Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Clinician-AI Collaboration in Endovascular Neurointerventions: In-Vitro User Study of Copilot Assistance Modes for Robotic Navigation Across the Autonomy Spectrum

Nikola Fischer
King's College London
Harry Robertshaw
King's College London
S.M.Hadi Sadati
King's College London
Xingyu Chen
King's College London
Weiyuan Deng
King's College London
Benjamin Jackson
King's College London
Michael Adofo Kwakye
King's College London
Sara Sciacca
King's College Hospital
Naga Kandasamy
King's College Hospital
Parthiban Balasundaram
King's College Hospital
Sahan Guruge
King's College Hospital
Charles Cummins
King's College Hospital
Joe Leyon
King's College Hospital
Nasr Abdelsalam
The Walton Centre
Matthew Benger
St George's Hospital

Andrew Falzon
St George's Hospital
Alejandro Granados
King's College London
Christos Bergeles
King's College London
Thomas Booth
thomas.booth@kcl.ac.uk

King's College London

Article

Keywords: AI, human–machine, shared autonomy, autonomous navigation, zero-shot sim-to-real transfer

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11729073/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Clinician-AI Collaboration in Endovascular Neurointerventions: *In-Vitro* User Study of Copilot Assistance Modes for Robotic Navigation Across the Autonomy Spectrum

Nikola Fischer[1], Harry Robertshaw[1], S.M.Hadi Sadati[1,2], Xingyu Chen[1], Weiyuan Deng[1], Benjamin Jackson[1], Michael Adofo Kwakye[1], Sara Sciacca[3], Naga Kandasamy[3], Parthiban Balasundaram[3], Sahan Guruge[3], Charles Cummins[3], Joe Leyon[3], Nasr Abdelsalam[4], Matthew Benger[5], Andrew Falzon[5], Alejandro Granados[1], Christos Bergeles[1†], and Thomas Booth[1,3*†]

***Abstract*—Mechanical thrombectomy (MT) is an effective treatment for acute ischemic stroke, yet access remains limited by the availability of specialist neurointerventional expertise. Telerobotic and AI-assisted systems may help address this challenge, but it remains unclear how clinicians interact with different levels of autonomy and how autonomous navigation systems perform in clinically realistic environments. This work introduces Copilot Assistance Modes (CAMs), a framework for clinician-AI collaboration in robotic endovascular navigation spanning the autonomy spectrum from manual operation to full autonomy. In the first *in-vitro* user study systematically comparing clinician-AI collaboration strategies, 11 expert and two student participants evaluated six navigation modes in a patient-specific vascular phantom under fluoroscopic guidance. Objective performance metrics were combined with subjective assessments, including workload, usability, and operator preference. Autonomous performance was additionally evaluated both *in silico* and *in vitro* to investigate long-horizon sim-to-real transfer. Hands-on manual navigation achieved the strongest overall performance (100% success). While fully autonomous navigation achieved 81% success in simulation, performance dropped to 23% *in vitro*. In contrast, several clinician-AI shared-autonomy modes achieved success rates of 92–97%, approximately four times higher than full autonomy and comparable to direct robotic control (95%), potentially demonstrating mitigation to sim-to-real domain shift when using imperfect AI through clinician-AI collaboration. Among the evaluated AI-driven approaches, user-directed task delegation and rotational shared control demonstrated the most favorable balance between performance and clinician acceptance. These findings plausibly suggest that, for robotic neurointerventional navigation, effective clinician-AI collaboration and allocation of control authority may currently be more important for reliable performance than increasing autonomy alone.**

[1]School of Biomedical Engineering & Imaging Sciences, King's College London, UK; [2]School of Engineering and Materials Science, Queen Mary University of London, UK; [3]Department of Neuroradiology, King's College Hospital NHS Foundation Trust, London, UK; [4]The Walton Centre NHS Foundation Trust, Liverpool, UK; [5]St George's University Hospitals NHS Foundation Trust, London, UK
*Corresponding author: thomas.booth@kcl.ac.uk
†These authors contributed equally as senior authors.
This work was supported by the Medical Research Council (MRC) Impact Acceleration Account (IAA) 2021 at King's College (Grant No. MR/X502923/1), by EPSRC [VITRRO / EP/Y024281/1 - ERC Proof of Concept] under the Horizon Europe Guarantee Extension, the King's-funded Centre for Doctoral Training on Surgical & Interventional Engineering, and the King's-China Scholarship Council [CSC No.: 202408330072]. Work in the Surgical & Interventional Engineering Validation Suite (mock OR) was supported by core funding from the Wellcome/EPSRC Centre for Medical Engineering [WT203148/Z/16/Z], a multi-user equipment grant for post-mortem evaluation of medical devices from Wellcome [218286/Z/19/Z], and the Wolfson Foundation [PR/ylr/md/21896].

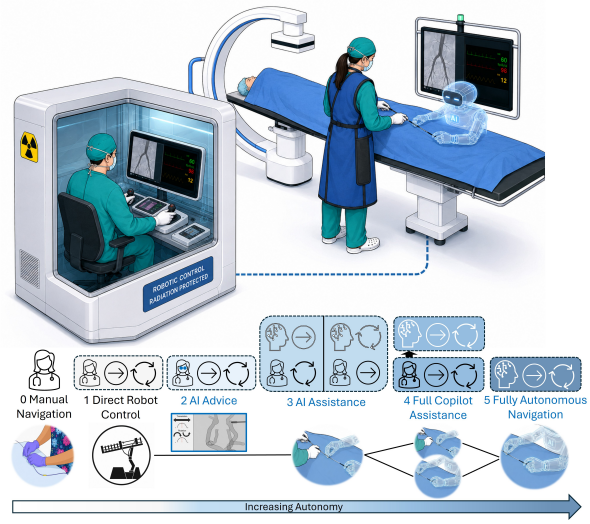


Figure 1. **Clinician-AI Collaboration for Neurointerventions:** Introducing the Copilot Assistance Modes (CAMs) including the conventional Manual (0) and Direct Robot Control (1) as well as the AI-driven modes (blue) AI Advice (2), AI Assistance (3), Full Copilot Assistance (4), and Fully Autonomous Navigation (5). The blue color gradient indicates increasing levels of autonomy, ranging from white (manual) over grey (teleoperation) to dark blue (highest autonomy).

***Index Terms*—human-AI, human–machine, shared autonomy, autonomous navigation, zero-shot sim-to-real transfer**

I. Introduction

STROKE is one of the leading causes of death and long-term disability worldwide, with a substantial socioeconomic burden on healthcare systems [1]. Acute ischemic stroke (AIS) remains a major global health burden, with approximately 12 million new cases and 7 million deaths annually [2]. Large vessel occlusion (LVO) strokes are particularly time critical, as reduced cerebral perfusion in a large territory accelerates considerable neuronal loss, making rapid recanalization essential for preserving neurological function and improving clinical outcomes. Mechanical thrombectomy (MT) has emerged as the reference standard treatment for eligible patients, significantly improving functional outcomes and reducing mortality compared to medical therapy alone [3]–[5].

Despite its clinical effectiveness, access to MT remains highly unequal. Because treatment success is critically dependent on minimizing the time from symptom onset to reperfusion, the limited availability of specialized neurointerventional centers and trained operators prevents many eligible patients from receiving timely treatment [6], thereby reducing the likelihood of favorable clinical outcomes [7]. The workforce shortage [8] is further exacerbated by the technical complexity of endovascular interventions, in which operators must navigate instruments (i.e., guidewires and catheters) through highly tortuous vasculature while achieving predictable control of the distal instrument tip through indirect manipulation of the two degrees of freedom (DoF) rotation φ and translation d at the proximal end. As a result, accurate navigation places considerable demands on clinician dexterity, hand-eye coordination, and procedural experience. To facilitate instrument handling, robotic (neuro)endovascular systems are emerging [9]. These systems enable precise teleoperated manipulation of endovascular instruments, potentially also allowing expert clinicians to perform procedures remotely from centralized stroke centers [10]. This paradigm may reduce geographic disparities by concentrating expertise while extending procedural reach to peripheral medical centers. In addition, robotic actuation may reduce radiation exposure and improve operator ergonomics [11], [12], even for clinicians performing procedures locally.

Beyond pure teleoperation, robotic systems also create opportunities for enhancement with artificial intelligence (AI) of the challenging navigation from the vascular access site to the intracranial occlusion. As this navigation requires operators to achieve predictable instrument tip behavior through indirect control and continuous adaptation, it remains a major procedural bottleneck and a potential source of serious complications (risks), including procedure-related vessel perforation (1%), dissection (2%), and distal embolization of the thrombus (9%) [13].

Outside of neurointervention, early efforts focused on automating constrained procedural subtasks [14]. Chen et al. presented an autonomous vascular access system using deep-learning-based robotic guidance, demonstrating the feasibility of AI-assisted vascular procedures [15]. Building on this for neurointervention, Abah et al. proposed a semi-autonomous catheter steering framework based on path-planar algorithms and preplanned optimal catheter configurations [16]. Yet, maintaining robust performance over long-horizon navigation tasks remains a key challenge in autonomous endovascular navigation, with recent work highlighting limitations in the generalization of reinforcement learning (RL) policies as task complexity and navigation duration increase [17].

Recent developments in AI-driven robotic control systems, also described as agentic control systems or AI copilots [18], have increasingly shifted attention toward collaborative paradigms that combine AI capabilities with clinician expertise, reflecting the emerging vision of surgical teams in which clinicians, AI systems, and robotic platforms operate as collaborative partners [19]. However, effective integration of human-AI collaboration remains an open challenge. In the endovascular domain, Yan et al. introduced a machine-learning-based human–machine collaborative control strategy for guidewire insertion that dynamically adapts to varying surgical conditions and incorporates a trust evaluation model to improve procedural safety and efficiency [20]. Such systems are increasingly viewed as shared autonomy agents designed to enhance procedural precision, reduce cognitive workload, and improve navigation efficiency. This way, AI-assisted robotic systems can help broaden access to care by supporting less experienced operators (i.e. novices) or even clinicians without specialized training during demanding procedural steps.

