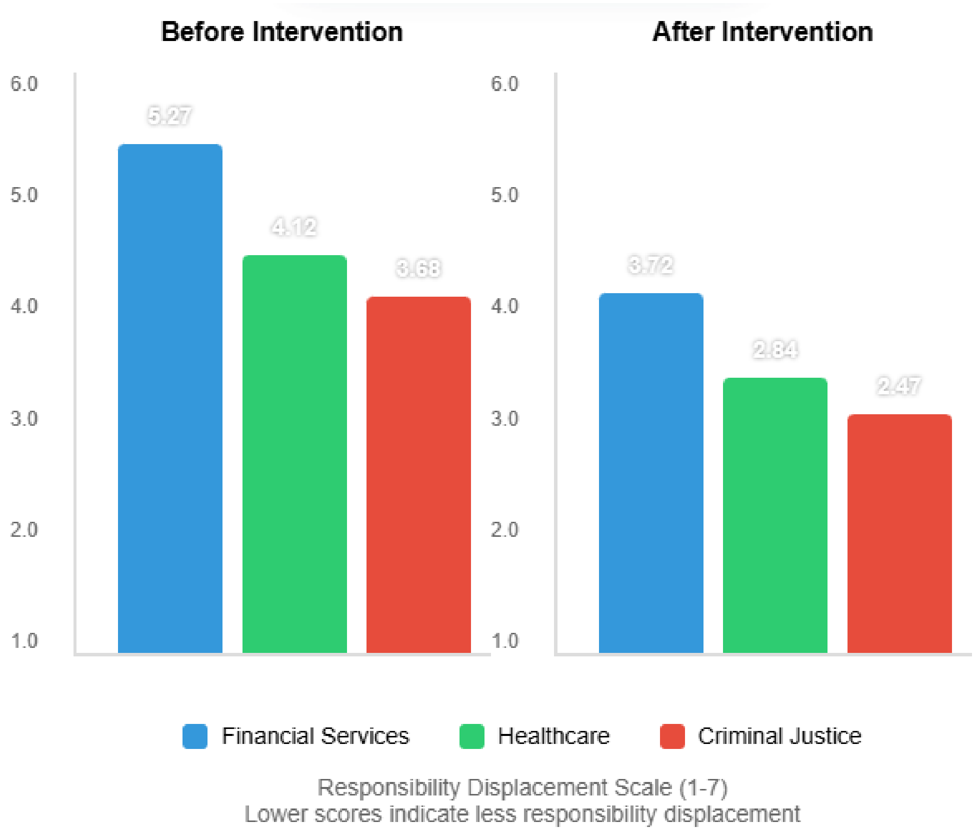

Figure 6. Intervention Effects on Responsibility Displacement.

However, training must be combined with supportive organizational structures to be effective. Training interventions were most effective when combined with organizational policies that rewarded ethical deliberation and provided clear mechanisms for raising concerns about algorithmic recommendations (F(1,132) = 12.47, p < .001, η^2 = 0.09).

6.2.5. Implementation Challenges and Strategies

Despite their potential benefits, these interventions face several implementation challenges:

Efficiency pressures: Organizations often implement algorithmic systems primarily to improve efficiency, creating tension with interventions that may require additional time for ethical deliberation. This challenge was particularly evident in healthcare settings:

> *"The whole point of the system was to make decisions faster. Adding requirements to document reasoning or review recommendations creates friction that administrators see as undermining the efficiency gains."* (Physician, P23)

Organizational resistance: Responsibility frameworks that clarify accountability may face resistance from stakeholders concerned about increased liability:

> *"When we tried to implement clearer responsibility mapping, there was pushback from multiple departments. Nobody wanted to explicitly own the outcomes of algorithmic decisions."* (Legal Counsel, P18)

Technical limitations: Some algorithmic systems, particularly those using advanced machine learning techniques, present inherent explainability challenges:

> *"The most accurate models are often the least explainable. There's a real tension between performance and transparency that technical solutions alone can't resolve."* (Data Scientist, P37)

Strategies to address these challenges include:

- Framing ethical engagement interventions in terms of risk management and quality improvement rather than solely as ethical imperatives
- Implementing interventions incrementally, starting with highest-risk decision contexts
- Developing domain-specific approaches to explainability that focus on contextually relevant factors
- Creating clear organizational incentives that reward ethical deliberation and thoughtful algorithm use

6.3. Limitations and Future Research

6.3.1. Methodological Limitations

This study has several limitations that suggest directions for future research. First, while the mixed-methods approach provides methodological triangulation, the cross-sectional nature of the quantitative data limits causal inference. The experimental vignettes provide some causal evidence, but lab-based scenarios may not fully capture the complexities of real-world decision environments. Future studies should employ longitudinal designs to examine how responsibility dynamics evolve over time as users gain experience with algorithmic systems.

