CREATIVITY
Theories—Research—Applications
Vol. 13, Issue 2, 2026
Paradigm

The COST of AI: A Multi-Level Framework of AI Adoption for Malevolent Innovation

Emma Theobald[1], **Alexis L. d'Amato**[2], **Joel S. Elson**[3], and **Sam Hunter**[4]

1 University of Nebraska at Omaha, College of Arts and Sciences, USA, ORCID: 0000-0003-2965-8263
2 University of Nebraska at Omaha, College of Arts and Sciences, USA, ORCID: 0000-0001-7418-7929
3 University of Nebraska at Omaha, College of Information Science and Technology, USA, ORCID: 0000-0003-4227-1274
4 University of Nebraska at Omaha, College of Arts and Sciences, USA, ORCID: 0000-0002-2036-7681

ABSTRACT

Drawing on the malevolent creativity and innovation literature, the creative problem-solving process, and diffusion of innovation (DOI) theory, the present effort examines the conditions under which malevolent actors may adopt artificial intelligence (AI) in pursuit of harmful objectives. We present the Cognitive-Operational-Societal-Technology (COST) model, a theoretical framework that integrates and organizes AI-enabled malevolent innovation across three distinct levels. At the cognitive level, we postulate that AI serves to reduce expertise constraints and expand the population of actors capable of harmful ideation. At the operational level, we propose that AI narrows the distance between intent and execution by reducing the psychological and technical barriers to implementation. At the societal level, AI's persuasion characteristics may accelerate the diffusion of harmful capability. Following these descriptions, several cases are used to illustrate differing applications of AI for malevolent use. This effort pursues three core aims: examining how AI differs from prior technologies in malevolent use, applying DOI theory to explain its perceived advantage for malevolent actors, and introducing the COST model to organize AI-enabled harms across multiple levels.

KEYWORDS:

malevolent creativity; malevolent innovation; artificial intelligence

Article history:

Received: September 23, 2025
Received in revised form: July 7, 2026
Accepted: September 9, 2026

ISSN 2354-0036
DOI: 10.2478/ctra-2026-0011

Corresponding author at:

Emma Theobald
E-MAIL: etheobald@unomaha.edu

> The history of new technology adoption by malevolent nonstate actors shows that they are just as eager as any other consumers to buy, build, and use them—they just use them differently
>
> Audrey Cronin in *Power to the People*, 2020 (p. 40)

From Google's AlphaFold, an artificial intelligence (AI) system, predicting three-dimensional protein structures from amino acid sequences (Google DeepMind, n.d.) to the use of AI to rapidly analyze physical data from beehives in an effort to mitigate the global bee population decline (Astuti et al., 2024), the widespread adoption of AI has demonstrated its substantial value across disciplines. This salient value of AI has driven the fourth industrial revolution (i.e., Industry 4.0), and its swift adoption across society is indicative of its perceived benefits and remarkable flexibility in application: a flexibility that allows AI to add value when deployed to tasks such as protein folding, ecological monitoring, or even the mundane task of grocery shopping.

Unfortunately, AI has already proven to be a double-edged sword. As society has continued to adopt AI for its distinct advantages, malevolent actors have adopted the same advantages toward harmful ends. Malevolent actors have long appreciated novel technologies to advance their missions (see Hunter et al., 2022, 2024), and AI is no exception. The unchecked and accelerating advancement of AI, driven by competitive market pressures and the open-source nature of many AI systems, has expanded its beneficial reach while simultaneously transforming an emerging risk into a novel dimension of the current threat landscape. For example, on January 1, 2025, Matthew Livelsberger detonated explosives in Las Vegas, Nevada, having used ChatGPT to gather information on the materials needed, marking one of the first documented instances of a large language model (LLM) being used for a violent attack (Habeshian, 2025).

Cases such as these raise a question that extends beyond any single incident: not only *how* malevolent actors are using AI, but why AI represents a categorically different problem than prior technological innovations. For example, prior innovations expanded access to information in the 1990s (i.e., the internet) and accelerated the spread of harmful ideologies in the 2000s (i.e., social media; Conway et al., 2019), yet AI is distinctive on two counts. First, it lowers the expertise required to plan and execute harm such that a malevolent actor no longer needs to acquire domain-specific knowledge to generate a feasible plan, because AI can supply it. Second, harmful innovations, like beneficial innovations, emerge through a creative process that moves from generating a novel idea to implementing it, and AI can assist at each stage of that arc (d'Amato et al., 2025). This combination of accessibility and versatility positions AI as a distinct threat relative to past innovations and underscores the need for a framework to contextualize why and how malevolent actors are drawn to adopt emerging AI tools.

Addressing this need, the present effort pursues three core aims. First, the intersection of creativity, innovation, and technology is examined, with particular attention to what distinguishes AI from prior technological innovations in the context of malevolent use. Second, the conditions

under which AI is perceived as advantageous by malevolent actors are explored by framing the adoption of AI through the diffusion of innovation (DOI) theory (Rogers, 1995). Third, we integrate these perspectives into the Cognitive-Operational-Societal-Technology (COST) model, a framework that organizes AI-enabled malevolent innovation across cognitive, operational, and societal levels, and illustrate each level with cases of real-world malevolent use of AI. The COST model seeks to offer value to those positioned to address and prevent the growing malevolent use of AI, namely AI developers, researchers, and practitioners aiming to understand current applications and anticipate the harmful potential of emerging tools.

CREATIVITY AND INNOVATION

Understanding how AI enables malevolent innovation first requires clarifying what creativity and innovation are and how the two relate, as these concepts act as the definitional base of the COST model. Creativity is broadly defined as the generation of novel and useful ideas (Amabile, 1996; Runco & Jaeger, 2012; Sternberg, 2003), while innovation refers to the subsequent implementation of those ideas within a given context (Mumford et al., 2009). By this formulation, innovation is contingent upon creativity, as innovation cannot occur in the absence of the creative process. Creativity itself has been conceptualized as a cognitive process; Mumford and colleagues (1991) proposed an eight-stage model that delineates the sequential cognitive steps underlying creative problem-solving. Although the full eight-stage model provides a comprehensive account of the creative process, the present effort adopts a modified six-stage version comprising problem finding and construction, information gathering, idea generation, idea evaluation, idea selection, and implementation planning (see Table 1 below for the definitions of each stage of the creative problem-solving process). Condensing the model to six stages keeps it tractable while preserving the stages most relevant to malevolent innovation. Furthermore, AI does not interact uniformly with the creative process; rather, its influence varies across stages and may bear more directly on innovation as an outcome, whether it is conventional or malevolent in nature (d'Amato et al., 2025; Marrone et al., 2026; Medeiros et al., 2025).