Despite these advances shown in [14]–[17], [19]–[22], existing AI research for robotic endovascular use cases such as MT, remains primarily focused on the development of increasingly autonomous navigation systems and improvements in navigation performance. In contrast, considerably less attention has been paid to understanding how clinicians and AI systems should collaborate during robotic endovascular procedures. As a result, it remains largely unclear how clinician-AI teams perform in robotic endovascular navigation settings, how operators prefer to interact with AI copilots in shared-control scenarios, and how different distributions of control authority influence navigation performance, workload, usability, and clinician perception.

Addressing these questions is critical for the clinical translation of AI-assisted robotic intervention, where successful adoption will likely depend not only on autonomous capabilities, but also on effective mechanisms for clinician-AI collaboration. Accordingly, this study investigates how different levels of AI-driven autonomy [23] influence objective navigation performance, operator workload, usability, and clinician perception during robotic endovascular navigation for MT (Figure 1).

Specifically, this work makes the following contributions:

- We introduce the Copilot Assistance Modes (CAMs), a framework for systematically investigating clinician-AI collaboration in robotic neurointerventional navigation. We instantiate this framework through multiple assistance paradigms representing guidance, shared control, delegated control, and fully autonomous navigation.
- We developed a comprehensive *in-vitro* evaluation framework that combines objective navigation outcomes with subjective assessments of workload, usability, and clinician perception for the evaluation of AI-assisted robotic interventions.
- We present, to our knowledge, the first *in-vitro* user study with clinical experts systematically comparing clinician-AI collaboration modes for robotic neurointerventional navigation under fluoroscopic guidance.
- We provide one of the first empirical investigations into how different allocations of control authority between clinicians and AI systems influence subjective and objective metrics, while also highlighting challenges associated with long-horizon autonomous navigation and zero-shot sim-to-real transfer.

II. Methods

The methodology is structured to progressively build the experimental framework used to investigate clinician-AI col-

laboration in robotic neurointerventional navigation. First, we define the navigation task and describe the *in-vitro* experimental platform, including the patient-specific vascular phantom, robotic manipulator, and user control interface. Next, we present the AI navigation agent that provides autonomous navigation capabilities. Building on these components, we introduce the Copilot Assistance Modes (CAMs), a framework for systematically studying different distributions of control authority between clinician and AI. Finally, we describe the user study design, outcome measures, and statistical analysis used to evaluate the proposed collaboration modes.

A. Endovascular Navigation Task

Endovascular navigation for a MT procedure involving a thrombus located in the distal cerebral vasculature typically begins at the femoral artery. From there, the device is advanced through the descending aorta to the aortic arch, followed by selective catheterization of the right common carotid artery (RCCA) or the left common carotid artery (LCCA) respectively, followed by subsequent progression into the internal carotid artery, and further advancement toward the intracranial target vessels – commonly the M1 segment of the middle cerebral artery (MCA).

The navigation task in this work was defined as advancing from the aortic arch to the target vessels right internal carotid artery (RICA) and right external carotid artery (RECA), respectively (Figure 2a). This segment is particularly relevant due to the anatomical complexity of the aortic arch branching with the need for precise instrument steering to ensure successful supra-aortic vessel selection.

Compared with the majority of autonomous endovascular navigation studies which are typically focused on tasks around the aortic arch [14], the investigated task represents an extended navigation horizon in both traversed distance and procedure duration. The estimated navigation distance for solely the aortic arch, from starting point to the brachiocephalic artery (BCA) is ≈250 mm, while our extended navigation toward RICA and RECA require a distance of ≈425 mm.

We used a patient-specific anatomical model reconstructed from clinical imaging data, namely a neck computed tomography angiography (CTA) scan extending from the aortic arch to the cerebral arteries and a body CTA scan encompassing the thoracic and abdominal vasculature, including the femoral arteries, descending aorta, and aortic arch. Ethical approval was obtained from the U.K. Health Research Authority Research Ethics Committee (London - South East REC; reference 24/LO/0057). The *in-vitro* test anatomy comprised a type I aortic arch without a bovine arch configuration, a reverse-curved RICA, and a high-riding BCA with a take-off angle of 141°. Recent evidence suggests that these anatomical characteristics are associated with reduced navigation success rates, with the reverse RICA curvature further contributing to longer procedure times [24]. Failure to negotiate these challenges may result in navigation errors and prolonged procedure times. Together, these factors make the navigation task sufficiently rigorous for evaluating clinician-AI collaboration.

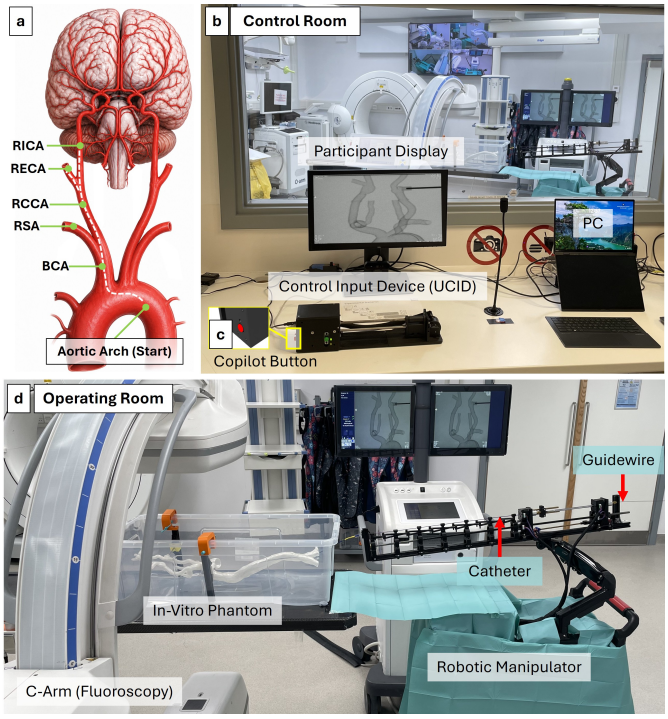


Figure 2. **Experimental Setup:** (a) Navigation path (white dashed line) from the descending aorta through the aortic arch and brachiocephalic artery (BCA) into the right common carotid artery (RCCA), bypassing the right subclavian artery (RSA), and proceeding to either the right internal carotid artery (RICA) or the right external carotid artery (RECA). (b) Control room from which the participants are teleoperating the robotic manipulator utilizing the User Control Input Device (UCID) [25]. (c) A red Copilot push-button functions as a dead-man switch, activating autonomous navigation. (d) Mock operating room with the robotic manipulator for catheter and guidewire [26], navigating within the *in-vitro* phantom under fluoroscopy imaging.

B. Experimental Setup for in-vitro Experiments

For the *in-vitro* testbed, we fabricated a 3D-printed vascular phantom model from translucent Polylactic Acid (PLA) based on the aforementioned patient anatomy (Section II-A). For endovascular instrumentation, a guidewire (Glidewire, 180 cm, 0.035”, Terumo, Tokyo, Japan), a multipurpose catheter with curved tip (Slip-Cath Beacon, 100 cm 0.044”, Penumbra, California, United States), and a Penumbra Neuron MAX (0.088”, Penumbra) were used in both the simulation (*in silico*) and the experimental testbed (*in vitro*). Only the guidewire was actively manipulated to reach the target, while both catheters remained static in the aortic segments below the arch to provide a stable base that facilitated precise guidewire manipulation.

The experiments were conducted in a mock operating room (OR) (Figure 2b–d), equipped with a mobile C-arm system (Cios Spin, Siemens Healthineers, Germany) with monoplane fluoroscopic imaging (mode: *vascular standard*, 30 pulses/s). A video stream of the fluoroscopic images was obtained using a video capture card. The tip position and two further points on the guidewire were tracked using a modified tracking algorithm [27] and provided to the AI agent (Section II-D). An unaltered video stream was displayed to participants in the adjacent control room, which was separated from the OR by radiation shielding and equipped with a viewing window

into the OR. All software components required for the experiments—including robot control, hardware communication, tracking, AI agent inference and response generation, and data logging—were executed on a notebook computer equipped with an Intel® Core™ Ultra 9 185H processor (2.50 GHz), 32 GB RAM, and integrated Intel® Arc™ Graphics, running a 64-bit Windows 11 operating system.

C. Robotic Manipulator and User Control Input Device

We used a custom-built endovascular robot to independently control the rotation φ_{Rob} and translation d_{Rob} of one instrument (Figure 2). Building upon a previous design [26], the robot features an enhanced control architecture with dedicated motor drivers (Tic 36v4, Pololu Corporation, USA). Participants operated the robot via a device-mimicking user control input device (UCID) [25] (Figure 2b). Through a slider-based mechanism, the interface allows operators to simultaneously input rotation (φ_{Inp}) and translation (d_{Inp}) with a single hand movement, thereby facilitating intuitive operation. An additional red push-button functions as a 'dead-man switch' (Figure 2c), enabling the Full Copilot Assistance functionality (Section II-E).

D. AI agent for Endovascular Navigation

The AI agent was trained exclusively in simulation, using the patient-specific anatomical model and single instrument (i.e., guidewire) control. During training, the agent had no access to the physical robotic system, fluoroscopic imaging and tracking data, or the physical characteristics of the *in-vitro* testbed (e.g., friction, control delays). The learned policy was subsequently deployed directly in the physical environment (*in vitro*) via zero-shot sim-to-real transfer, without prior hardware-based evaluation or fine-tuning.

