Extended Abstract

Technology Paternalism and the Reactance Deficit - Are Our Natural Protective Mechanisms Failing Us?

by Michelle Görlitz

With rapidly advancing artificial intelligence (AI) technology, the potential for paternalistic artificial agents grows. Autonomy restrictions have a detrimental impact on psychological well-being, regardless of the source of the restriction. What remains unclear is if the source is also irrelevant when it comes to effectively using our natural protective mechanisms, reactance reactions. This thesis investigates a critical and understudied psychological implication of such systems: Whether users maintain the capacity to detect and resist autonomy threats posed by paternalistic AI. Specifically, it introduces the concept of a *reactance deficit*, understood as the absence of the motivational state of reactance typically elicited in response to perceived restrictions of freedom, as described by Psychological Reactance Theory (PRT). The objective of this research is to enable future interactions with secured human autonomy, and thereby well-being, while benefiting from enhanced capabilities of AI-supported decision-making.

To investigate the complex relationship between AI paternalism, autonomy perceptions, and reactance reactions a mixed-method, scenario-based online experiment was conducted to empirically test the impact of AI paternalism on perceived human autonomy, perceived system autonomy and elicited reactance. Participants (N = 137) were randomly assigned to five experimental conditions representing increasing levels of AI-imposed paternalistic restriction (between-subjects factor). The control condition (C1) experienced no restrictions at all, while the experimental conditions interacted with AI-systems that provided either non-restrictive recommendations (C2), information about the restrictions and the option to reverse them (C3), non-reversible, informed restrictions (C4), and strongly paternalistic uninformed choice-restriction (C5). One interactive Figma prototype of a clothing web shop was designed per experimental condition to expose the participants to a realistic, interactive environment with an AI system integration. The system was introduced as an assistant, helping to make financial decisions based on the user's spending history and income.

The within-subjects factor assessed the dependent variables (perceived human autonomy, perceived system autonomy, and reactance) both before and after

participants received informational materials about AI-supported decision-making and its paternalistic potential.
To enrich the quantitative data further qualitative insights were generated by asking two open-ended questions inquisitioning general sentiment towards AI-supported decision-making.
The study's findings provide evidence for the presence of a reactance deficit, showing that individuals may have difficulty utilizing their natural protective mechanisms, reactance reactions, when confronted with highly paternalistic AI systems.
Surprisingly, the results also challenge core assumptions of the PRT, revealing that reactance reactions can be sparked without significant impact on the individual's perceived human autonomy, as no experimental condition showed significant changes comparing the two measures of perceived human autonomy. In terms of perceived system autonomy on the other hand, a significant increase in the two strongest paternalistically restrictive conditions (C4 and C5) indicates that the participants went through a belief updating process, adapting their initial underestimation of the system's autonomy. Similar effects were found for these two conditions for the reactance measures, which significantly increased. This shows that highly paternalistic AI systems elevate perceptions of system agency but do not provoke the expected reactance response, suggesting the presence of a reactance deficit for the two strongest paternalistic conditions of this study.
Qualitative data further indicates a more reflected engagement with the topic for participants who interacted with the AI system compared to the participants who interacted with the website without the influence of the AI system (C1). Participants exposed to AI paternalism (C2-C5) demonstrated more thoughtful engagement with the ethical and functional implications of AI in decision-making. Yet, post-intervention sentiment remained ambivalent as participants expressed concerns about manipulation and overreliance on AI, while also acknowledging its utility for optimized decision-making. This provides promising incentives for experience-based learning approaches to ensure individuals are psychologically and cognitively equipped to recognize and resist manipulative system behaviors.
From a theoretical perspective, these findings challenge the linear assumptions of PRT for AI-imposed restrictions and they contribute to an emerging body of work at the intersection of human-technology interaction and AI ethics by providing empirical support for the presence of a reactance deficit and stressing the societal necessity for protecting democratic values and human autonomy in an AI-driven world. The thesis concludes by advocating for further interdisciplinary investigation into the mechanisms of belief updating, autonomy perception, and strategies for fostering awareness in human-AI interaction.