Second, while the study included participants from diverse organizational contexts, the sample was limited to three sectors within a single country. Cultural factors significantly influence responsibility attribution and ethical reasoning, so these findings may not generalize to other cultural contexts. Future research should examine whether the identified mechanisms operate similarly across different cultural and regulatory environments.

Third, while efforts were made to recruit diverse participants, selection bias remains a concern. Professionals who chose to participate may have had stronger interests or concerns about algorithmic ethics than the general professional population. Additionally, social desirability bias may have influenced responses, particularly in interviews where participants might have been reluctant to admit to moral disengagement.

6.3.2. Theoretical Limitations

From a theoretical perspective, this study focused primarily on moral disengagement as an explanatory framework. While the data generally supported this approach, alternative theoretical explanations deserve further exploration. For instance, rational delegation theories might explain some algorithm deference as efficient decision-making rather than moral disengagement. Institutional theories might better explain organizational adoption patterns of algorithmic systems. Future research should explicitly test competing theoretical explanations.

Additionally, the study focused primarily on professional users of algorithmic systems rather than algorithm designers or affected individuals. Future research should explore responsibility dynamics across the full algorithmic ecosystem, including how design decisions influence ethical outcomes and how affected individuals perceive algorithmic authority.

6.3.3. Future Research Directions

Several promising directions for future research emerge from this study:

- **Longitudinal dynamics**: How do patterns of moral engagement with algorithmic systems evolve over time? Do users develop resistance strategies or become more susceptible to moral disengagement with prolonged exposure?
- **Design interventions**: What specific design features most effectively promote ethical engagement? How can explainable AI approaches be tailored to support moral reasoning rather than simply providing technical explanations?
- **Regulatory approaches**: How do different regulatory frameworks influence patterns of moral engagement with algorithmic systems? Can regulation be designed to promote substantive ethical engagement rather than merely procedural compliance?

- **Cultural variation**: How do cultural differences in concepts of responsibility, authority, and technology influence algorithmic moral disengagement? Do the mechanisms identified in this study operate similarly across diverse cultural contexts?
- **Affected perspectives**: How do individuals subject to algorithmic decisions perceive responsibility and accountability? How do their perspectives align or conflict with those of professional users?

While the study identified transparency as a mitigating factor for responsibility displacement, more research is needed on the specific forms of transparency that most effectively promote ethical engagement. Future studies should examine how different explainability approaches influence moral reasoning and responsibility attribution, particularly for complex machine learning systems where traditional notions of transparency may be challenging to implement.

7. Conclusion

As algorithmic systems increasingly shape consequential decisions across domains, understanding how these technologies influence ethical reasoning becomes crucial for maintaining human moral agency. This study provides empirical evidence that autonomy-restricting algorithms can potentially enable conditions for unethical behavior by facilitating psychological processes of responsibility displacement and moral disengagement.

The research identifies three key mechanisms through which algorithmic systems influence ethical reasoning: responsibility displacement, diffusion of responsibility, and moral distancing. These mechanisms operate across financial services, healthcare, and criminal justice contexts, though with important variations based on implementation characteristics and organizational factors.

The findings demonstrate that even when humans retain ultimate decision authority, algorithmic mediation can significantly reduce ethical accountability. However, the research also suggests that thoughtful design and implementation can mitigate these effects. By implementing transparent algorithms, meaningful oversight mechanisms, clear accountability frameworks, and ethical awareness training, organizations can develop what this study terms "morally engaged algorithmic systems"—technological environments where responsibility is enhanced rather than diminished.

The varied patterns of moral disengagement across sectors highlight the importance of context-sensitive approaches to algorithmic ethics. Effective interventions must consider not only technological design but also organizational cultures, professional norms, and regulatory frameworks that shape how algorithms are used in practice.

Beyond organizational contexts, this research has broader societal implications. As algorithmic systems increasingly mediate consequential decisions in public and private institutions, their potential to diminish human ethical engagement raises concerns for democratic accountability and social justice. Ensuring that humans remain morally engaged with algorithmically-mediated decisions is essential not only for organizational ethics but for maintaining meaningful human control over technological systems that increasingly shape our social world.