Our simulation environment follows previous work based on the stEVE framework and SOFA [17], [28], where a Navien 0.058" support catheter and 0.035" guidewire were modeled using beam-based finite element representations. The patient-specific vascular anatomy was generated by fusing neck and body CTA scans to create complete femoral-to-cerebral arterial trees. To improve generalization, vasculatures were augmented through random scaling and rotation, while guidewire–vessel interaction parameters were tuned to replicate realistic device behavior and support subsequent transfer to physical experiments [29]. AI agent observations were intentionally limited to the coordinates of the three most distal tracked guidewire points ($P_1, P_2, P_3 \in \mathbb{R}^2$), the target location, previous device positions, and the previous action, without access to vessel geometry. Using this method, one can later extract these device tip points also from live fluoroscopic images [30] during MT, providing the inputs needed for the AI agent to navigate the vasculature in 3D. The action space comprised proximal guidewire rotation and translation commands, constrained to clinically realistic speed limits of 180°/s and 40 mm/s [28], respectively, enabling investigation of autonomous endovascular navigation using only device-tip tracking information.

We employed the soft actor-critic (SAC) RL algorithm [31] with multi-worker distributed training to optimize the control policy for the autonomous navigation AI agent.

TABLE I
KEY HYPERPARAMETERS FOR SOFT ACTOR-CRITIC ALGORITHM.

Parameter	Value	Parameter	Value
LSTM size	500	Batch size	32
MLP layers	4	Replay buffer	10^6
MLP hidden units	900	Exploration noise	0.25
Learning rate	2.2×10^{-4}	Training steps	10M
Discount factor γ	0.99	Warm-up steps	500K
Target update rate τ	0.005	Update ratio	0.05

LSTM, long short-term memory; MLP, multi-layer perceptron.

The policy and value networks employed recurrent architectures to capture temporal dependencies in the observation space. Both the twin Q-networks and policy network consisted of a long short-term memory (LSTM) head followed by a multi-layer perceptron (MLP), with key architectural and training hyperparameters summarized in Table I. Training was performed using SAC with Adam optimization, linear learning-rate scheduling, and action normalization. A 500K-step warm-up phase using uniformly random actions preceded exploration and policy learning for up to 10M steps, with periodic evaluation following the protocol of [28].

Success was evaluated using three metrics: task success rate, normalized path ratio, and procedure time [32]. Task success rate served as the primary criterion for selecting the optimal evaluation checkpoint during RL training [17]. Based on this metric, the checkpoint corresponding to 8M exploration steps was selected for all subsequent analyses. The AI agent was then evaluated both *in silico* and *in vitro* (RICA: $n = 100$; RECA: $n = 100$) to assess autonomous performance in the absence of human intervention. Training was conducted on an NVIDIA DGX A100 system (Santa Clara, CA, USA) equipped with eight GPUs and required approximately 11 hours to complete 8M exploration steps.

E. Copilot Assistance Modes (CAMs)

AI-driven robotic navigation can be understood as existing along a continuum of autonomy, ranging from manual control of robotic instruments, through shared or semi-autonomous assistance, and ultimately toward fully autonomous navigation. Along this spectrum, the role of the clinician gradually shifts from low-level control of endovascular devices toward higher-level supervision and decision-making, while the AI system assumes increasing responsibility for motion generation and trajectory execution.

To systematically investigate how control should be shared between clinicians and AI systems, we introduced CAMs, a framework for studying clinician-AI collaboration across progressively increasing levels of autonomy (Figure 1). These interaction modes were informed by discussions with our clinical partners and represent clinically motivated approaches to sharing control between human operators and AI systems. Each mode corresponded to a specific allocation of control authority between the operator and the AI agent, thereby enabling the isolated evaluation of guidance, shared control, delegated control, and fully autonomous navigation paradigms (TABLE II, Figure 3). CAMs (0) and (1) constituted conventional modes while CAMs (2)–(5) were AI-driven modes.

TABLE II
TAXONOMY OF COPILOT ASSISTANCE MODES (CAMS)

#	CAM	Rotation	Translation
0	Manual (baseline)	Manual φ_{Man}	Manual d_{Man}
1	Direct Robot Control	Teleoperated φ_{Inp}	Teleoperated d_{Inp}
2	AI Advice	Teleoperated φ_{Inp}	Teleoperated d_{Inp}
3	**AI Assistance**	Splitting Robotic DOF	
3a	AI Translation	Teleoperated φ_{Inp}	AI agent d_{AI}
3b	AI Rotation	AI agent φ_{AI}	Teleoperated d_{Inp}
4	**Full Copilot Assistance**	Splitting Navigation Path	
4a	Predefined Sections	$[\varphi_{\mathrm{Inp}}, \varphi_{\mathrm{AI}}]$	$[d_{\mathrm{Inp}}, d_{\mathrm{AI}}]$
4b	Operator-Defined Sections	$[\varphi_{\mathrm{Inp}}, \varphi_{\mathrm{AI}}]$	$[d_{\mathrm{Inp}}, d_{\mathrm{AI}}]$
5	Fully Autonomous Navigation	AI agent φ_{AI}	AI agent d_{AI}

Manual (0): As baseline, the operator controlled the guidewire hands-on, i.e. manually $[\varphi_{\mathrm{Man}}, d_{\mathrm{Man}}]$ from within the OR.

Direct Robot Control (1): As robotic baseline, the operator controlled the robot (i.e., teleoperation) with the UCID providing φ_{Inp} and d_{Inp} from a radiation-protected control room (Figure 2b).

AI Advice (2): This mode represented perceptual assistance. The operator retained full control of both DoF while the AI agent only provided action recommendation via the imaging video stream enhanced with an AI-Advice user interface (UI) (Figure 4a–b). AI-Advice indicated recommended direction and normalized amplitude of movement, based on the recommended normalized actions for rotation rate ω_{AI} and translation velocity v_{AI} of the AI agent.

AI Assistance (3): This mode represented shared control through DoF allocation. The AI agent controlled one robotic DoF, i.e. either AI Translation (3a) or AI Rotation (3b). The remaining DoF was controlled by the operator. The operator's position commands $[\varphi_{\mathrm{Inp}}, d_{\mathrm{Inp}}]$ served as the primary input, whereas the velocity actions generated by the AI agent $[\omega_{\mathrm{AI}}, v_{\mathrm{AI}}]$ were projected onto, and combined with, the operator's position command, resulting in a combined control input $[\varphi_{\mathrm{Inp}}, d_{\mathrm{AI}}]$ for AI Translation (3a) and $[\varphi_{\mathrm{AI}}, d_{\mathrm{Inp}}]$ for AI Rotation (3b).

Full Copilot Assistance (4): This mode represented temporary task delegation. The AI agent assumed control of both robotic DoF with $[\varphi_{\mathrm{AI}}, d_{\mathrm{AI}}]$ whenever activated by the operator through holding the Copilot push-button on the UCID (Figure 2c). In Full Copilot–Predefined Sections (4a), full assistance was activated at a predefined point, spanning the segment from the aortic arch to the RCCA. In Full Copilot–Operator-Defined Sections (4b), the operator could freely choose when to activate and deactivate full assistance.

Fully Autonomous Navigation (5): This mode represented complete task delegation. The AI agent navigated autonomously from start to target without human interaction.

Together, the CAMs provided a structured framework for evaluating how different distributions of control authority influence objective performance and subjective operator perception, thereby enabling systematic investigation of clinician-AI collaboration in robotic neuroendovascular intervention.

F. Hypothesis and Statistical Analysis

We hypothesized that shared autonomy and copilot-assisted control modes improve navigation performance relative tele-

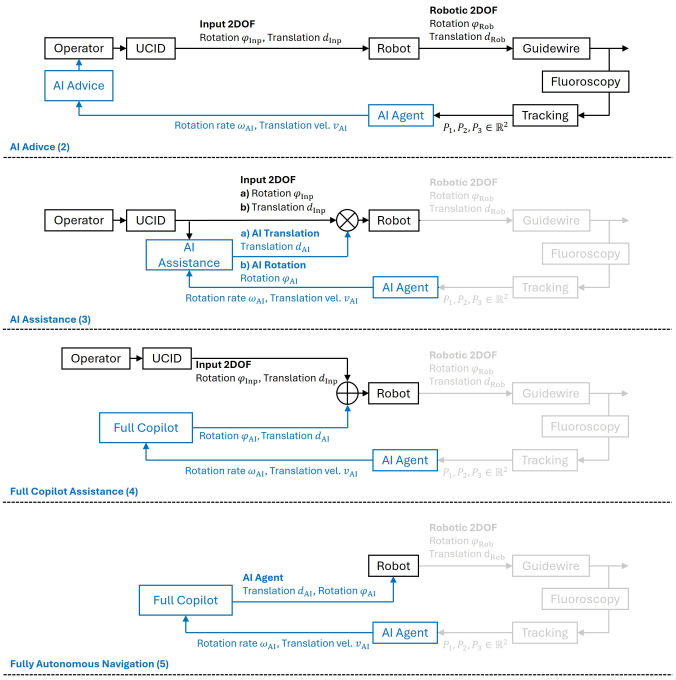


Figure 3. **Control Architectures for AI-Driven Modes:** Collaborative control loops for Copilot Assistance Modes (CAMs) (2)–(5), with AI components highlighted in blue. The operator controls the robot via the User Controller Input Device (UCID), which manipulates the guidewire within the 3D-printed vascular phantom. Instrument tip coordinates are extracted from the fluoroscopic image stream and provided to the AI agent. Grey components in CAMs (3)–(5) denote identical components. The symbol "⊗" represents the combination of inputs in AI Assistance (3), whereas "⊕" denotes a mutually exclusive selection between inputs in Full Copilot Assistance (4).