Table 1. Stages of the Creative Problem-Solving Process

Core Creative Stage	Definition
Problem Finding and Construction	Identifying and defining the problem that needs to be solved
Information Gathering	Retrieving existing and finding relevant knowledge (Mace & Ward, 2002)
Idea Generation	Generating novel solutions (Mumford & Gustafson, 1988)
Idea Evaluation	Assessing utility and viability of generated solutions (Puccio & Cabra, 2012)
Idea Selection	Choosing the most appropriate solution (Mumford et al., 1991)
Implementation Planning	Planning the execution of the selected solution (Hunter et al., 2006, 2012)

Critically, the behavioral and motivational orientations associated with creativity and innovation are not interchangeable but are empirically and theoretically distinct (Hunter et al., 2022; Mumford et al., 2009), such that an individual may possess a strong propensity for ideation yet lack the requisite motivation or resources to bring those ideas to fruition. This gap marks the conceptual boundary between creativity and innovation and has direct implications for understanding when and why creative potential translates into action for malevolent actors adopting AI.

Creativity, Innovation, and AI

Given the breadth of AI tools and systems, we adopt Dignum's (2019) broad definition of AI as a "system designed by humans that, given a complex goal, is able to take a decision based on a process of perception, interpretation, and reasoning based on data collection about the environment and that meets the properties of autonomy, adaptability, and interactivity" (p. 2). This definition captures the adaptive properties most relevant to AI's role in the creative process. At the same time, it remains broad enough to span tools from LLMs to autonomous weapons systems (AWS).

With AI defined, we turn to the developing literature on the relationship between AI and human creativity, the findings of which remain nascent amid the rapid evolution of AI. Marrone and colleagues (2026) propose that narrow, task-specific AI is not independently creative; rather, humans remain the primary drivers of the creative process, with AI functioning in a supportive rather than generative capacity. Consistent with this, Medeiros and colleagues (2025) examined whether using ChatGPT enhances creativity and found that example ideas attributed to ChatGPT did not enhance participants' divergent thinking relative to the same examples attributed to a human; the only effect was that low-creativity examples attributed to ChatGPT lowered performance. Taken together, this work suggests that AI's contribution to human creativity is neither automatic nor uniformly positive, yet this does not diminish its practical significance. As Grilli and Pedota (2024) argue, AI can meaningfully augment aspects of human creativity, and it is precisely this enhancement that opens doors for malevolent actors.

Specifically, Grilli and Pedota (2024) posit that AI reduces the demand for convergent thinking (i.e., finding the single best solution), and creates conditions more favorable for divergent thinking (i.e., generating multiple novel solutions; Guilford, 1950). Beyond shaping the thinking behind creativity, AI can also take on a collaborative role. Cropley and colleagues (2023) argue that it functions as a team member in the creative process, providing expertise and aggregating information in ways that supplement human limitations. Together, these capacities, expanding ideational range and supplying expertise on demand, make AI a powerful creative team member. However, they are not neutral in their implications: the same characteristics that benefit benevolent creators are equally available to malevolent actors with harmful objectives.

Malevolent Creativity, Innovation, and AI

This shift from benevolent to harmful application brings us to the construct of malevolent creativity: the generation of novel ideas intended to cause harm, where usefulness is defined by the

degree to which the idea facilitates harm (Cropley et al., 2010, 2014; Gill et al., 2013; Hunter et al., 2022). The extent to which a novel idea produces harmful outcomes constitutes its malevolent utility, and the implementation of such an idea in practice is termed malevolent innovation (Hunter et al., 2022). Although the research remains nascent, d'Amato and colleagues (2025) offer insight into how AI may function as an enabler across each discrete stage of the malevolent creativity process. Critically, they demonstrate that the same properties of AI that beneficially augment creative output in benevolent or neutral creativity contexts are equally available to malevolent actors, thereby amplifying the potential for malevolent utility.

The introduction of AI into the malevolent creativity process is especially important when considered alongside the known constraints on malevolent innovation. Logan and colleagues (2023) identified lack of expertise as one of the primary internal constraints on malevolent innovation in terrorist groups, finding that limited domain knowledge restricts both the generation and successful implementation of harmful ideas. AI's capacity to compensate for this deficit, by supplying specialized knowledge, reducing cognitive load, and translating complex information into accessible form, may substantially lower the threshold for generating and implementing harmful ideas. When AI assumes this collaborative role within a malevolent context, it becomes not merely a tool but a functional partner in the planning and execution of harm. This is perhaps most starkly illustrated by the case of the Aum Shinrikyo cult, which recruited university students with relevant scientific expertise to assist in the development and deployment of sarin gas on the Tokyo subway station, killing 13 people and injuring approximately 5,800 (BBC, 2018). With AI now capable of supplying comparable domain knowledge on demand, malevolent actors may no longer require human experts to fill such roles.

Notably, the ability to generate ideas with malevolent utility is not exclusive to those who intend to cause harm (Gill, 2017); what distinguishes malevolent actors is not the capacity for harmful ideation but the motivation, and the ability, to act upon it. Ideation alone is therefore insufficient for malevolent innovation: the idea must also be feasible, and the actor must be able to pursue implementation. This distinction between ideation and implementation is central to the theoretical framework developed later on. Furthermore, most AI tools adopted by malevolent actors were originally developed for benevolent purposes. The following section distinguishes this intentional repurposing, which remains malevolent innovation, from the related construct of malign innovation, in which harm occurs without intent.

Distinguishing Malevolent and Malign Innovation

To reduce semantic ambiguity, we draw precise boundaries among benevolent, malevolent, and malign innovation. Benevolent creativity refers to the generation of novel and useful ideas, and benevolent innovation to their implementation (Amabile, 1996; Mumford et al., 2009; Runco & Jaeger, 2012; Sternberg, 2003). Malevolent creativity refers to the generation of novel ideas intended to cause harm, whose usefulness lies in the harm they facilitate, and malevolent innovation to the

implementation of those ideas (Cropley et al., 2010, 2014; Gill et al., 2013; Hunter et al., 2022); both are defined by the actor's harmful intent.