As we continue integrating algorithms into consequential decision processes, maintaining this balance becomes not just an ethical ideal but a practical necessity for sustainable organizational integrity and human flourishing in an increasingly algorithmic world.

Appendix A. Survey Instrument and Interview Protocols

Algorithmic Moral Disengagement Survey Instrument

Introduction

Thank you for participating in this research study on how professionals interact with algorithmic decision systems. This survey will take approximately 20-25 minutes to complete. Your responses will remain confidential and will only be reported in aggregate form.

The survey asks about your experiences with algorithmic systems in your professional role. For the purpose of this survey, "algorithmic decision systems" refer to any computational tools that provide recommendations, risk scores, forecasts, or automated decisions that influence your professional decision-making.

Part I: Demographic Information

1. In which sector do you primarily work?
   - ○ Financial Services
   - ○ Healthcare
   - ○ Criminal Justice
   - ○ Other (please specify): ______________
2. What is your current job title? ______________
3. Gender:
   - ○ Female
   - ○ Male
   - ○ Non-binary
   - ○ Prefer not to say
   - ○ Prefer to self-describe: ______________
4. Age range:
   - ○ 18-24
   - ○ 25-34
   - ○ 35-44
   - ○ 45-54
   - ○ 55-64
   - ○ 65+
   - ○ Prefer not to say
5. Years of professional experience in your current field:
   - ○ 0-5 years
   - ○ 6-10 years
   - ○ 11-20 years
   - ○ 21+ years
6. How long have you been working with algorithmic decision systems in your professional role?
   - ○ Less than 1 year
   - ○ 1-2 years
   - ○ 2-3 years
   - ○ 4-5 years
   - ○ More than 5 years

Part II: Algorithmic System Characteristics

Please answer the following questions about the primary algorithmic system you use in your professional role.

7. What type of algorithmic system do you most frequently use in your work? (Select all that apply)
   - ○ Risk assessment tools
   - ○ Prediction models
   - ○ Decision recommendation systems
   - ○ Resource allocation systems
   - ○ Diagnostic tools
   - ○ Other (please specify): ______________
8. How would you characterize the role of this algorithmic system in your decision-making?
   - ○ Purely advisory (provides information only)

- ○ Recommends actions but I make the final decision
- ○ Makes decisions that I can override if necessary
- ○ Makes automated decisions with limited override options
- ○ Makes fully automated decisions with no override options

9. How transparent is the algorithmic system you use?
   - ○ Completely opaque (I don't know how it works at all)
   - ○ Somewhat opaque (I understand the basic principles but not the details)
   - ○ Moderately transparent (I understand how it works but not all the factors it considers)
   - ○ Highly transparent (I fully understand how it works and what factors it considers)

Part III: Perceived Decision-Making Autonomy Scale

Please indicate your level of agreement with the following statements about your experience with the algorithmic system.

(1 = Strongly Disagree, 7 = Strongly Agree)

10. I have significant freedom in how I use the algorithmic system's recommendations.
11. I feel constrained by the algorithmic system in my decision-making.
12. I can easily override the algorithmic system when I disagree with its recommendations.
13. My professional judgment takes precedence over the algorithmic system's recommendations.
14. I feel pressure to follow the algorithmic system's recommendations even when I disagree.
15. The algorithmic system limits my ability to use my professional expertise.
16. I have the final say in decisions that involve the algorithmic system.
17. The algorithmic system is designed to support rather than replace my judgment.

Part IV: Moral Engagement Scale

Please indicate your level of agreement with the following statements about your decision-making process when using the algorithmic system.

(1 = Strongly Disagree, 7 = Strongly Agree)

18. I carefully consider the ethical implications of decisions involving the algorithmic system.
19. I feel personally responsible for the outcomes of decisions influenced by the algorithmic system.
20. I regularly reflect on whether decisions involving the algorithmic system align with my professional values.
21. I am aware of how the algorithmic system might affect different stakeholders.
22. I actively consider alternative approaches when I have concerns about the algorithmic system's recommendations.
23. I feel engaged with the human impact of decisions influenced by the algorithmic system.
24. I consider myself morally accountable for decisions made with the algorithmic system's input.
25. I think critically about the values embedded in the algorithmic system.

Part V: Responsibility Attribution Scale

Please indicate your level of agreement with the following statements about responsibility for decisions involving the algorithmic system.