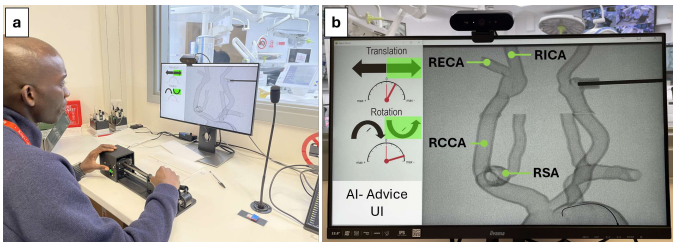


Figure 4. **AI-Advice User Interface (UI)**: (a) Participant navigating the guidewire via robotic manipulator with the User Controller Interface Device (UCID) from the control room. (b) AI Advice is provided by the UI next to the fluoroscopic image. Direction of motion is highlighted in green for both translation and rotation while the red needle in the dial gauges quantify the amplitude of motion. Note, that labels for the right subclavian artery (RSA), right common carotid artery (RCCA), right external carotid artery (RECA), and right internal carotid artery (RICA) were not shown on the UI to the operator during the experiments.

operated control while maintaining the clinician's sense of agency and supervisory control. Accordingly, we assessed whether Full Copilot Assistance (4) enhances navigation performance relative to Direct Robot Control (1), using Manual (0) as the reference baseline condition. Informed by the recommendations of the Stakeholder Taskforce for AI-assisted Robotic Thrombectomy (START) position statement [32], we evaluated navigation performance using complementary measures of task success, navigation efficiency, and navigation errors. The primary outcome metrics comprised task success rate (i.e., reaching the target anatomy) and time-to-target

(T2T), which served as the measure of navigation effectiveness and efficiency. The secondary outcome metric was the number of navigation errors, reflecting deviations from the intended navigation path during vessel selection and target approach. Navigation errors were defined as any instance of probing or navigating into an incorrect vessel branch. Specifically, errors comprised i) overshooting into the ascending aorta towards the aortic valve, ii) navigating into the right subclavian artery (RSA), and iii) falsely probing RICA or RECA.

To ensure sufficient statistical power for detecting differences in T2T, the required number of repetitions per condition was determined through an *a priori* power analysis (Supplementary Material A). The study was planned as a within-subject design with $N = 10$ participants and 6 repetitions per condition, allowing sufficient power to detect differences in navigation time and error count between Full Copilot Assistance (4) and the conditions Manual (0) and Direct Robot Control (1).

The remaining CAMs, AI Advice (2) and AI Assistance (3), were included as exploratory conditions with 4 repetitions per participant.

T2T and errors were analyzed using successful trials only. Due to the expected small number of participants, differences among CAMs were assessed using a permutation-based one-way ANOVA. To account for the study design, CAM condition assignments were permuted 10,000 times while preserving the participant-cohort structure. The ANOVA F statistic was used to evaluate whether the mean outcome differed across CAMs. Statistical significance was assessed at $\alpha = 0.05$. Post-hoc pairwise comparisons were conducted using two-sided permutation tests, and the Holm step-down correction was applied to control the family-wise error rate across comparisons within each outcome. Corrected values are reported as p_{Holm}, with statistical significance defined as $p_{\mathrm{Holm}} < 0.05$.

G. Participants

The final participant cohort ($N = 13$) consisted of 10 neurointerventional radiologists, including 6 UK consultants (US attending equivalents) and 4 fellows with 2–18 years of neurointerventional experience, 1 general interventional radiology resident, and 2 third-year medical students. Participants were stratified into four subgroups based on procedural target assignment and experience level. All participants provided written informed consent in accordance with the Declaration of Helsinki. The participant briefing is provided in Supplementary Material B.

H. Experimental Procedure

The objective of the task was to navigate the guidewire from its starting point in the aortic arch into the target anatomy (RICA or RECA) under different CAMs. A maximum time limit of $240\,\mathrm{s}$ was imposed for each experimental run. To mitigate potential ordering and learning effects, the sequence of evaluated CAMs was randomized across participants, with the exception of Manual (0), which was consistently performed last.

I. Survey Design

Following the experimental trials, participants completed a survey designed to capture subjective evaluations of the different CAMs. The survey assessed perceived workload, usability, and overall preference for each in order to complement the objective performance measures collected during the experiments. Subjective workload was assessed using a modified version of the NASA Task Load Index (NASA-TLX). To reduce respondent burden and facilitate repeated assessments within the surgical simulation workflow, all workload dimensions were rated on a 5-point Likert scale (1 = low workload, 5 = high workload).

Usability was assessed using the System Usability Scale (SUS), a standardized and widely adopted 10-item questionnaire with 5-point Likert-scale responses [33]. The SUS provides a reliable measure of perceived usability and ease of interaction with each CAM and responses were converted to the standard SUS score ranging from 0–100, with higher scores indicating better perceived usability. Usability outcomes were interpreted according to the ISO 9241-11 definition of usability, encompassing effectiveness, efficiency, and user satisfaction in the specified context of use [34].

Statistical analysis of NASA-TLX and SUS mirrored that of T2T described in Section II-F, comprising a two-sided permutation-based one-way ANOVA followed by pairwise two-sided permutation tests with Holm step-down correction.

In addition to the standardized questionnaires, participants were asked to comparatively rank the evaluated CAM according to their overall preference and perceived effectiveness. This ranking was included to capture qualitative operator preferences that may not be fully reflected by workload or usability scores alone.

Additional survey items assessed participants' experiences with and perceptions of the UCID, as well as their evaluation of the realism of the *in-vitro* phantom.

III. Results

A. Objective Performance Metrics

1) ***Success Rate****:* Figure 5 shows the run outcomes for all CAMs. Fully Autonomous Navigation (5) achieved the lowest success rate (23%), followed by AI Translation (3a) (38%). All remaining robotic CAMs achieved success rates between 92% (Full Copilot–Predefined Sections (4a)) and 97% (Full Copilot–Operator-Defined Sections (4b)) while Direct Robot Control (1) without any AI assistance gave 95%. All robotic CAMs exhibited time-out rates between 2%–6% including Direct Robot Control (1). Compared with Fully Autonomous Navigation (5), all assisted and manual conditions demonstrated substantially improved performance. Fails in AI Translation (3a) occurred when the AI agent trapped the instrument in an incorrect vessel branch. A single fail run also occurred in Full Copilot–Predefined Sections (4a), when the AI agent autonomously retracted the instrument from the scene. An in-depth analysis for Fully Autonomous Navigation (5) can be found in the Supplementary Material C.

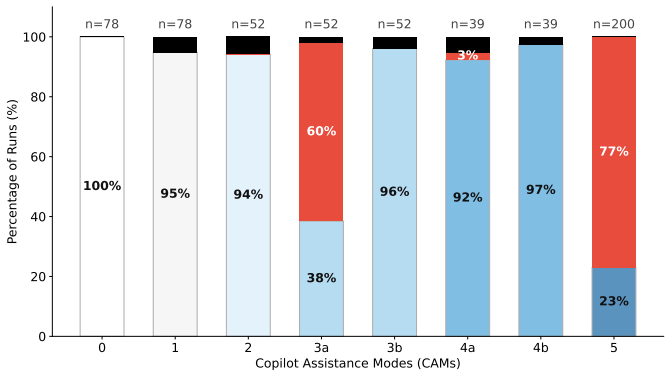


Figure 5. **Success Rates Comparison:** Stacked bars show the proportions of successful, failed (red), and timed-out (black) runs for each of the Copilot Assistance Modes (CAMs) (TABLE II). Percentages within the bars indicate success and failure rates, and values above the bars indicate the total number of runs (n). For each participant, CAMs (0), (1), and (4), i.e. 4a and 4b combined, were tested in 6 runs, whereas CAMs (2), (3a), and (3b) were each tested in 4 runs. Fully Autonomous Navigation (5) was evaluated without participant involvement in 200 autonomous runs (100 targeting the right internal carotid artery (RICA) and 100 targeting the right external carotid artery (RECA)). The blue color gradient indicates increasing levels of autonomy, ranging from white (manual) over grey (teleoperation) to dark blue (highest autonomy).

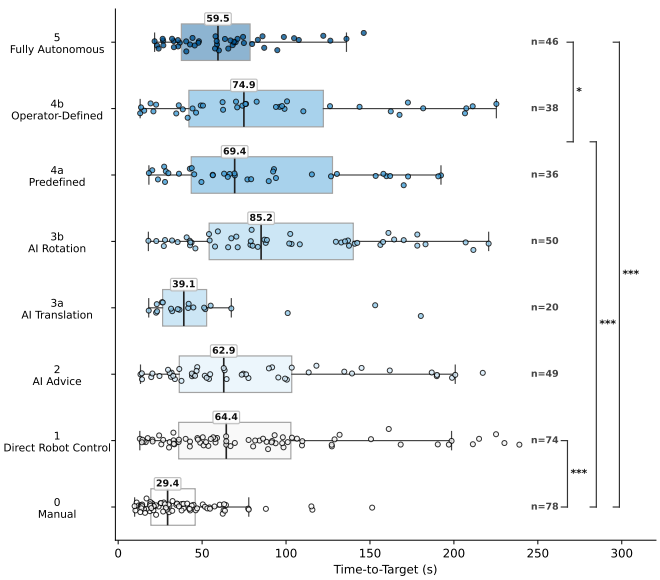


Figure 6. **Time-to-Target (T2T):** Distribution of T2T for all successful runs of each Copilot Assistance Mode (CAM). n values denote the number of successful runs per condition. Significance levels are denoted based on Holm-adjusted p-values: $(*)$, $(0.01 \leq p_{\text{Holm}} < 0.05)$; $(***)$, $(p_{\text{Holm}} < 0.001)$. The blue color gradient indicates increasing levels of autonomy, ranging from white (manual) over grey (teleoperation) to dark blue (highest autonomy).

