Federated Explainable Hierarchical Imitation Learning (FEHIL) for Homecare Robotics

Harrison Field

***Abstract*—Data privacy and explainability concerns represent a barrier to the widespread adoption of AI-assisted care robotics. To address this, the application of Federated Leaning (FL) to an explainable Imitation Learning (IL) framework is explored to better understand the trade-offs between data privacy, explainability and performance in robotics systems. Here, robotic liquid pouring is used as an example to evaluate the performance of IL with the federated approach, which we denote as Federated Explainable Hierarchical Imitation Learning, or FEHIL. Through this process, it is observed that improving data privacy can yield comparable model performance while maintaining explainability at the cost of longer training.**

***Index Terms*—Federated Learning, EHIL, Data Privacy, Explainability, Robotics, Imitation Learning, Behaviour Cloning.**

I. Introduction

This report begins by investigating the barriers to adoption regarding robotics in healthcare. The justification for robotics in healthcare and the primary barriers to adoption are represented in section I, with a more detailed breakdown of the contextual and technical background of this paper provided in section II. The primary contribution of this paper is the combination of two techniques: Explainable Hierarchical Imitation Learning (EHIL)[77] and Federated Learning (FL)[52], to develop an AI-driven robotic pouring system that inherently incorporates principles of explainability and data privacy within its foundational design. We term this combination of techniques Federated Explainable Hierarchical Imitation Learning (FEHIL), explained in detail in section III.

A. Background and motivation

According to the Care Quality Commission UK, the National Health Service (NHS) is undergoing a continued deficit of available staff. Service providers are losing the battle to attract and retain enough staff, with 9 out of 10 NHS leaders warning of a social care workforce crisis in their area. Care homes are finding it particularly difficult to attract and retain registered nurses, with qualified workers moving to better-paid jobs under more favourable conditions. As a result, care homes have had to stop providing nursing care. Among the providers who reported workforce pressures to the commission, 87% of care home providers and 88% of homecare providers declared recruitment challenges. Only 43% of NHS staff asserted they could meet all the conflicting demands on their time at work [73].

Deficits in available carers are observed internationally. Nursing Home Abuse Justice, a USA based organisation committed to combating abuse within care home facilities, claims nursing homes are understaffed because of; staff turnover, labour costs, false reports of staffing levels, overtime pay issues, and crucially a lack of applicants in many cases [65]. In Europe, research has been conducted into the relationship between understaffing of nurses and patient safety in hospitals[30] with the explicit purpose of increasing the knowledge of understaffing of hospital nurses, and the consequences that understaffing may have on patient safety [30].

Amongst this shortage of care workers is an ageing population, where increasing proportions of an area's population require some form of assistance with their lives due to the ageing process. This can range from occasional assistance with tasks to full time care, resulting in reduced independence for the patient and increased burden on care institutions. According to the World Health Organization, independent living defines good ageing as "the ability to perform functions related to daily living"[58]. Furthermore, an EU innovation policy concerning European healthcare systems positions independent living as a way of relieving European healthcare systems of their understaffing burdens, with the idea that technology innovations will assist patients at home and prevent unnecessary hospitalisation. Specifically, robotics is championed as an area with promise to allow elderly people to live independently for longer[30].

Politically speaking, robotics is seen as a vital future market in which no governmental body can afford to have its jurisdiction lag behind [30]. With the socio-economic realities of the healthcare system as they are, Dr. Renda makes a compelling argument in favor of incorporating AI and robotics into the healthcare sector. He identifies several critical factors supporting this stance, including the challenges posed by an aging population, workforce shortages in healthcare, the need for more personalized care approaches, concerns about healthcare resource utilization, increased prevalence of lifestyle diseases, extended diagnostic processes, and the disproportionate surge in chronic and non-communicable diseases among underserved communities [19].

To the patient at home, service robots would be the most likely solution. A service robots is a "robot in personal use or professional use that performs useful tasks for humans or equipment"[35]. From a patient-centric perspective, service robots can fulfill essential functions, such as providing comfort care and offering personal assistance to patients, including helping with mobility, feeding, and everyday activities [63]. In this scenario, the patient would receive care from the robot entirely at home, where the robot is able to assist in either the delivery or actual medical care to patients [63]. If service robots demonstrate their effectiveness in this context, their

widespread implementation could yield advantages such as enhanced productivity, consistent service quality, and decreased staffing expenses [43]. Robots in this setting would typically enable organisations to gather data from the setting, analyse the data and adapt care to changing needs as those needs change in real time [43].