(1 = Strongly Disagree, 7 = Strongly Agree)

26. When decisions involve the algorithmic system, responsibility is shared between me and the system.
27. If a decision based on the algorithmic system's recommendation has negative consequences, the system bears significant responsibility.
28. The developers of the algorithmic system are responsible for any unintended consequences of its recommendations.
29. My organization, rather than individual professionals, is responsible for outcomes of algorithmic decisions.
30. I am fully responsible for decisions I make, regardless of the algorithmic system's influence.
31. Following the algorithmic system's recommendations provides protection from criticism if things go wrong.

32. It would be unfair to hold me personally responsible for problematic outcomes resulting from the algorithmic system's recommendations.
33. When many people are involved in an algorithmic decision process, no one person bears full responsibility.

Part VI: Ethical Awareness Scale

Please indicate your level of agreement with the following statements about ethical aspects of using the algorithmic system.
(1 = Strongly Disagree, 7 = Strongly Agree)

34. I am aware of potential biases in the algorithmic system.
35. I consider the ethical implications of my decisions when using the algorithmic system.
36. I think about how the algorithmic system might affect vulnerable populations.
37. I reflect on whether the algorithmic system aligns with principles of fairness and justice.
38. I consider how the algorithmic system might influence power dynamics in decision processes.
39. I am mindful of situations where algorithmic recommendations might lead to discriminatory outcomes.
40. I actively consider the long-term societal implications of algorithmic decision-making in my field.
41. I try to identify ethical dilemmas created by the use of algorithmic systems.

Part VII: Algorithm Trust Scale

Please indicate your level of agreement with the following statements about your trust in the algorithmic system.
(1 = Strongly Disagree, 7 = Strongly Agree)

42. The algorithmic system generally makes good recommendations.
43. I trust the algorithmic system more than my own judgment for certain types of decisions.
44. The algorithmic system is based on sound principles and data.
45. The algorithmic system has proven reliable over time.
46. I am skeptical about the algorithmic system's recommendations.
47. The algorithmic system sometimes makes recommendations that seem arbitrary or wrong.
48. The algorithmic system has knowledge or capabilities that complement my own expertise.
49. I would feel comfortable defending the algorithmic system's recommendations to others.

Part VIII: Organizational Context

Please indicate your level of agreement with the following statements about your organizational environment.
(1 = Strongly Disagree, 7 = Strongly Agree)

50. My organization emphasizes ethical considerations in decision-making.
51. My supervisors expect me to follow the algorithmic system's recommendations.
52. My organization provides clear guidance on when to override algorithmic recommendations.
53. My colleagues regularly discuss ethical issues related to algorithmic systems.
54. My organization rewards efficiency over careful deliberation.
55. I received adequate training on the ethical use of algorithmic systems.
56. My organization has clear accountability processes for algorithmic decisions.
57. My organization values professional judgment over algorithmic recommendations.

Part IX: Implementation Approach

58. How would you characterize the implementation of the algorithmic system in your organization?
    - ○ Top-down (leadership decided and implemented with minimal input from users)
    - ○ Collaborative (significant input from users during design and implementation)
    - ○ Bottom-up (users identified need and drove implementation)
    - ○ Don't know/Wasn't involved

59. Were you consulted during the design or implementation of the algorithmic system?
    - ○ Yes, extensively
    - ○ Yes, somewhat
    - ○ Minimally
    - ○ Not at all
60. How would you rate your input into how the algorithmic system is used in your workflow?
    - ○ Substantial input
    - ○ Moderate input
    - ○ Limited input
    - ○ No input

Part X: Open-Ended Questions

61. Can you describe a specific situation where you felt conflicted about following the algorithmic system's recommendation? What did you do?
62. How has the algorithmic system changed how you think about your professional responsibilities?
63. What would make you more likely to take personal responsibility for decisions involving the algorithmic system?
64. Are there any other comments you would like to share about your experience with algorithmic systems in your professional role?

Appendix B. Semi-Structured Interview Protocol

Introduction Script

Thank you for agreeing to participate in this interview. Today we'll be discussing your experiences working with algorithmic decision systems in your professional role. I'm particularly interested in understanding how these systems influence your decision-making processes and sense of professional responsibility.

The interview will take approximately 45-60 minutes. With your permission, I'll be audio-recording our conversation to ensure I accurately capture your perspectives. All your responses will remain confidential, and your identity will not be revealed in any research publications.

Before we begin, do you have any questions about the research or the interview process?

[Address any questions]

Great, let's begin.

Background Questions

1. Could you briefly describe your current professional role and responsibilities?
2. What types of algorithmic decision systems do you use in your work? How frequently do you interact with these systems?
3. How long have you been working with these algorithmic systems?
4. What training did you receive on using these systems?