*2) **Time-to-Target (T2T)**:* Figure 6 shows the distribution of T2T. Overall, there was no significant difference for T2T between Direct Robot Control (1) and any of the AI modes (2)–(4) that were not fully autonomous. Medians ranged from $29.5\,\text{s}$ for Manual (0) to $85.2\,\text{s}$ for AI Rotation (3b). A permutation-based ANOVA revealed a significant effect of CAM on T2T, $F(6, 338) = 10.643$, $p_{\text{perm}} < 0.0001$ with a large omnibus effect, $\eta^2 = 0.159$ and $\omega^2 = 0.144$. Holm-corrected pairwise comparisons identified two set of significant effects. First, Manual (0) resulted in significantly lower T2T values than Direct Robot Control (1), Full Copilot Assistance (4), and Fully Autonomous Navigation (5) (all $p_{\text{Holm}} = 0.0006$). The corresponding median T2T values were $29.45\,\text{s}$, $64.40\,\text{s}$, $72.15\,\text{s}$, and $59.50\,\text{s}$, respectively. Second, a significant difference was observed between medians of Full Copilot Assistance (4) and Fully Autonomous Navigation (5) ($72.15\,\text{s}$ vs. $59.50\,\text{s}$; $p_{\text{Holm}} = 0.0303$). However, the T2T of AI Translation (3a) and Fully Autonomous Navigation (5) should be interpreted cautiously, given the relatively small number of successful runs.

*3) **Errors**:* Figure 7 shows the distribution of navigation errors per run and per minute. Navigation errors, as defined in Section II-F, comprise deviations into incorrect vessel branches or bifurcations. All pairwise comparisons between Direct Robot Control (1) and the the AI-driven modes (2), (3a), (3b), (4a), and (4b) were non-significant ($p = 0.590$, 0.761, 0.836, 0.802, and 0.191, respectively).

Median error counts ranged from 0.00–1.00 errors per run, while median error rates ranged from 0.00–$0.70\,\text{min}^{-1}$. The lowest values were observed for Manual (0) and Fully Autonomous Navigation (5), whereas the highest values occurred in Full Copilot–Predefined Sections (4a). For CAMs (0)–(5), a permutation-based ANOVA revealed a significant effect of CAM on errors per run ($F(6, 338) = 3.818$, $p_{\text{perm}} = 0.0008$, $\eta^2 = 0.063$, $\omega^2 = 0.047$), but not on errors per minute ($F(6, 338) = 1.179$, $p_{\text{perm}} = 0.3148$, $\eta^2 = 0.021$, $\omega^2 = 0.003$). Holm-corrected pairwise tests for errors per run showed that Manual (0) differed from Direct Robot Control (1) and Full Copilot Assistance (4) (both $p_{\text{Holm}} = 0.0003$). Median errors per run were 0.00 for Manual (0) and 1.00 for both Direct Robot Control (1) and Full Copilot Assistance (4). No significant difference was found between Direct Robot Control (1) and Full Copilot Assistance (4) ($p_{\text{Holm}} = 0.2965$). Likewise, no significant differences were observed among AI Advice (2), AI Translation (3a), and AI Rotation (3b), whose median errors per run were 1.00, 0.50, and 1.00, respectively (corresponding median errors per minute: $0.41\,\text{min}^{-1}$, $0.17\,\text{min}^{-1}$, and $0.51\,\text{min}^{-1}$).

*4) **AI Usage Analysis**:* For Full Copilot–Operator-Defined Sections (4b), where participants could freely use the AI agent throughout the task, we additionally analyzed the spatial distribution of AI agent activity by recording the start and end positions of each utilization interval. This allowed us to characterize participants' engagement with the AI agent and quantify its contribution during the navigation task. Figure 8 shows the spatial distribution of AI activation and deactivation. AI was most frequently activated at the start of the procedure in the aortic arch, and to a lesser extent around the RSA-RCCA bifurcation. Deactivation points formed a dense cluster in the RICA (for RECA targets) and another near the RSA-RCCA bifurcation including within the RSA origin.

B. Survey

To ensure that the subjective evaluations reflected the views of individuals with relevant clinical expertise, only those participants with an interventional radiology background were included (11/13). The complete survey results, including all

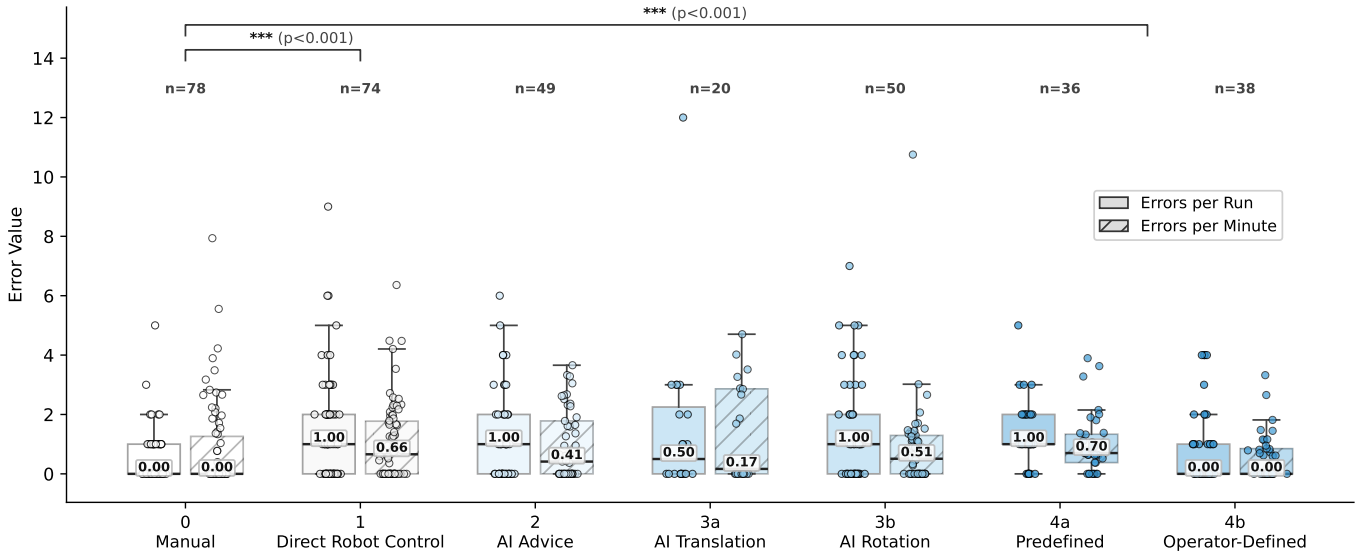


Figure 7. **Errors:** Distribution of navigation errors per run (filled) and per minute (hatched) for all successful runs of each Copilot Assistance Mode (CAM). Navigation errors correspond to deviations into incorrect vessel branches (Section II-F). n denotes the number of successful runs per CAM. Fully Autonomous Navigation (5) is absent because any navigation error resulted in an unsuccessful trial. Unlike the clinician-supervised modes, in this mode, the AI agent could not recover from these errors, leaving no successful trials to display. The blue color gradient indicates increasing levels of autonomy, ranging from white (manual) over grey (teleoperation) to dark blue (highest autonomy).

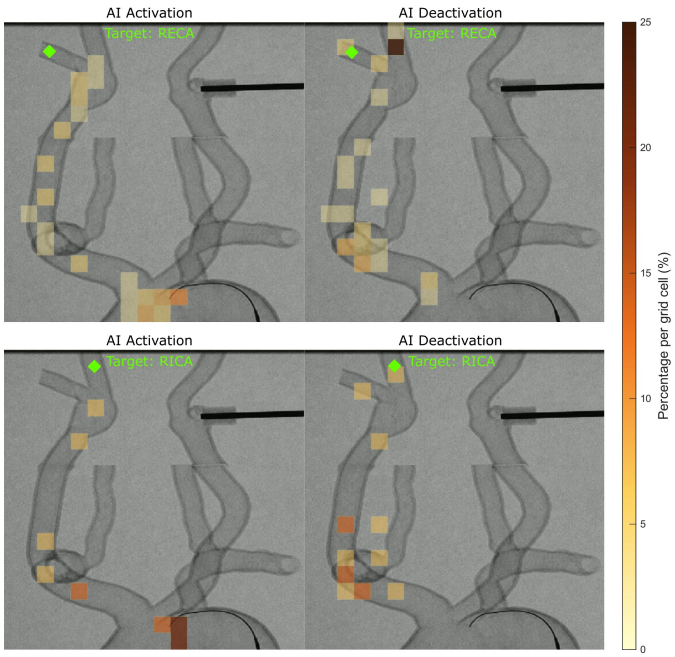


Figure 8. **AI activation and deactivation during Full Copilot–Operator-Defined Sections (4b):** Heatmaps depict the spatial distribution of AI activations (left column) and deactivations (right column) during navigation to the right external carotid artery (RECA, top row) and right internal carotid artery (RICA, bottom row). Values represent the percentage of AI-assisted interactions per grid cell and are overlaid on fluoroscopic phantom images. Warmer colors indicate a higher frequency of AI activation or deactivation within the corresponding region.

free-text responses, are available in the Supplementary Material D.

1) Robot Controller Interface UCID: Several clinicians (64%, $7/11$) reported prior exposure to the UCID [25], while $4/11$ reported first-time use. Free-text responses indicated that the interface was intuitive, easy to learn, and reasonably aligned with clinically familiar guidewire manipulation with continuous control. It was also valued for the ability to coordinate translation and rotation with one hand, highlighting its potential for practical application, e.g., performing additional tasks with the other hand. In contrast, participants reported limitations related to delays between input commands and robot movement, particularly during rotational instrument manipulation; the restricted range of continuous translational motion imposed by the UCID stroke length; and reduced tactile feedback.

2) Phantom Realism: The phantom was rated as realistic by most clinicians, with a mean realism score of $4.1/5.0$, median score of $4.0/5.0$ and range [3.0–5.0]. Most clinicians (82%, $9/11$) rated realism as 4.0 or 5.0. According to free-text responses, the positive ratings were mainly due to the vessel geometry, branch localization, overall anatomical shape, and guidewire behavior. Functionally, it was described as similar to real neurointerventional navigation, including familiar challenges such as the guidewire preferentially entering the RSA. However, participants also identified important limitations of the phantom model as not fully reproducing *in-vivo* conditions [32], particularly blood flow, pulsatility, vessel compliance, smaller branch detail, and realistic friction.