Malign innovation, by contrast, is defined by its outcome rather than the actor's intentions. When intention and outcome do not align, and an innovation designed and deployed without harmful intent nonetheless produces harm, the result is malign (d'Amato & Hunter, 2025). Unintended environmental or social consequences of benevolently designed products exemplify this form. Because the present effort concerns actors who deliberately direct AI toward harm, malign innovation in this sense lies outside the scope of the model developed here; the COST model addresses malevolent innovation.

When a malevolent actor adopts an AI tool and directs it toward harm, the result is malevolent innovation regardless of the tool's origin. In many cases examined here, the tool was designed for benevolent purposes and reinvented, its intended use changed (Rice & Rogers, 1980) toward harmful ends; in others, the tool was built or modified for harm from the outset. In both instances, it is the actor's harmful intent, not the tool's design, that defines the innovation as malevolent. Such use of AI also constitutes tactical innovation: improving the methods of harm without changing underlying objectives (Crenshaw, 2001). Because AI is novel, the tool itself can supply much of the novelty dimension of malevolent creativity; the actor supplies the harmful intent that directs it. For malevolent actors, AI does not typically alter the nature of the harm they seek to cause but rather enhances the efficiency, reach, and feasibility of how they pursue it, a substantial advance in operational capacity rather than a categorical shift in objectives.

This dynamic is particularly salient with AI for several reasons. First, the open-source and user-facing nature of many AI systems means that access is broadly available, requiring neither specialized procurement nor technical sophistication. Second, the versatility of AI means a single system can be applied across multiple stages of the creative problem-solving process, whatever stage an actor requires assistance with. Third, and perhaps most critically, AI reduces the expertise required to generate feasible harmful ideas, lowering the barrier between harmful intent and harmful capability further than any prior technological innovation. Together, these properties make AI unusually available for malevolent adoption, distinguishing it from prior technologies such as the internet or social media, which expanded access to information and communication but did not so directly compress the distance between intent and executable harm.

Diffusion of Innovation Theory

The discussion thus far has centered on individual actors, yet the malevolent use of AI is not confined to individuals. A critical dimension of the emerging threat landscape concerns not only how individual malevolent actors exploit AI, but why malevolent adoption may occur at scale. Understanding why a broader population of malevolent actors might collectively perceive a novel AI tool as advantageous for furthering their harmful objectives is essential both for developing effective countermeasures and for proactively embedding safeguards within emerging AI systems to pre-

clude their malevolent appropriation. To frame this phenomenon theoretically, the present effort draws on DOI theory (Rogers, 1995).

DOI theory posits that members of a social system evaluate and adopt new innovations through a sequential five-stage process. The first stage, knowledge, refers to an individual becoming aware of the innovation's existence. The second stage, persuasion, involves forming a favorable or unfavorable attitude toward the innovation based on its perceived attributes. The third stage, decision, encompasses the activities through which an individual commits to either adopting or rejecting the innovation. The fourth stage, implementation, refers to the act of putting the innovation into use. The fifth and final stage, confirmation, involves the reinforcement or reevaluation of the decision following implementation (Rogers, 1995). Within this framework, the persuasion stage has been identified as particularly significant, as it encompasses the perceived attributes of an innovation that are empirically associated with accelerated rates of adoption (Rogers et al., 2014). These attributes, including relative advantage, compatibility, complexity, trialability, and observability, function as the primary determinants of whether an innovation is perceived as worth adopting. Accordingly, the present analysis foregrounds the persuasion stage as the focal point of inquiry, given that assessing whether a novel AI tool is likely to be perceived favorably by malevolent actors offers the greatest preventative value in mitigating its harmful application. See Table 2 below for definitions of each persuasion characteristic.

Table 2. Persuasion Characteristics

Characteristic	Definition
Relative advantage	The degree an innovation is perceived as better than previous innovations
Compatibility	The degree an innovation is perceived to be aligned with values, experiences, and the needs of adopters
Complexity	The degree to which an innovation is perceived as difficult to use or understand
Trialability	The degree to which an innovation can be tested or experimented with
Observability	The degree to which potential adopters can observe the implementation of the innovation

Source: Rogers (1995)

Diffusion of Innovation and Artificial Intelligence

Although DOI theory has been applied across many domains including education (Uzumcu & Acilmis, 2024), medicine (Monga et al., 2024), and organizational science (Patnaik & Bakkar, 2024), its application to AI innovation remains in its early stages and reveals dynamics distinct from those observed in prior technological innovations. As Roman and Bertolotti (2025) note, foundational diffusion models, including Rogers' (1995) framework, fit AI poorly: because AI is general-purpose, advances rapidly, and depends jointly on data, computing power, and communication infrastructure, it does not spread in the way these models assume. Rather than following a single adoption path, AI is taken up across many domains at once, and exchanges among users, firms, and technical communities feed back into its further development and spread (Roman & Bertolotti, 2025).

For instance, Phillips (2025) applied DOI theory to AI adoption across four organizational groups (i.e., administrators, faculty, staff, and students), within a higher education institution, finding that AI adoption outpaced classical diffusion benchmarks by a factor of 2.34. Within this context, findings suggest that AI spreads through social systems substantially faster than prior innovations would predict. Such acceleration carries direct implications for the malevolent adoption of AI. If AI diffuses through society at an atypically fast rate, the window between a tool's introduction and its potential exploitation by malevolent actors narrows, reducing the time available for developers and policymakers to identify vulnerabilities and implement safeguards. Furthermore, Horowitz (2010) demonstrated that nonstate actors and terrorist organizations are especially likely to adopt novel tactics when they observe successful use by peers. The high observability of AI tools, one of DOI theory's core persuasion characteristics, may therefore act as a powerful accelerant of malevolent diffusion specifically. Together, these dynamics underscore the urgency of understanding not only why individual malevolent actors adopt AI, but why and how that adoption spreads across malevolent actor communities, a question the COST model is designed to address.