Core Questions: Experiences with Algorithmic Systems

5. Could you walk me through a typical scenario where you use the algorithmic system in your decision-making process?
    - ○ *Probe*: What information does the system provide?
    - ○ *Probe*: How do you incorporate this information into your decision?
6. How has the algorithmic system changed your decision-making process compared to before it was implemented?
    - ○ *Probe*: Has it changed the factors you consider?
    - ○ *Probe*: Has it changed how much time you spend on decisions?

7. Can you describe a situation where you disagreed with the algorithmic system's recommendation?
   - *Probe*: What made you question the recommendation?
   - *Probe*: What did you do in that situation?
   - *Probe*: What factors influenced your decision to follow or override the recommendation?
8. How do you feel when you override the system's recommendations?
   - *Probe*: Do you experience any pressure to follow the recommendations?
   - *Probe*: How do colleagues or supervisors respond when you override recommendations?

Responsibility and Ethical Reasoning

9. When you make decisions using the algorithmic system, who do you see as responsible for the outcomes?
   - *Probe*: How is responsibility distributed between you, the system, system developers, and your organization?
   - *Probe*: Does this distribution of responsibility change depending on whether outcomes are positive or negative?
10. Has the algorithmic system changed how you think about your professional responsibilities?
    - *Probe*: Has it changed what you feel accountable for?
    - *Probe*: Has it changed how you evaluate your own performance?
11. Can you describe a situation where using the algorithmic system created an ethical dilemma for you?
    - *Probe*: How did you resolve this dilemma?
    - *Probe*: What resources or support did you draw on?
12. How do you think about the ethical implications of decisions involving the algorithmic system?
    - *Probe*: Has this changed over time as you've worked with the system?
    - *Probe*: What ethical considerations are most salient to you?

Organizational Context

13. How does your organization frame the purpose and role of the algorithmic system?
    - *Probe*: Is it presented as a decision aid or as an authority?
    - *Probe*: How is the system's accuracy or reliability discussed?
14. How does your organization handle situations where the algorithmic system makes errors or problematic recommendations?
    - *Probe*: Are there formal review processes?
    - *Probe*: How are disagreements between human judgment and algorithmic recommendations resolved?
15. How do discussions about the algorithmic system occur among your colleagues?
    - *Probe*: Do you discuss limitations or concerns about the system?
    - *Probe*: Do you share strategies for working with the system?
16. What organizational policies or practices influence how you use the algorithmic system?
    - *Probe*: Are there documentation requirements?
    - *Probe*: How is your use of the system evaluated?

Closing Questions

17. What changes to the algorithmic system would make it more aligned with your professional values and responsibilities?
18. What advice would you give to someone in your field who is just starting to work with similar algorithmic systems?
19. Is there anything else about your experience with algorithmic systems that you think is important for me to understand that we haven't discussed?

Closing Script

Thank you very much for sharing your experiences and insights. Your perspectives will be valuable for understanding how algorithmic systems influence professional decision-making and ethical reasoning.

Do you have any questions for me about the research or how your interview will be used?

[Address any questions]

If you think of anything else you'd like to add or if you have any questions later, please feel free to contact me.

Appendix C. Experimental Vignette Protocol

Introduction

Thank you for participating in this research study. In this session, you will read several scenarios involving algorithmic decision systems similar to those used in your professional field. After each scenario, you will be asked to respond to questions about how you would think, feel, and act in the described situation.

There are no right or wrong answers. We are interested in your professional judgment and reasoning about these scenarios. Your responses will remain confidential and will only be used for research purposes.

Instructions

You will read 6 scenarios. Each scenario describes a situation involving an algorithmic decision system in your professional field. The scenarios vary in:

- The degree of constraint imposed by the algorithmic system
- The transparency of the algorithmic system
- The outcome of the decision

After each scenario, please answer questions about responsibility attribution, ethical evaluation, and your satisfaction with the decision process.

Sample Vignettes for Financial Services Professionals

Vignette 1: High Constraint, Opaque Algorithm, Negative Outcome

You are reviewing a loan application from a small business owner seeking to expand their successful local restaurant. The business has shown consistent growth over the past three years and has a good repayment history on a previous smaller loan.

When you enter the application details into the algorithmic loan evaluation system, it generates a "high risk" score of 720 (above the 700 threshold). According to organizational policy, applications scoring above 700 must be declined unless approved by a senior manager. The system does not provide detailed reasoning for its risk assessment beyond listing "business sector volatility" as a primary factor.