3) NASA-TLX: The heatmap in Figure 9 illustrates the perceived workload across control conditions. Physical demand remained consistently low across all modes, with mean scores ranging from 1.45–1.73 on a 5-point scale, indicating that neither robotic control nor AI assistance imposed substantial physical burden during guidewire manipulation. Greater variation was observed for mental demand, effort, frustration, and perceived performance. A permutation-based one-way

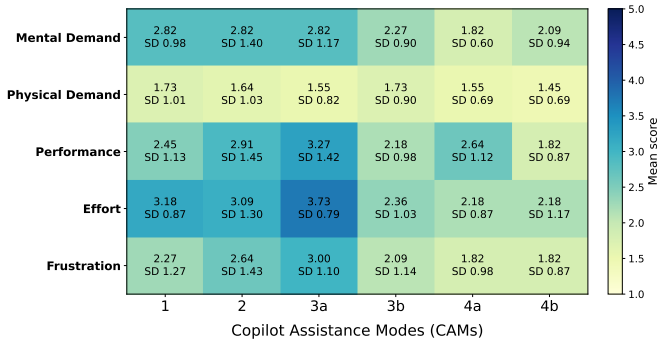


Figure 9. **Subjective Workload:** Heatmap of NASA-TLX ratings by subscale and condition. Values in each cell are mean ± SD (displayed on two lines) on a 1–5 scale, with the Performance subscale reverse-coded to align interpretation (higher score = higher perceived workload).

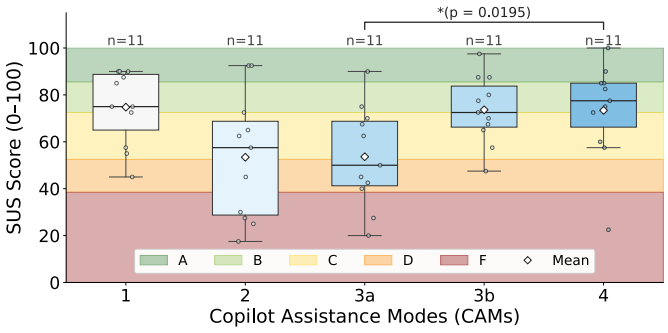


Figure 10. **System Usability:** Distribution of System Usability Score (SUS) for all CAM (expert clinician participants only, $N = 11$): boxplots show mean, median, and IQR (whiskers: 1.5×IQR); points show individual participants, shaded bands indicate standard SUS grade ranges (A–F). The SUS ratings for each item ranged from 0–5 and were converted to a 0–100 scale. The blue color gradient indicates increasing levels of autonomy, ranging from grey (teleoperation) to dark blue (highest autonomy).

ANOVA revealed an overall effect of control condition on NASA-TLX scores ($F(5, 60) = 2.77$, $p_{\text{perm}} = 0.0217$), with a moderate effect size ($\eta^2 = 0.187$; $\omega^2 = 0.118$). However, no pairwise comparison remained statistically significant after Holm correction for multiple comparisons (all $p_{\text{Holm}} \geq 0.05$). The smallest adjusted p-value was observed for the comparison between AI Rotation (3b) and AI Advice (2) ($p_{\text{Holm}} = 0.0825$). Therefore, the differences described below should be interpreted as descriptive trends rather than statistically significant pairwise effects. Among the AI-driven modes, AI Translation (3a) exhibited the highest mean workload ratings, including effort (3.73 ± 0.79), frustration (3.00 ± 1.10), and reverse-coded performance workload (3.27 ± 1.42). This trend is consistent with qualitative feedback suggesting that operators found it challenging to coordinate manual rotational input with autonomous translational motion. In contrast, AI Rotation (3b) showed a trend toward lower ratings of mental demand, effort, frustration, and performance-related workload – especially when compared to Direct Robot Control (1). This trend is consistent with qualitative feedback suggesting that the operator could wait for the AI to rotate correctly before the operator translated the guidewire forwards into the correct bifurcation branch. The lowest overall workload ratings were observed for Full Copilot Assistance (4), with both Full Copilot–Predefined Sections (4a) and Full Copilot–Operator-Defined Sections (4b) associated with low levels of mental demand, effort, frustration, and performance-related workload.

4) System Usability Scale (SUS): Figure 10 and Table III summarize the SUS scores across modes. Scores were converted to a 0–100 scale and interpreted using standard SUS benchmarks, where mean values above 68 indicate above-average usability and values between 50–68 indicate marginal usability. Manual operation was not assessed using the SUS. Mean SUS scores ranged from 53.4 ± 26.5 (AI Advice (2)) to 74.8 ± 15.9 (Direct Robot Control (1)). According to SUS benchmark classifications, Direct Robot Control (1), AI Rotation (3b), and Full Copilot Assistance (4) achieved mean scores in the good usability range, whereas AI Advice (2) and AI Translation (3a) fell within the marginal usability range. Across 11 participants and five CAMs, a permutation-based one-way repeated-measures ANOVA revealed a significant effect of setup on SUS scores ($F(4, 40) = 4.17$, $p_{\text{perm}} = 0.0076$, $\eta^2 = 0.29$, $\omega^2 = 0.22$). Holm-corrected comparisons identified a significant difference only between AI Translation (3a) and Full Copilot Assistance (4) ($p_{\text{Holm}} = 0.021$).

When considering median SUS scores, Full Copilot Assistance (4) achieved the highest overall usability rating (77.5), followed by Direct Robot Control (1) (75.0) and AI Rotation (3b) (72.5), consistent with the ranked order shown in Table III.

Qualitative feedback was broadly consistent with these findings. Users reported that AI Rotation (3b) and Full Copilot Assistance (4) reduced workload and improved navigation during challenging procedural segments, although concerns regarding transparency, occasional unexpected AI behavior, and overall system reliability remained. Participants rated Direct Robot Control (1) favorably but noted the loss of tactile feedback and limited advantages over manual control for experienced operators. The comparatively lower SUS scores of AI Advice (2) and AI Translation (3a) were frequently attributed to difficulties integrating AI advice into real-time navigation and coordinating manual rotation with AI-controlled translation.

TABLE III
SYSTEM USABILITY SCALE (SUS) SCORES BY COPILOT ASSISTANCE MODES (CAMS)

#	CAM*	**Median**	Mean ± SD	Range	Grade
4	Full Copilot	**77.5**	73.4 ± 20.9	22.5–100.0	B
1	Direct Robot Control	**75.0**	74.8 ± 15.9	45.0–90.0	B
3b	AI Rotation	**72.5**	73.6 ± 14.4	47.5–97.5	B
2	AI Advice	**57.5**	53.4 ± 26.5	17.5–92.5	C
3a	AI Translation	**50.0**	53.6 ± 21.3	20.0–90.0	C

*From highest to lowest median SUS score. Standard grade ranges (A–F).

5) Ranking: TABLE IV presents the participant preference rankings, ordered according to median rank. Manual (0) was the most preferred condition overall, with the lowest mean rank (1.64 ± 1.29) and the strongest first-choice frequency ($8/11$). Among the telerobotic and AI-driven modes, Direct Robot Control (1) and Full Copilot Assistance (4) were the most competitive alternatives with a mean rank of 3.00 ± 1.18 and 3.09 ± 1.70, respectively. Their preference patterns

TABLE IV
RANKING-BASED EVALUATION OF COPILOT ASSISTANCE MODES (CAMS)

#	CAM*	**Median (IQR)**	Mean ± SD	1st Choice	Last Choice
0	Manual	**1 [1.0–1.5]**	1.64 ± 1.29	8/11	0/11
1	Direct Robot Control	**3 [2.0–3.5]**	3.00 ± 1.18	0/11	0/11
4	Full Copilot	**3 [2.0–4.5]**	3.09 ± 1.70	2/11	1/11
2	AI Advice	**4 [3.0–6.0]**	4.18 ± 1.78	1/11	4/11
3b	AI Rotation	**4 [4.0–4.5]**	4.09 ± 0.70	0/11	0/11
3a	AI Translation	**6 [4.5–6.0]**	5.00 ± 1.41	0/11	6/11

*From lowest to highest median. Lower values indicate stronger preference.

differed: Direct Robot Control (1) was ranked in the top three by $8/11$ clinicians. and no participant ranked it as the first or last choice; In contrast, preferences for Full Copilot Assistance (4) were more variable, with $2/11$ participants ranking it first and $1/11$ ranking it last. Notably, these rankings at the extremes originated from consultant-level participants, suggesting greater heterogeneity in preferences within this subgroup. The least favorable ranking overall showed AI Translation (3a) with a median rank of 6 and ranked last by $6/11$ participants.

IV. DISCUSSION

A. Interpretation of Copilot Assistance Modes (CAMs)

The evaluated CAMs revealed distinct trade-offs in the distribution of control authority between clinicians and the AI agent. Increasing autonomy did not consistently improve objective performance or clinician acceptance. Instead, the perceived utility of each mode appeared to depend on whether autonomous behaviour was comprehensible, predictable and compatible with the clinician's ongoing control strategy.

AI Guidance: AI Advice (2) received a moderate workload profile but a marginal SUS score, ranking second to last among the evaluated modes. Participants reported difficulty following the guidance while simultaneously monitoring fluoroscopic imaging and controlling the guidewire, describing it as rapidly changing, unclear, counterintuitive and affected by latency. This may reflect the agent's low-level control strategy, which generates repeated forward–backward and rotational movements to optimise guidewire pose before advancing. Presenting these instantaneous actions directly to the operator obscures the higher-level navigation strategy and may make the agent appear erratic. Several participants consequently appeared to reduce their reliance on the guidance after an initial attempt to follow it.

Nevertheless, a minority perceived the mode as beneficial, suggesting potential for the underlying concept with improved interaction design. Rather than streaming low-level actions, future systems could communicate higher-level navigation intentions and integrate them into the clinician's primary field of attention, for example by overlaying directional cues at the guidewire tip within the fluoroscopic image. Haptic guidance through the UCID, such as *Guided Control* [35], could provide an alternative communication channel with lower visual demand.