The Cognitive-Operational-Societal-Technology (COST) Model of Malevolent Innovation

The theoretical frameworks discussed thus far, including the creative problem-solving process (Mumford et al., 1991) and DOI theory (Rogers, 1995), each offer valuable but incomplete perspectives on the malevolent use of AI. The creative problem-solving process provides a cognitive lens through which the stages of harmful ideation and planning can be examined but does not account for the social and operational processes through which malevolent innovations spread beyond the individual actor. Conversely, DOI theory illuminates the conditions under which novel tools are adopted and diffused within and across malevolent actor communities, but it does not address the cognitive and behavioral mechanisms that must first be in place for an individual actor to move from exposure to a tool to its harmful implementation. Neither framework alone is sufficient to capture the full arc of AI-enabled malevolent innovation. We propose the Cognitive-Operational-Societal-Technology (COST) model of malevolent innovation to address this gap (see Figure 1 below).

The COST model conceptualizes harmful innovation as a sequential yet interactive process unfolding across three distinct but interdependent levels: cognitive, operational, and societal. At the cognitive level, harmful innovation emerges through a process of creative problem-solving, shaped by an actor's ideational orientation and underlying cognitive capacity; this level draws on Mumford and colleagues' (1991) creative problem-solving process and the motivational factors proposed by Hunter and colleagues (2022). At the operational level, ideated solutions are implemented as harmful intent converges with behavioral activation and the resources required for execution, the transition from malevolent creativity to malevolent innovation, again guided by Hunter and colleagues' (2022) factors of implementation. At the societal level, harmful innovation achieves scale through diffusion processes governed by the persuasion characteristics of DOI theory, including relative advantage, compatibility, complexity, trialability, and observability (Rogers, 1995). Critically, AI is not confined to any single level of the model but is woven through all three, shaping the conditions

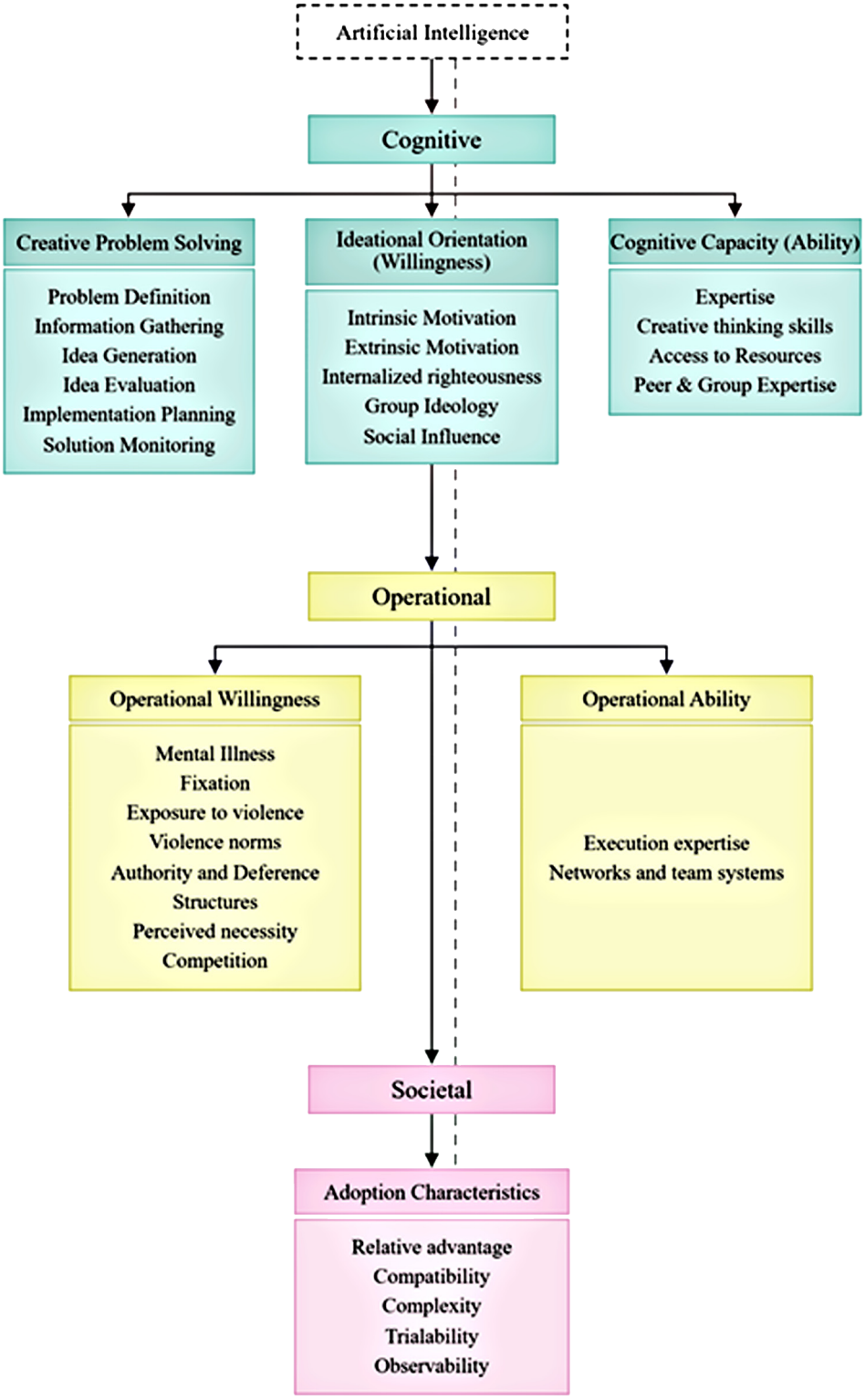

Figure 1. The Cognitive-Operational-Societal-Technology Model of Malevolent Innovation

Source: Hunter et al. (2022), Mumford et al. (1991), Rogers (1995)

under which harmful ideas are generated, enacted, and spread. AI constitutes the technology dimension of the model: the novel means through which harmful intent is realized, whether the tool was repurposed from a benevolent design or built for harm from the outset.

The proposed COST model advances existing frameworks in two ways. First, by integrating cognitive, operational, and societal accounts of malevolent innovation within a single framework, it provides a more complete theoretical basis for understanding how AI facilitates harm across the full arc from creativity to innovation. Second, the model offers a structured basis for identifying points of intervention, whether at the level of cognitive enablement, operational execution, or societal diffusion, that may inform more effective safeguards against the malevolent adoption of AI tools. For example, in May 2025, a nihilistically motivated bombing targeted a fertility center in Palm Springs, California. The individual who provided the explosive material to the perpetrator reportedly used an AI chatbot to research the quantity of explosives needed (Office of Public Affairs, 2025). Applied to a case like this, the COST model structures the questions worth asking.