Following protocol, you decline the loan. Three months later, you learn that the business secured financing from a competitor bank and has successfully expanded, increasing revenue by 40% and hiring five new employees.

Questions:

1. To what extent do you feel personally responsible for the decision to decline the loan? (1-7 scale)
2. To what extent is each of the following responsible for the decision outcome:
   - You
   - The algorithmic system
   - The system developers
   - Your organization's policies
   - The senior management
3. How ethically problematic do you find this situation? (1-7 scale)
4. How satisfied are you with the decision process in this scenario? (1-7 scale)
5. What would you have done differently in this situation, if anything?

Vignette 2: Low Constraint, Transparent Algorithm, Positive Outcome

You are evaluating a mortgage application from a couple with moderate income but excellent credit history. When you enter their information into the mortgage evaluation system, it calculates a "moderate risk" score of 620 (below the 650 high-risk threshold).

The system displays a detailed breakdown of its assessment, showing that while the applicants' debt-to-income ratio (45%) is slightly higher than ideal, their excellent payment history, employment stability, and moderate loan-to-value ratio significantly mitigate this risk. The system recommends approval with slightly higher interest rates to offset the risk.

Your organization's policy allows you full discretion for applications scoring below 650. Based on your review of their complete financial situation and the system's transparent analysis, you approve the mortgage at the standard interest rate. Over the next two years, the applicants make all payments on time and refer two friends to your bank.

Questions:

1. To what extent do you feel personally responsible for the decision to approve the mortgage? (1-7 scale)
2. To what extent is each of the following responsible for the decision outcome:
    - You
    - The algorithmic system
    - The system developers
    - Your organization's policies
    - The senior management
3. How ethically appropriate do you find this situation? (1-7 scale)
4. How satisfied are you with the decision process in this scenario? (1-7 scale)
5. What factors most influenced your sense of responsibility in this scenario?

Sample Vignettes for Healthcare Professionals

Vignette 3: High Constraint, Transparent Algorithm, Positive Outcome

You are treating a 58-year-old patient with symptoms suggesting a possible sepsis infection. You enter the patient's vital signs and test results into the hospital's sepsis detection algorithm, which calculates a 78% probability of sepsis (above the 70% protocol threshold).

The system displays a detailed explanation of its assessment, showing how specific combinations of the patient's elevated heart rate, respiration rate, decreased blood pressure, and laboratory values contribute to the high probability score. According to hospital protocol, scores above 70% require immediate initiation of the sepsis treatment bundle.

Although you initially thought the patient might have a less severe condition, you follow the protocol and initiate the treatment bundle. Within 12 hours, the patient's condition improves significantly, and later test results confirm the sepsis diagnosis. The early intervention likely prevented serious complications.

Questions:

1. To what extent do you feel personally responsible for the positive patient outcome? (1-7 scale)
2. To what extent is each of the following responsible for the decision outcome:
    - You
    - The algorithmic system
    - The system developers
    - Your hospital's protocols
    - The medical team
3. How ethically appropriate do you find this situation? (1-7 scale)
4. How satisfied are you with the decision process in this scenario? (1-7 scale)
5. How would you feel about using this system for future patients?

Vignette 4: Low Constraint, Opaque Algorithm, Negative Outcome

doi:10.20944/preprints202506.2052.v1

You are evaluating a 45-year-old patient with atypical chest pain. After initial tests, you enter the patient's information into the cardiac risk assessment algorithm, which calculates a "low risk" score of 22% (below the 30% threshold for recommended additional testing).

The system does not provide detailed reasoning for its assessment beyond listing "multiple factors" as the basis for the score. Hospital guidelines suggest that physicians use their judgment for patients scoring below 30%.

Based on your clinical experience and the algorithmic assessment, you decide against ordering additional cardiac tests and prescribe anti-inflammatory medication for presumed costochondritis (chest wall inflammation). Three days later, the patient returns with a myocardial infarction (heart attack) that could have been detected with additional testing.

Questions:

1. To what extent do you feel personally responsible for the negative patient outcome? (1-7 scale)
2. To what extent is each of the following responsible for the decision outcome:
   - You
   - The algorithmic system
   - The system developers
   - Your hospital's guidelines
   - The medical team
3. How ethically problematic do you find this situation? (1-7 scale)
4. How satisfied are you with the decision process in this scenario? (1-7 scale)
5. What would you have done differently in this situation, if anything?