Shared Control: The two configurations of AI Assistance (3), AI Translation (3a) and AI Rotation (3b), produced markedly different responses despite both distributing control authority between the clinician and the AI agent. AI Translation (3a) received a marginal SUS score and the least favorable workload profile, with the highest ratings for effort, frustration and perceived performance demand. Its relatively short T2T should be interpreted cautiously, as this metric included only successful trials and therefore excluded the substantial number of failures associated with its 38% success rate. The difficulties with AI Translation (3a) may be related to the disruption of the natural coupling between translation and rotation. Participants could not anticipate the AI agent's actions, while the AI agent could advance the guidewire before the operator had completed the required rotational adjustments. This reduced coordination between the two control dimensions. In contrast, AI Rotation (3b) received a comparatively high SUS score and favorable workload profile. Participants reported that retaining direct robotic control over translation provided greater procedural influence and allowed them to slow or stop before critical bifurcations. Several also perceived the AI agent as more effective at rotational manipulation than themselves. However, uncertainty regarding the AI agent's intentions and coordination between translation and rotation remained concerns.

For both configurations, the AI agent had been optimized primarily for autonomous navigation rather than clinician–AI collaboration. Its actions were therefore generated independently of operator intent, requiring clinicians to adapt to the AI agent rather than the reverse. The contrasting responses to AI Translation (3a) and AI Rotation (3b) suggest that effective shared autonomy depends not only on the allocation of control authority but also on coordination between autonomous behavior and operator intent. Collaboration-aware policies that incorporate the clinician's ongoing behavior or inferred intention may therefore support more effective human–AI coordination.

User-Directed Delegation: Full Copilot Assistance (4) achieved a favorable workload profile and one of the highest SUS scores when Full Copilot–Predefined Sections (4a) and Full Copilot–Operator-Defined Sections (4b) were considered together, and received the most favorable overall ranking among the AI-driven modes. Participants particularly valued Full Copilot–Operator-Defined Sections (4b), which allowed them to invoke assistance selectively according to the anatomical context and task demands. Assistance could be used during challenging segments or during relatively straightforward anatomy to accelerate progression, with direct robot control resumed for critical steps or when performance deteriorated.

This flexibility did not eliminate concerns regarding autonomous performance. Participants reported anatomy-dependent failures, including unintended entry into the RSA, failure to reach the target and additional recovery work when the AI agent left the guidewire in an unfavorable position. The activation patterns provide further insight into how participants incorporated autonomous control. In Full Copilot–Operator-Defined Sections (4b), assistance was most frequently activated near the beginning of the procedure, indicating early willingness to delegate irrespective of the target vessel. Because bypassing the RSA into the RCCA

was particularly challenging, activation frequently occurred near this region following unsuccessful unaided attempts. Deactivation locations were more heterogeneous, consistent with participants disengaging the AI agent when it selected an incorrect branch such as RICA when targeting the RECA, or RSA when targeting the RCCA.

Together, these observations suggest that clinicians treated autonomous control as an opportunistic resource rather than a continuously active controller: they delegated control when autonomous behavior appeared beneficial and resumed direct robot control when it became unreliable. This preserved clinician authority while allowing selective use of AI assistance may explain the comparatively favorable perception of Full Copilot Assistance (4) relative to AI Advice (2) and AI Assistance (3). The findings further suggest that effective delegation requires not only reliable autonomous performance but also sufficient transparency for clinicians to anticipate when assistance is likely to be beneficial and recognize when it should be withdrawn. Communicating the agent's intended action or confidence could therefore support more informed delegation decisions.

B. Implications for Clinician-AI Collaboration

A key insight emerging from our results is that the shared-autonomy modes allowed compensation for limitations of the autonomous navigation agent. Fully Autonomous Navigation (5) achieved a success rate of only 23%, yet several shared-autonomy modes built on the same underlying policy reached success rates of 92–97%, comparable to or exceeding Direct Robot Control (1) (Section III-A1, Figure 5). This disparity suggests that overall performance depended less on the autonomous policy itself and more on how clinicians were able to supervise, interpret, or selectively deploy AI assistance. Thus, the limitation of sim-to-real domain shift shown here was largely mitigated. The strongest example was Full Copilot–Operator-Defined Sections (4b), which slightly exceeded the success rate of Direct Robot Control (1) despite relying on the same AI policy that performed poorly in isolation. This finding suggests that operator-directed task delegation may allow clinicians to exploit beneficial autonomous behaviors while intervening when AI actions become suboptimal. This interpretation is further supported by the T2T results. Although Fully Autonomous Navigation (5) exhibited the lowest success rate, completion times among successful trials were among the shortest observed (Section III-A2, Figure 6), indicating that the AI agent possessed useful navigation competence when successful but lacked the robustness required for reliable autonomous operation.

Taken together, these findings indicate that an essential advantage of shared autonomy may not be to augment clinician performance directly, but to make imperfect AI systems clinically usable. Human oversight therefore appears to be more than a temporary safeguard until autonomous navigation becomes sufficiently reliable. Instead, clinician involvement may constitute an integral component of overall system performance. Designing clinician-AI collaboration strategies that are robust to imperfect autonomy may conceivably prove more important for clinical translation than pursuing higher levels of autonomy alone. Rather than representing an intermediate step toward full autonomy, shared-autonomy frameworks may constitute a long-term paradigm in which clinical expertise and AI capabilities complement one another, combining human robustness with the efficiency and task-specific strengths of autonomous systems.

C. Study Limitations

This study represents an initial investigation of clinician-AI collaboration in robotic neurointervention. Despite prospective power calculations, the small expert cohort (10 interventional neuroradiologists and 1 interventional radiologist), substantial inter-operator variability, and modest differences between assistance modes limited the detection of statistically significant effects. Nevertheless, the recruited cohort represents almost 10% of the UK's interventional neuroradiology workforce (123 in 2025) [8], highlighting the practical challenges of conducting large-scale studies in this highly specialized field.

The relatively small sample size also limited the ability to investigate differences between participant subgroups. Whilst clinical experience was considered a potential confounder, our sensitivity analysis suggested that subgroup composition did not influence performance outcomes (Supplementary Material A).

Furthermore, the generalizability of the findings is inherently limited, as the study examined a single navigation task within one patient-specific vascular anatomy. While the selected case contained clinically relevant challenges, it cannot represent the full diversity of neurointerventional procedures. Future work should therefore evaluate the proposed CAMs across a broader range of anatomies, navigation tasks, and clinical scenarios to determine whether the observed trends persist under varying procedural demands, where differences in workload, navigation complexity, and the value of AI assistance may become more pronounced.

Additional limitations arise from the experimental platform itself. The patient-specific phantom provided a controlled and reproducible environment for comparing CAMs, but did not capture physiological factors such as blood flow, vessel deformation, pulsatility, and realistic instrument–vessel interactions, all of which may influence navigation dynamics and affect both clinician and AI performance [32]. Similarly, participants reported perceived latency and reduced rotational precision when using the robotic control interface compared with manual manipulation. However, because the same robotic platform and interface were used across all robotic CAMs, it is plausible that these effects and further systematic errors are unlikely to account for the relative differences observed between assistance modes. In summary, while the comparative findings regarding clinician-AI collaboration are likely robust within the study setting, their translation to clinical practice should be validated in more realistic experimental and clinical environments.

D. Sim-to-Real Gap

Success rates decreased substantially from 81% *in silico* to 23% *in vitro*, demonstrating a pronounced sim-to-real perfor-

mance gap (Section III-A1). However, successful autonomous runs remained comparatively fast (Section III-A2, Figure 6), suggesting that locally effective navigation behaviors were preserved. One possible interpretation of these findings is that the principal challenge was not navigation efficiency itself, but maintaining robust performance over extended trajectories in the physical environment. This interpretation is consistent with the long-horizon, zero-shot nature of the investigated navigation task (Section II-A). Over such distances, small control inaccuracies, tracking errors, and sim-to-real mismatches can accumulate and ultimately result in procedural failure. Notably, several shared-autonomy modes achieved substantially higher success rates despite relying on the same underlying policy, suggesting that clinician supervision can compensate for a proportion of these failures. Nevertheless, improving sim-to-real transfer, state estimation, and long-horizon decision-making will likely be essential before fully autonomous robotic neurointerventions can be deployed reliably in clinical practice.

V. Conclusion

This work introduced Copilot Assistance Modes (CAMs), a framework for systematically investigating clinician-AI collaboration in robotic endovascular navigation, and presented the first *in-vitro* user study comparing multiple assistance paradigms across the autonomy spectrum with clinical experts. The implemented AI-driven modes did not outperform conventional robotic teleoperation, while hands-on manual navigation remained the strongest-performing mode overall. Nevertheless, several shared-autonomy approaches achieved performance comparable to robotic teleoperation despite limitations of the deployed AI agent, plausibly suggesting that effective clinician-AI collaboration may depend less on increasing autonomy and more on the allocation of control authority between human operators and AI systems. Notably, all shared-autonomy modes substantially outperformed the AI agent operating alone, indicating that certain clinician-AI collaboration strategies are robust to imperfect AI. Among the evaluated approaches, operator-directed task delegation and AI-assisted rotation showed the most favorable overall combination of objective performance, workload, usability, and clinician preference. The study further provides one of the first empirical investigations of long-horizon autonomous endovascular navigation and highlights the substantial challenges associated with sim-to-real transfer. Overall, the findings indicate that shared-autonomy frameworks may have the potential to provide a practical pathway toward the clinical adoption of AI-assisted endovascular robotics. Future progress will likely depend on improving the robustness, generalizability, and clinical reliability of autonomous navigation systems while preserving meaningful clinician oversight.