Given the sensitive nature of medical data and the physical risks of robots interacting with patients, some have claimed that humans should always be ultimately responsible for decisions made by robot-governing AI [19]. Concerning the ethical considerations pertaining to robotics in healthcare, it is worth noting that the challenge lies not in comprehensive identification of all ethical concerns (as elucidated in sections I-B and II), but rather in the formulation of effective implementation strategies [19].

The widespread implementation of AI-governed robotics systems brings with it new legal challenges as well as ethical ones. The absence of existing liability or ethical frameworks makes it difficult for law enforcement agencies to keep up with technological innovation in this area [19]. The deployment of AI in healthcare may assign an elevated responsibility to corporations, clinics, and public entities in acquiring, utilizing, and safeguarding patient health data, giving rise to concerns regarding privacy and data security during implementation [55]. Because of this, AI and robotics technologies intended for care scenarios should be designed in such a way that promotes responsible use from the beginning without stifling innovation. The next section (I-B) further explores data privacy and explainability in care robotics, two areas identified as key barriers to adoption which academic research is well suited to addressing.

B. Problem statement and research objectives

Over the past 20 years, AI-governed robotics in healthcare has evolved in the discourse surrounding EU work programs and policy agendas. It has gone from a virtually nonexistent topic to a firmly established contingent product of a range of technological, social and political processes [48]. During this time, commercial robotics and AI have undergone considerable development, enabling enterprises to manufacture and distribute products on an unprecedented scale [18]. Assistive robots differ from their commercial counterparts in many ways, some key differences include operating in uncontrolled environments (patient's homes), physical human interaction, and the use of sensitive medical data. Due to these considerations, roboticists must focus on creating robots capable of operating as self-monitoring independent entities and enhancing their behaviors through experiential learning outcomes [51]. These ideas resonate with the rest of this section, where data privacy and explainability are further explored as key areas for development if robotics in healthcare is to be widely adopted.

To excell at a given task, service robots often require knowledge or the ability to gather it about their patient [63]. This presents the core privacy issue in that the patients' personally identifying information such as; voice, image, name, medication, etc is both gathered and considered highly sensitive. One issue with gathering sensitive data is that private organisations often implement care robotics and AI systems, meaning that any data gathered is managed by a private entity. Recently, this has resulted in the poor protection of privacy manifesting in the misuse of data [55][19].

In 2016, private company DeepMind partnered with the Royal Free London NHS Trust Foundation to develop Machine Learning (ML) techniques capable of assisting with acute kidney injury [16]. Critics of the partnership stated that patients did not have adequate access or control over their data and that privacy impacts were not adequately discussed. Because of the outcry, DeepMind's parent company Alphabet took over the project and as a result the patient data was moved to the USA from the UK [55]. Regarding the issue, Murdoch emphasizes that the potential to transfer large volumes of private patient data to another jurisdiction is a novel aspect of big data, especially when integrating commercial healthcare AI. This situation arises due to the unequal distribution of technological expertise within prominent tech companies, which can result in public institutions becoming increasingly reliant and less cooperative in health tech implementation. Furthermore, Murdoch underscores the significance of server and computer location and ownership in storing and accessing patient health data for healthcare AI purposes. Regulatory measures should mandate that patient data remains within its originating jurisdiction, with limited exceptions[55].

Exploring this area further, a 2018 American survey of 4,000 adults found that only 11% were willing to share health data with tech companies, with 72% willing to share the same data with physicians. The public opinion of tech companies' proficiency with data was clear, with only 31% "somewhat confident" or "confident" in tech companies' data security [55]. From this, we can discern that the American public is not comfortable with it's health data being shared with private companies. Despite the public desire to have their personal and private information retained, hospitals are engaging in business deals to share patient data with private entities in the tech sector [24][34][72][7]. Alongside this, healthcare data breaches have become more common in recent years, with AI algorithms contributing to the growing inability of countries and organisations to protect health information [55]. A number of recent studies have shown how emerging computational strategies can be used to identify individuals in health data repositories managed by public or private institutions, including data that has been anonymized and had all identifiers removed [55] [13][33][56][23][37].