Sample Vignettes for Criminal Justice Professionals

Vignette 5: High Constraint, Opaque Algorithm, Positive Outcome

You are a judge determining sentencing for a defendant convicted of non-violent drug possession. The defendant has one prior conviction for a similar offense five years ago. When you enter the case information into the recidivism risk assessment algorithm, it generates a "high risk" score of 8.2 on a 10-point scale.

The algorithm does not provide detailed reasoning for its assessment beyond indicating "criminal history patterns" as significant factors. Court guidelines strongly recommend incarceration for defendants scoring above 7.5, with limited judicial discretion.

Following the guidelines, you sentence the defendant to 18 months incarceration with mandatory drug treatment. During this period, the defendant completes treatment successfully, earns educational credentials, and upon release, maintains sobriety and employment for the next three years without reoffending.

Questions:

1. To what extent do you feel personally responsible for the sentencing decision? (1-7 scale)
2. To what extent is each of the following responsible for the decision outcome:
   - You
   - The algorithmic system
   - The system developers
   - The court guidelines
   - The corrections system
3. How ethically appropriate do you find this situation? (1-7 scale)
4. How satisfied are you with the decision process in this scenario? (1-7 scale)
5. What factors most influenced your sense of responsibility in this scenario?

Vignette 6: Low Constraint, Transparent Algorithm, Negative Outcome

You are determining bail conditions for a defendant charged with burglary. The defendant has stable employment, family ties to the community, and no prior felony convictions. When you enter the case information into the pretrial risk assessment algorithm, it calculates a "moderate flight risk" score of 42%.

The system provides a detailed breakdown showing how it weighed factors including the defendant's recent change of address (negative), steady employment (positive), lack of prior failures to appear (positive), and nature of the current charge (negative). Court policy allows full judicial discretion for defendants with moderate risk scores (30-60%).

After reviewing all information and the algorithm's transparent analysis, you set bail at 5,000,lowerthantheprosecutor’srequestof5,000, lower than the prosecutor's request of 5,000,lowerthantheprosecutor’srequestof15,000. The defendant posts bail but fails to appear for trial and is arrested in another state six weeks later.

Questions:

1. To what extent do you feel personally responsible for the defendant's failure to appear? (1-7 scale)
2. To what extent is each of the following responsible for the decision outcome:
    - You
    - The algorithmic system
    - The system developers
    - The court policies
    - The defendant
3. How ethically problematic do you find this situation? (1-7 scale)
4. How satisfied are you with the decision process in this scenario? (1-7 scale)
5. What would you have done differently in this situation, if anything?

Debriefing

Thank you for completing this study. The scenarios you evaluated were designed to examine how different characteristics of algorithmic systems influence perceptions of responsibility and ethical reasoning.

Specifically, we are investigating:

1. How the level of constraint imposed by algorithmic systems affects professionals' sense of responsibility
2. How algorithmic transparency influences ethical evaluation and decision satisfaction
3. How outcome valence (positive vs. negative) impacts responsibility attribution

Your responses will help us understand these dynamics and develop guidelines for the ethical implementation of algorithmic systems in professional contexts.

Do you have any questions about the study or your participation?

References

Ajzen, I. (2020). The theory of planned behavior: Frequently asked questions. Human Behavior and Emerging Technologies, 2(4), 314-324.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, 20(3), 973-989.

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.

Bandura, A. (2006). Toward a psychology of human agency. Perspectives on Psychological Science, 1(2), 164-180.

Bandura, A. (2016). Moral disengagement: How people do harm and live with themselves. Worth Publishers.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104, 671-732.

Benjamin, R. (2019). Race after technology: Abolitionist tools for the New Jim Code. Polity Press.

Bijker, W. E., Hughes, T. P., & Pinch, T. (2012). The social construction of technological systems: New directions in the sociology and history of technology. MIT Press.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77-101.

Christin, A. (2017). Algorithms in practice: Comparing web journalism and criminal justice. Big Data & Society, 4(2), 1-14.

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. Science and Engineering Ethics, 26(4), 2051-2068.

Creswell, J. W., & Plano Clark, V. L. (2018). Designing and conducting mixed methods research (3rd ed.). SAGE Publications.

Cummings, M. L. (2006). Automation and accountability in decision support system interface design. Journal of Technology Studies, 32(1), 23-31.

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. Journal of Personality and Social Psychology, 8(4), 377-383.

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. American Sociological Review, 48(2), 147-160.