Author Contribution

S.M.H.S., C.B., and T.B. conceived the idea of a neuroendovascular copilot. N.F. designed the study, implemented the hardware and software components, performed pre-testing, conducted the experiments, analyzed the data, and wrote the majority of the manuscript. H.R. trained the AI agent and contributed to manuscript writing. X.C. and W.D. supported experimental execution, data and survey analysis, and manuscript writing. B.J. contributed to experimental planning and developed the UCID. S.M.H.S. contributed to the development of concepts for the CAMs, experimental planning, and the robotic manipulator. M.A.K., S.S., N.K., P.B., S.G., C.C., J.L., N.A., M.B., and A.F. provided clinical feedback during framework development and experimentation. A.G. contributed to manuscript revision and editing. C.B. and T.B. served as equal senior supervisors of the project and provided critical feedback and manuscript review throughout the study. All authors read and approved the final manuscript.

Competing Interest

All authors declare no financial or non-financial competing interests.

Acknowledgments

The authors thank Duane James, Gayathri Nantharatnam, and Marty Rajaratnam for their support in facilitating this study and the operation of the mock operating room, as well as Raghav Khanna, Adam Warwicker, and Yi-Yi Liu for their help on manufacturing printed components, and Christoph Eyberg for providing the code base for automatic tool tracking.

Large Language Models (LLMs) were used to assist with the structure, grammar, and spelling of this manuscript and for the creation of visualizations. All outputs were reviewed, edited, and verified by the authors.

References

[1] V. L. Feigin *et al.*, "Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the global burden of disease study 2019," *The Lancet Neurology*, vol. 20, no. 10, pp. 795–820, 2021. [Online]. Available: https://doi.org/10.1016/S1474-4422(21)00252-0

[2] ——, "World stroke organization: Global stroke fact sheet 2025," *International Journal of Stroke*, vol. 20, no. 2, pp. 132–144, 2025.

[3] M. Bendszus *et al.*, "Endovascular thrombectomy for acute ischaemic stroke with established large infarct: multicentre, open-label, randomised trial," *The Lancet*, vol. 402, no. 10414, pp. 1753–1763, 2023.

[4] M. Goyal *et al.*, "Endovascular thrombectomy after large-vessel ischaemic stroke: A meta-analysis of individual patient data from five randomised trials," *The Lancet*, vol. 387, pp. 1723–1731, 2016.

[5] R. G. Nogueira *et al.*, "Thrombectomy 6 to 24 hours after stroke with a mismatch between deficit and infarct," *New England Journal of Medicine*, vol. 378, pp. 11–21, 2018.

[6] K. S. Asif *et al.*, "Mechanical thrombectomy global access for stroke (mt-glass): A mission thrombectomy (mt-2020 plus) study," *Circulation*, vol. 147, no. 16, pp. 1208–1220, 2023. [Online]. Available: https://www.ahajournals.org/doi/abs/10.1161/CIRCULATIONAHA.122.063366

[7] J. L. Saver, "Time is brain—quantified," *Stroke*, vol. 37, no. 1, pp. 263–266, 2006. [Online]. Available: https://www.ahajournals.org/doi/abs/10.1161/01.STR.0000196957.55928.ab

[8] The Royal College of Radiologists, "Clinical radiology workforce census 2025 report," London, UK, 2025. [Online]. Available: https://www.rcr.ac.uk/media/n1fjvrv4/rcr-2025-clinical-radiology-workforce-census-report.pdf

[9] W. Crinnion *et al.*, "Robotics in neurointerventional surgery: a systematic review of the literature," *Journal of NeuroInterventional Surgery*, vol. 14, no. 6, pp. 539–545, 2022. [Online]. Available: https://jnis.bmj.com/content/14/6/539

[10] Y. C. Shen, A. S. Kim, and R. Y. Hsia, "Treatments and patient outcomes following stroke center expansion," *JAMA Network Open*, vol. 7, no. 11, p. e2444683, 2024.

[11] L. W. Klein *et al.*, "Occupational health hazards in the interventional laboratory: Time for a safer environment," *Society of Interventional Radiology*, vol. 250, pp. 538–544, 2009.

[12] R. D. Madder *et al.*, "Impact of robotics and a suspended lead suit on physician radiation exposure during percutaneous coronary intervention," *Cardiovascular Revascularization Medicine*, vol. 18, pp. 190–196, 2017.

[13] O. A. Berkhemer *et al.*, "A randomized trial of intraarterial treatment for acute ischemic stroke," *New England Journal of Medicine*, vol. 372, no. 1, pp. 11–20, 2015. [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJMoa1411587

[14] H. Robertshaw *et al.*, "Artificial intelligence in the autonomous navigation of endovascular interventions: a systematic review," *Frontiers in Human Neuroscience*, vol. Volume 17 - 2023, 2023. [Online]. Available: https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2023.1239374

[15] A. Chen *et al.*, "Deep learning robotic guidance for autonomous vascular access," *Nature Machine Intelligence 2020 2:2*, vol. 2, pp. 104–115, 2 2020. [Online]. Available: https://www.nature.com/articles/s42256-020-0148-7

[16] C. Abah *et al.*, "Self-steering catheters for neuroendovascular interventions," *IEEE Transactions on Medical Robotics and Bionics*, 2024.

[17] H. Robertshaw *et al.*, "Toward AI Autonomous Navigation for Mechanical Thrombectomy Using Hierarchical Modular Multi-Agent Reinforcement Learning (HM-MARL)," *IEEE Robotics and Automation Letters*, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/11395986

[18] G. Wang, L. Bai, and H. Ren, "How can reasoning capability empower the ai copilot robot in endoscopic surgery," *npj Digital Medicine*, vol. 9, p. 447, 2026. [Online]. Available: https://doi.org/10.1038/s41746-026-02827-8

[19] A. Granados *et al.*, "Evolving surgical teams in the age of artificial intelligence and robotics," *Frontiers in Science*, vol. Volume 4 - 2026, 2026. [Online]. Available: https://www.frontiersin.org/journals/science/articles/10.3389/fsci.2026.1783803

[20] Y. Yan *et al.*, "Machine learning-based surgical state perception and collaborative control for a vascular interventional robot," *IEEE Sensors Journal*, vol. 22, pp. 7106–7118, 4 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9721846/

[21] Y. Xian *et al.*, "A semi-autonomous stereotactic brain biopsy robotic system with enhanced surgical safety and surgeon-robot collaboration," *IEEE Transactions on Biomedical Engineering*, vol. 70, pp. 3288–3299, 12 2023.

[22] J. Zhang *et al.*, "Ai co-pilot bronchoscope robot," *Nature Communications 2024 15:1*, vol. 15, pp. 1–13, 1 2024. [Online]. Available: https://www.nature.com/articles/s41467-023-44385-7

[23] G.-Z. Yang *et al.*, "Medical robotics—regulatory, ethical, and legal considerations for increasing levels of autonomy," *Science Robotics*, vol. 2, no. 4, p. eaam8638, 2017. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.aam8638

[24] H.-R. Wu *et al.*, "Vascular geometry characterization for ai-based endovascular navigation," *International Journal of Computer Assisted Radiology and Surgery*, 2026. [Online]. Available: https://doi.org/10.1007/s11548-026-03742-9

[25] B. Jackson *et al.*, "Manual, joystick, or haptic control? an in vitro comparison of navigation strategies for robotic interventional neuroradiology procedures," *arXiv preprint arXiv:2607.07253*, 2026.

[26] H. Sadati *et al.*, "Autonomous endovascular navigation with a long ctr: In-vitro studies," in *Proc. Hamlyn Symp. Med. Robot*, 2025, pp. 115–116.

[27] C. Eyberg *et al.*, "A ros2-based testbed environment for endovascular robotic systems," *Current Directions in Biomedical Engineering*, vol. 8, no. 1, pp. 89–92, 2022. [Online]. Available: https://doi.org/10.1515/cdbme-2022-0023

[28] L. Karstensen *et al.*, "Learning-based autonomous navigation, benchmark environments and simulation framework for endovascular interventions," *Computers in Biology and Medicine*, vol. 196, p. 110844, 9 2025.

[29] H. Robertshaw *et al.*, "Autonomous navigation of catheters and guidewires in mechanical thrombectomy using inverse reinforcement learning," *International Journal of Computer Assisted Radiology and Surgery*, vol. 19, pp. 1569–1578, 2024. [Online]. Available: https://doi.org/10.1007/s11548-024-03208-w

[30] H. Robertshaw, Y. Hao, W. Deng *et al.*, "Toward real-time autonomous navigation: Transformer-based catheter tip tracking in fluoroscopy," *International Journal of Computer Assisted Radiology and Surgery*, vol. 21, pp. 1071–1079, 2026. [Online]. Available: https://doi.org/10.1007/s11548-026-03647-7

[31] H. Robertshaw *et al.*, "Toward safe autonomous robotic endovascular interventions using world models," 2026. [Online]. Available: https://arxiv.org/abs/2604.20151

[32] ——, "A position statement on endovascular models and effectiveness metrics for mechanical thrombectomy navigation, on behalf of the stakeholder taskforce for artificial intelligence–assisted robotic thrombectomy (start)," *Journal of the American Heart Association*, vol. 15, p. e044931, 4 2026, doi: 10.1161/JAHA.125.044931. [Online]. Available: https://doi.org/10.1161/JAHA.125.044931

[33] J. Brooke, "Sus: A "quick and dirty" usability scale," in *Usability Evaluation in Industry*, P. W. Jordan, B. Thomas, B. A. Weerdmeester, and I. L. McClelland, Eds. Taylor & Francis, 1996, pp. 189–194.

[34] *ISO 9241-11: Ergonomics of Human-System Interaction – Part 11: Usability: Definitions and Concepts*, International Organization for Standardization Std., 2018.

[35] C. Yang *et al.*, "A review of human-machine cooperation in the robotics domain," *IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS*, vol. 52, 2022. [Online]. Available: https://doi.org/10.1109/THMS.2021.3131684.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- NF2026AIcopilotinvitrouserstudySUPPMATERIAL.pdf
- NF2026AIcopilotinvitrouserstudyVIDEO.mp4.mp4