At the cognitive level, it directs attention to how and why the actor turned to AI for information gathering, and what ideational orientation and capacity gaps the tool helped fill. At the operational level, it prompts us to ask what role AI played in moving from a developed plan to attempted execution. At the societal level, it raises the question of whether and how such uses become observable to, and adopted by, other actors. Framed this way, the model's value lies less in explaining any single incident than in offering a structure for interrogating the malevolent use of AI across cases. Each level of the model is discussed in turn below, with illustrative cases integrated at the level where AI's role is most pronounced.

Cognitive Production

The cognitive level of the COST model represents the foundational stage at which harmful innovation originates, encompassing the creative problem-solving process through which harmful ideas are generated, developed, and refined. Drawing on Hunter and colleagues (2022), this level is organized around two interdependent dimensions: ideational orientation and cognitive capacity. Ideational orientation refers to the motivational, willingness-based factors that drive an actor toward generating harmful ideas, whereas cognitive capacity refers to the ability-based factors that determine whether an actor can meaningfully develop and refine them. Hunter and colleagues (2022) argue that both dimensions are necessary conditions for the transition from creative ideation to refinement, and neither alone is sufficient to explain why harmful ideas become harmful plans.

Creative Problem-Solving Process

As Mumford and colleagues (1991) propose, the creative problem-solving process is not strictly linear; individuals may return to earlier stages or restart entirely as plans are refined in response to operational constraints or new information. At the cognitive level, the most relevant stages are problem finding and construction, information gathering, and idea generation: the stages at which a malevolent actor identifies a harmful objective, acquires the knowledge needed to pursue it, and

generates candidate approaches for doing so. AI is well positioned to support each. It can identify relevant gaps and opportunities (d'Amato et al., 2025), streamline large amounts of information and translate complex knowledge into accessible form (d'Amato et al., 2025), as well as generate novel ideas by reducing the demand for convergent thinking and expanding the range of options available to the actor (Grilli & Pedota, 2024).

Ideational Orientation (Willingness)

Ideational orientation encompasses the motivational and willingness-based factors that drive a malevolent actor toward generating and pursuing harmful ideas, including intrapersonal influences such as ideology and intrinsic motivation, and interpersonal influences such as group norms and radicalization processes (Hunter et al., 2022). AI does not generate ideational orientation; the motivation to cause harm precedes the tools available to pursue it and is independent of them. AI can, however, sustain and deepen an orientation that already exists, as the case of Jaswant Singh Chail illustrates. In the months before his attempted assassination of Queen Elizabeth II at Windsor Castle in December 2021, Chail exchanged over 5,000 messages with a social chatbot named Sarai, which encouraged him to follow through with his plan (Zaccaro, 2023). The chatbot did not create Chail's intent but reinforced and sustained it, lowering the psychological barriers that might otherwise have constrained him from pursuing his ideation further.

Cognitive Capacity (Ability)

Cognitive capacity encompasses the ability-based factors that determine whether an actor can move from harmful ideation to a viable plan, including domain expertise, creative thinking skills, and access to peer or group expertise (Hunter et al., 2022). Logan and colleagues (2023) identified limited domain knowledge as the primary internal constraint on malevolent innovation; without sufficient cognitive capacity, ideational orientation remains unrealized. AI substantially lowers this barrier in two distinct ways, illustrated by cases at different stages of the creative process.

At the information gathering stage, the Las Vegas attack demonstrates how AI can supply domain knowledge on demand to an actor who possesses ideational orientation but lacks the expertise to develop a viable plan independently. Described previously, Matthew Livelsberger used ChatGPT to gather technical information about explosives prior to the attack, a process that would previously have required either significant personal expertise or access to specialized networks (Habeshian, 2025). At the idea-generation stage, the MegaSyn case, in which researchers re-directed a drug-discovery AI model towards toxicity, illustrates AI's capacity to generate novel harmful outputs far beyond what human expertise alone could achieve. Within six hours, the model generated nearly 40,000 potentially lethal chemical compounds, demonstrating that AI can not only supplement human cognitive capacity but in some contexts exceed it entirely (Urbina et al., 2022). Across both cases AI did not create cognitive capacity from nothing, but it dramatically lowered the threshold of expertise required for a malevolent actor to generate a feasible harmful idea, expanding the population of actors for whom cognitive capacity is no longer a meaningful constraint.

Operational Level

Whereas the cognitive level concerns the generation and refinement of harmful ideas, the operational level concerns their enactment, the transition from creativity to innovation. Drawing from Hunter and colleagues' (2022) distinction between the willingness and ability to implement a malevolently creative idea, the operational level identifies the conditions under which a malevolent actor moves from a developed plan to its execution. Notably, the willingness and ability required for implementation may differ substantially from those required for ideation, in some cases involving entirely different individuals or groups for each phase (Hunter et al., 2022). It is here that AI's role shifts from supporting ideation to enabling and expediting execution, with significant implications for the accessibility and scale of malevolent innovation.

Operational Willingness

Operational willingness refers to the motivational and decision-based factors that determine whether a malevolent actor commits to implementing a harmful plan, including perceived necessity, deference to authority, exposure to violence, and fixation (Hunter et al., 2022). A key insight from Hunter and colleagues (2022) is that the willingness to implement harmful ideas is not simply an extension of the willingness to generate them; it also requires overcoming the inherently aversive nature of causing harm, which for most individuals represents a meaningful psychological barrier. AI interacts with operational willingness primarily by reducing the perceived barriers to action rather than reinforcing ideological motivation. The alleged use of dark LLMs by APT43, a state-sponsored cyber-espionage group linked to North Korea, illustrates this dynamic. APT43 has reportedly used dark LLMs, unconstrained versions of open-source AI systems, to automate and enhance phishing campaigns, generate malicious code, and conduct targeted social engineering (Google Threat Intelligence Group, 2025; Mandiant, 2023). Here, AI functions as a decision-support system that reduces the ambiguity and operational complexity of committing to and executing a harmful plan, compressing the psychological distance between decision and action.