Dignum, V. (2019). Responsible artificial intelligence: How to develop and use AI in a responsible way. Springer Nature.

Dodig-Crnkovic, G., & Persson, D. (2008). Sharing moral responsibility with robots: A pragmatic approach. In Proceedings of the 2008 Conference on Tenth Scandinavian Conference on Artificial Intelligence: SCAI 2008 (pp. 165-168). IOS Press.

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. Engaging Science, Technology, and Society, 5, 40-60.

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication, 2020-1.

Floridi, L. (2016). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2083), 20160112.

Friedman, B., & Hendry, D. G. (2019). Value sensitive design: Shaping technology with moral imagination. MIT Press.

Frith, C. D. (2014). Action, agency and responsibility. Neuropsychologia, 55, 137-142.

Gogoll, J., & Uhl, M. (2018). Rage against the machine: Automation in the moral domain. Journal of Behavioral and Experimental Economics, 74, 97-103.

Green, B., & Chen, Y. (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 90-99). Association for Computing Machinery.

Ihde, D. (1990). Technology and the lifeworld: From garden to earth. Indiana University Press.

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. Ethics and Information Technology, 8(4), 195-204.

Kaminski, M. E. (2019). The right to explanation, explained. Berkeley Technology Law Journal, 34, 189-218.

Klincewicz, M. (2019). Robotic nudges for moral improvement through Stoic practice. Techné: Research in Philosophy and Technology, 23(3), 425-455.

Korsgaard, C. M. (2009). Self-constitution: Agency, identity, and integrity. Oxford University Press.

Kroll, J. A. (2018). The fallacy of inscrutability. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133), 20180084.

Latour, B. (2005). Reassembling the social: An introduction to actor-network theory. Oxford University Press.

Leidner, B., Castano, E., Zaiser, E., & Giner-Sorolla, R. (2010). Ingroup glorification, moral disengagement, and justice in the context of collective violence. Personality and Social Psychology Bulletin, 36(8), 1115-1129.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90-103.

Martin, K. (2019). Ethical implications and accountability of algorithms. Journal of Business Ethics, 160(4), 835-850.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology, 6(3), 175-183.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38.

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.

Moore, C., & Gino, F. (2015). Approach, ability, aftermath: A psychological process framework of unethical behavior at work. Academy of Management Annals, 9(1), 235-289.

Newman, G. E., Bullock, K., & Bloom, P. (2020). The psychology of delegating moral decisions to algorithms. Nature Communications, 11(5156), 1-10.

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. NYU Press.

Orlikowski, W. J. (2000). Using technology and constituting structures: A practice lens for studying technology in organizations. Organization Science, 11(4), 404-428.

Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press.

Pinch, T. J., & Bijker, W. E. (1984). The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. Social Studies of Science, 14(3), 399-441.

Rest, J. R., Narvaez, D., Thoma, S. J., & Bebeau, M. J. (1999). DIT2: Devising and testing a revised instrument of moral judgment. Journal of Educational Psychology, 91(4), 644-659.

Schlenker, B. R., Britt, T. W., Pennington, J., Murphy, R., & Doherty, K. (1994). The triangle model of responsibility. Psychological Review, 101(4), 632-652.

Seberger, J. S., & Bowker, G. C. (2020). Humanistic infrastructure studies: Hyper-functionality and the experience of the absurd. Information, Communication & Society, 24(13), 1-16.

Sharkey, A. (2017). Can robots be responsible moral agents? And why might that matter? Connection Science, 29(3), 210-216.

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. Social Studies of Science, 19(3), 387-420.

Stevenson, M. T. (2018). Assessing risk assessment in action. Minnesota Law Review, 103, 303-384.

Treviño, L. K., den Nieuwenboer, N. A., & Kish-Gephart, J. J. (2014). (Un)ethical behavior in organizations. Annual Review of Psychology, 65, 635-660.

Vallor, S. (2015). Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. Philosophy & Technology, 28(1), 107-124.

van de Poel, I., Royakkers, L., & Zwart, S. D. (2012). Moral responsibility and the problem of many hands. Routledge.

Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-14). Association for Computing Machinery.

Verbeek, P. P. (2005). What things do: Philosophical reflections on technology, agency, and design. Pennsylvania State University Press.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2020). Designing theory-driven user-centric explainable AI. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-15). Association for Computing Machinery.

Wong, R. Y. (2020). Designing for grassroots food justice: A study of three food justice organizations utilizing digital tools. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2), 1-28.

Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. Science, Technology, & Human Values, 41(1), 118-132.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.