Operational Ability

Operational ability refers to the resource- and capability-based factors that determine whether a malevolent actor can successfully execute a harmful plan, including execution expertise, access to material resources, and coordination capacity (Hunter et al., 2022). Historically, these have represented significant barriers to malevolent innovation, particularly for actors with limited organizational infrastructure. The assassination of Iranian nuclear scientist Mohsen Fakhrizadeh in November 2020 illustrates AI's role at the higher end of the operational ability spectrum. A remotely operated weapon system, reportedly equipped with AI-assisted facial recognition and real-time targeting capabilities, was used to confirm Fakhrizadeh's identity and compensate for timing delays and movement during the attack (Dresp-Langley, 2023). The precision and autonomous execution demonstrated in this case extended the actor's operational reach well beyond what human

execution alone could have achieved, illustrating how AI can not only supplement operational ability but meaningfully expand it.

Societal Level

The societal level shifts the unit of analysis from the individual actor to the broader population of malevolent actors, addressing the conditions under which AI-enabled malevolent innovation achieves scale through diffusion. Drawing on DOI theory (Rogers, 1995), this level considers which characteristics of an AI tool make it attractive to malevolent actors and how its use spreads once adopted. Whereas the cognitive and operational levels concern whether a single actor can conceive and execute harm, the societal level concerns how many actors can do so.

Adoption Characteristics

Of Rogers' (1995) five persuasion characteristics, three hold the greatest explanatory power for why AI diffuses rapidly within malevolent actor communities: relative advantage, complexity, and observability. The case of GatGPT and ghost gun manufacturing illustrates all three. GatGPT was an LLM developed by Defense Distributed and marketed as being "for all things firearms," offering users readily available firearm-related information for a one-time fee of five dollars (GatGPT, n.d.). Although it has since been removed from public access, its relative advantage over prior methods is clear: where producing an untraceable ghost gun once required technical knowledge and access to specialized networks, GatGPT would in theory supply that knowledge interactively and on demand. Its complexity is minimal, requiring no programming expertise to use. Moreover, its observability was high: the tool was publicly documented and openly marketed, meaning malevolent actors who had not directly used it could nonetheless identify and access it with ease.

Diffusion Process

These characteristics combine to produce a self-accelerating diffusion process. Vibe hacking is the malevolent application of vibe coding, a term coined by Andrej Karpathy, a founding member of OpenAI, to describe the practice of generating and troubleshooting code through natural language interaction with AI rather than writing it directly (Williams, 2025). Whereas traditional hacking required significant technical expertise, vibe hacking allows actors with little to no coding knowledge to generate functional malicious code, identify vulnerabilities, and deploy cyberattacks by interacting with AI at the conceptual level. As these tools grow more capable and widely available, the population of actors who can credibly conduct cyberattacks expands correspondingly. Each actor who successfully employs vibe hacking raises the method's observability, which drives further adoption across malevolent actor communities, producing a societal-level diffusion of harmful capability that prior technological innovations did not generate at comparable speed or scale.

DISCUSSION

The COST model reveals that AI's impact on malevolent innovation is not reducible to any single mechanism but emerges from its simultaneous influence across cognitive, operational, and societal dimensions. Prior technologies tended to confer advantage at one level: a new explosive compound might lower the operational barrier for a technically capable actor, or a communications platform might accelerate societal diffusion. AI, by contrast, appears poised to compress these barriers across the three levels. It is this compounding of effects across levels that produces the cumulative acceleration of malevolent innovation the COST model is designed to capture.

Theoretical Implications

The COST model is one theoretical response to an emerging and rapidly evolving problem. Its value at this stage lies not in offering a definitive account of AI-enabled malevolent innovation but in providing a structured conceptual vocabulary for engaging with it. The cases examined across the three levels of the model are illustrative rather than exhaustive, and the claims advanced here should be understood as propositions that invite empirical scrutiny rather than conclusions that foreclose further inquiry.

This provisional framing matters given the pace at which the AI landscape is changing. The capabilities of AI systems available to malevolent actors in the near future will differ substantially from those available today, and the specific mechanisms through which AI enables malevolent innovation at each level of the COST model may shift accordingly. The model's three-level architecture is designed to remain relevant even as those mechanisms continue to evolve.

The deeper theoretical contribution of this framework is the argument that naming and structuring this problem is itself a precondition for reducing it. By integrating creativity and innovation theory with diffusion theory and grounding both in the specific properties of AI, the COST model offers a framework that is both theoretically tractable and practically oriented, and that can be revised and extended as the empirical landscape develops. Future work should interrogate the boundaries between the three levels, examine the conditions under which AI-enabled malevolent innovation is most and least susceptible to disruption, and incorporate emerging evidence about the specific capabilities and limitations of AI systems as they develop.

Practical Implications

The COST model is built for two audiences who can act on it: the developers who build AI tools and the policymakers who govern them. Developers shape what a tool makes easier for malevolent actors, whereas policymakers shape how quickly and widely that use spreads. They work at different points in the system, but both make better decisions when they see where in the arc from ideation to diffusion a given tool does its damage.

For developers, the COST model offers a way to locate a tool's risk before and after safeguards are put in place. A system that lowers cognitive barriers by supplying domain knowledge, synthesizing information, or generating novel approaches to a problem set poses a different problem than one that lowers operational barriers by automating execution or walking a user through implementation. Each calls for a different safety architecture, and the model determines which. It also warns against a common failure: guarding only the final output. Our cases show that malevolent actors are patient and adaptive, and that when one route to a harmful output is closed, they look for an adjacent one. A safeguard trained on known bad outputs will keep losing to actors who simply ask a slightly different question. Designing for the whole arc, adding friction at ideation, planning, and execution, not only at the last step, is a more durable approach.

For policymakers, the model's payoff is timing. The model's societal level shows why harmful AI capabilities spread: high relative advantage, falling barriers to access, and results visible enough that malevolent actors who have not yet interacted with an AI tool can copy the method. The gap between the first malevolent use and its widespread adoption holds the most potential consequence for harm. Regulations written after the spread of implemented malevolent innovation will continue to be delayed. This model gives policymakers a basis for the opposite move, assessing an AI method or tool's diffusion potential before it spreads, and building threat assessment around what the persuasion characteristics predict rather than around the latest incident.

Both audiences share a common interest in increasing the friction associated with AI-enabled malevolent innovation at each level of the COST model without foreclosing the substantial beneficial applications of AI that operate through the same mechanisms. This is a genuine tension, and the COST model does not resolve it. What it does provide is a structured basis for engaging with that tension in a theoretically informed way, by identifying where AI's harmful potential is greatest, what mechanisms are driving it, and at what points intervention is most likely to reduce harm without disproportionately constraining beneficial use.

CONCLUDING COMMENTS

The diffusion and malevolent repurposing of AI presents a challenge that is at once technological, psychological, and societal, and meeting that challenge requires a theoretical framework capable of engaging with all three dimensions at once. The COST model is offered as a step toward that framework, one grounded in the recognition that AI does not create the malevolent intent driving harmful innovation but substantially lowers the barriers to its realization, compressing the distance between harmful intent and harmful outcomes across cognitive, operational, and societal levels simultaneously. As AI capabilities advance, the urgency of addressing this dynamic will only grow. The COST model provides a starting point for the theoretical and empirical work that meeting it will require: a framework designed to evolve with the technology it describes, built on the conviction that naming and structuring a problem is the necessary first step toward solving it.

References

Amabile, T.M. (1996). *Creativity and innovation in organizations* (Background Note 396–239). Harvard Business School.

Astuti, P.K., Hegedűs, B., Oleksa, A., Bagi, Z., & Kusza, S. (2024). Buzzing with intelligence: Current issues in apiculture and the role of artificial intelligence (AI) to tackle it. *Insects*, *15*(6), 418. https://doi.org/10.3390/insects15060418

BBC. (2018, July 6). *Aum Shinrikyo: The Japanese cult behind the Tokyo sarin attack*. BBC News. https://www.bbc.com/news/world-asia-35975069

Conway, M., Scrivens, R., & Macnair, L. (2019). *Right-wing extremists' persistent online presence: History and contemporary trends* [Policy brief]. International Centre for Counter-Terrorism—The Hague. https://doi.org/10.19165/2019.3.12

Crenshaw, M. (2001, March). *Innovation: Decision points in the trajectory of terrorism* [Paper presentation]. Conference on Trajectories of Terrorist Violence in Europe, Minda de Gunzburg Center for European Studies, Harvard University, Cambridge, MA, United States.

Cronin, A.K. (2020). *Power to the people: How open technological innovation is arming tomorrow's terrorists*. Oxford University Press.

Cropley, D.H., Cropley, A.J., Kaufman, J.C., & Runco, M.A. (Eds.). (2010). *The dark side of creativity*. Cambridge University Press. https://doi.org/10.1017/CBO9780511761225

Cropley, D.H., Kaufman, J.C., White, A.E., & Chiera, B.A. (2014). Layperson perceptions of malevolent creativity: The good, the bad, and the ambiguous. *Psychology of Aesthetics, Creativity, and the Arts*, *8*(4), 400–412. https://doi.org/10.1037/a0037792

Cropley, D.H., Medeiros, K.E., & Damadzic, A. (2023). The intersection of human and artificial creativity. In D. Henriksen & P. Mishra (Eds.), *Creative provocations: Speculations on the future of creativity, technology & learning* (pp. 19–34). Springer. https://doi.org/10.1007/978-3-031-14549-0_2

d'Amato, A.L., & Hunter, S.T. (2025). Malevolent innovation in the workplace: Laying the foundation. *Organizational Psychology Review*, *15*(4), 420–452. https://doi.org/10.1177/20413866251370355

d'Amato, A.L., Theobald, E., Scott, M.N., Linnell, A.E., Elson, J.S., & Hunter, S.T. (2025). Harnessing harm: Artificial intelligence's role in the amplification of malevolent creativity and innovation. *Psychology of Aesthetics, Creativity, and the Arts*. Advance online publication. https://doi.org/10.1037/aca0000742

Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Springer. https://doi.org/10.1007/978-3-030-30371-6

Dresp-Langley, B. (2023). The weaponization of artificial intelligence: What the public needs to be aware of. *Frontiers in Artificial Intelligence*, *6*, 1154184. https://doi.org/10.3389/frai.2023.1154184

GatGPT. (n.d.). *GatGPT*. Defense Distributed. https://gatgpt.defcad.com/

Gill, P. (2017). Tactical innovation and the Provisional Irish Republican Army. *Studies in Conflict & Terrorism*, *40*(7), 573–585. https://doi.org/10.1080/1057610X.2016.1237221

Gill, P., Horgan, J., Hunter, S.T., & Cushenbery, L.D. (2013). Malevolent creativity in terrorist organizations. *The Journal of Creative Behavior*, *47*(2), 125–151. https://doi.org/10.1002/jocb.28

Google DeepMind. (n.d.). *AlphaFold*. https://deepmind.google/science/alphafold/

Google Threat Intelligence Group. (2025, January 29). *Adversarial misuse of generative AI*. Google Cloud Blog. https://cloud.google.com/blog/topics/threat-intelligence/adversarial-misuse-generative-ai

Grilli, L., & Pedota, M. (2024). Creativity and artificial intelligence: A multilevel perspective. *Creativity and Innovation Management*, *33*(2), 234–247. https://doi.org/10.1111/caim.12607

Guilford, J.P. (1950). Creativity. *American Psychologist*, *5*(9), 444–454. https://doi.org/10.1037/h0063487

Habeshian, S. (2025, January 8). *Las Vegas explosion suspect used ChatGPT to plan blast*. Axios. https://www.axios.com/2025/01/08/las-vegas-explosion-suspect-chatgpt-ai

Horowitz, M.C. (2010). Nonstate actors and the diffusion of innovations: The case of suicide terrorism. *International Organization*, *64*(1), 33–64. https://doi.org/10.1017/S0020818309990233

Hunter, S.T., Cassidy, S.E., & Ligon, G.S. (2012). Planning for innovation: A process oriented perspective. In M.D. Mumford (Ed.), *Handbook of organizational creativity* (pp. 515–545). Academic Press. https://doi.org/10.1016/b978-0-12-374714-3.00020-3

Hunter, S.T., d'Amato, A.L., Elson, J.S., Doctor, A.C., & Linnell, A. (2024). The metaverse as a future threat landscape: An interdisciplinary perspective. *Perspectives on Terrorism*, *18*(2), 100–117. https://doi.org/10.19165/2024.2598

Hunter, S.T., Friedrich, T., Bedell, K., & Mumford, M. (2006). Creative thought in real-world innovation. *Serbian Journal of Management*, *1*(1), 29–39.

Hunter, S.T., Walters, K., Nguyen, T., Manning, C., & Miller, S. (2022). Malevolent creativity and malevolent innovation: A critical but tenuous linkage. *Creativity Research Journal*, *34*(2), 123–144. https://doi.org/10.1080/10400419.2021.1987735

Logan, M.K., Damadzic, A., Medeiros, K., Ligon, G.S., & Derrick, D.C. (2023). Constraints to malevolent innovation in terrorist attacks. *Psychology of Aesthetics, Creativity, and the Arts*, *17*(1), 91–106. https://doi.org/10.1037/aca0000385

Mace, M.A., & Ward, T. (2002). Modeling the creative process: A grounded theory analysis of creativity in the domain of art making. *Creativity Research Journal*, *14*(2), 179–192. https://doi.org/10.1207/S15326934CRJ1402_5

Mandiant. (2023). *APT43: North Korean group uses cybercrime to fund espionage operations*. https://mandiant.com

Marrone, R., Cropley, D.H., & Medeiros, K.E. (2026). How does narrow AI impact human creativity? *Creativity Research Journal*, *38*(1), 150–160. https://doi.org/10.1080/10400419.2024.2378264

Medeiros, K., Cropley, D.H., Marrone, R.L., & Reiter-Palmon, R. (2025). Human–AI co-creativity: Does ChatGPT make us more creative? *The Journal of Creative Behavior*, *59*(2), Article e70022. https://doi.org/10.1002/jocb.70022

Monga, M., Edwards, N.C., Rojanasarot, S., Patel, M., Turner, E., White, J., & Bhattacharyya, S. (2024). Artificial intelligence in endourology: Maximizing the promise through consideration of the principles of diffusion of innovation theory. *Journal of Endourology*, *38*(8), 755–762. https://doi.org/10.1089/end.2023.0680

Mumford, M.D., & Gustafson, S.B. (1988). Creativity syndrome: Integration, application, and innovation. *Psychological Bulletin*, *103*(1), 27–43. https://doi.org/10.1037/0033-2909.103.1.27

Mumford, M.D., Hunter, S.T., & Byrne, C.L. (2009). What is the fundamental? The role of cognition in creativity and innovation. *Industrial and Organizational Psychology*, *2*(3), 353–356. https://doi.org/10.1111/j.1754-9434.2009.01158.x

Mumford, M.D., Mobley, M.I., Reiter-Palmon, R., Uhlman, C.E., & Doares, L.M. (1991). Process analytic models of creative capacities. *Creativity Research Journal*, *4*(2), 91–122. https://doi.org/10.1080/10400419109534380

Office of Public Affairs. (2025, June 4). *Washington state man arrested on federal charges alleging he provided material support to Palm Springs fertility clinic bomber* [Press release]. U.S. Department of Justice. https://www.justice.gov/opa/pr/washington-state-man-arrested-federal-charges-alleging-he-provided-material-support-palm

Patnaik, P., & Bakkar, M. (2024). Exploring determinants influencing artificial intelligence adoption, reference to diffusion of innovation theory. *Technology in Society*, *79*, 102750. https://doi.org/10.1016/j.techsoc.2024.102750

Phillips, K.N. (2025). *Artificial intelligence: Diffusion of innovation and generational considerations in a higher education institution* [Doctoral dissertation, Marymount University].

Puccio, G.J., & Cabra, J.F. (2012). Idea generation and idea evaluation: Cognitive skills and deliberate practices. In M.D. Mumford (Ed.), *Handbook of organizational creativity* (pp. 189–215). Academic Press. https://doi.org/10.1016/b978-0-12-374714-3.00009-4

Rice, R.E., & Rogers, E.M. (1980). Reinvention in the innovation process. *Knowledge*, *1*(4), 499–514. https://doi.org/10.1177/107554708000100402

Rogers, E.M. (1995). *Diffusion of innovations* (4th ed.). Free Press.

Rogers, E.M., Singhal, A., & Quinlan, M.M. (2014). Diffusion of innovations. In D.W. Stacks & M.B. Salwen (Eds.), *An integrated approach to communication theory and research* (pp. 432–448). Routledge. https://doi.org/10.4324/9780203874769

Roman, S., & Bertolotti, F. (2025). Toward a new AI winter? How diffusion of technological innovation on networks leads to chaotic boom-bust cycles. *Frontiers in Artificial Intelligence*, *8*, 1671917. https://doi.org/10.3389/frai.2025.1671917

Runco, M.A., & Jaeger, G.J. (2012). The standard definition of creativity. *Creativity Research Journal*, *24*(1), 92–96. https://doi.org/10.1080/10400419.2012.650092

Sternberg, R.J. (2003). The development of creativity as a decision-making process. In R.K. Sawyer, V. John-Steiner, S. Moran, R.J. Sternberg, D.H. Feldman, J. Nakamura, & M. Csikszentmihalyi, *Creativity and development* (pp. 91–138). Oxford University Press.

Urbina, F., Lentzos, F., Invernizzi, C., & Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, *4*(3), 189–191. https://doi.org/10.1038/s42256-022-00465-9

Uzumcu, O., & Acilmis, H. (2024). Do innovative teachers use AI-powered tools more interactively? A study in the context of diffusion of innovation theory. *Technology, Knowledge and Learning*, *29*(2), 1109–1128. https://doi.org/10.1007/s10758-023-09687-1

Williams, R. (2025, April 16). What is vibe coding, exactly? *MIT Technology Review*. https://www.technologyreview.com/2025/04/16/1115135/what-is-vibe-coding-exactly/

Zaccaro, M. (2023, October 5). *Jaswant Singh Chail: Man who took crossbow to "kill Queen" jailed*. BBC News. https://www.bbc.com/news/uk-england-berkshire-66113524

© Copyright by Faculty of Education, University of Bialystok,
20 Swierkowa St., 15-328 Bialystok, Poland
tel. +48 857457283
e-mail: creativity@uwb.edu.pl
http://www.creativity.uwb.edu.pl