While these case studies do not relate to homecare robotics specifically, the need for security in this sector has been made clear. A recent white paper published by the UK-RAS network directly calls for innovations in healthcare robotics, to imbue these devices with privacy and explainability from the start. Among the main concerns are the dangers of compromising patient data confidentiality, data reliability and the potential weaponisation of homecare robotics systems [50].

The case studies mentioned here along with the academic

literature serve as principle inspiration for the work conducted in this report. The misuse of patient data negatively impacts public opinion surrounding AI in robotics health care, while the growing ability to invade patient privacy through ML methods has the potential to do real harm to patients. These outcomes are unacceptable, the patients in need of care deserve to have their privacy, safety and dignity respected to the highest standards. If entities continue to disregard their impact on those to whom they provide care, the ramifications for health technology could be stifling in the long term. This scenario would simultaneously continue to harm current patients and deny future ones the benefits of continued innovation. Motivating this report is the imperative for a paradigm shift in healthcare data management practices, which necessitates a recalibration towards heightened patient agency in the stewardship of their own data assets.

In order to achieve a better system of robotics in healthcare, it has been argued by Pierce that robotized systems should be aligned with values such as transparency, accountability, explicability, auditability and traceability, and neutrality or fairness [19]. These values will be present throughout this report, with a key focus on accountability and explicability. Transparency, auditability and traceability are not directly addressed in this paper due to their omnipresent nature in the implementation process, going beyond academic research. While the same can be said for accountability and explicability, this report explores the possibility of imbuing systems with these qualities inherently from their design.

Explicability in ML systems refers to the "black box" problem, which pertains to the opacity and lack of interpretability exhibited by certain complex ML models. In the realm of academic discourse, this issue revolves around the challenge of comprehending and elucidating the inner workings, decision-making processes, and feature attributions of such models. In the realm of healthcare robotics, the issue of opacity extends to the utilization and manipulation of health and personal data when adequate safeguards are lacking. In response to this concern, numerous researchers have been working on creating interpretable AI models designed for seamless integration into medical practices [55][1]. When considering AI-driven care robots, one must consider this matter not only from the standpoint of system development but also from the perspective of the patient. It's crucial that both the development team and the user/patient possess an understanding of how decisions that impact them are made. This is especially pertinent in the context of systems operating with some degree of autonomy [19].

In summary, the design of an AI-governed robotics care system should inherently prioritize explicability and data privacy. This report leverages the EHIL technique as an exemplar of transparent AI solutions (further elaborated in sections II-A and III). EHIL, a hierarchical machine learning technique tailored for achieving explicability in care robotics during behavioral cloning [77], serves as a starting point for implementing robust data privacy measures. Initiating with an explainable approach allows for the introduction of data privacy methods and subsequent model retraining to observe their effects. The objective is to preserve both model performance and explainability while enhancing data privacy, marking a significant stride towards empowering patients in managing their own data assets.

Regarding the data privacy measures explored in this report, Federated Learning (detailed in sections II-C and III) transforms the machine learning model training process into a distributed system across numerous devices. Practically, this ensures that datasets remain localized, eliminating the risk of central data aggregation and potential misuse. This approach not only aligns with the principles of continuous learning, as advocated by Matsuzaki and Lindemann [51], but also directly addresses data privacy concerns mentioned in the case studies within this section and complies with the recommendations of Murdoch, who emphasizes that patient data should remain within the jurisdiction of its origin [55].

When integrating Federated Learning into the EHIL framework, the hypothesis is that the FEHIL model will demonstrate performance comparable to the original non-federated EHIL model. While the EHIL model's detailed explanation is provided in sections II-A and III, it is pertinent at this stage to understand that it comprises multiple machine learning models trained on image data for performing classification and regression tasks. Specifically, in the classification scenario, the FEHIL model is expected to achieve accuracy within a 5% margin of the EHIL approach, while in regression tasks, the federated model is anticipated to achieve a Mean Squared Error (MSE) within 0.05 of EHIL's performance. These predictions are founded on the research discussed in section II-C, where experiments akin to those planned in this project are found in [61] and [49].

In order to complete the report and experiment outlined here, it will be necessary to acquire benchmark results by training the EHIL model, prepare the Federated Learning environment, design the Federated Learning implementation, prepare the EHIL approach for Federated Learning, organise the data for federated learning, train the FEHIL model, compare and interrogate the results of the federated approach with the original EHIL, draw conclusions, discuss the impact of the work with it's achievements and limitations, and propose future research directions.

To the best of our knowledge, this is the first exploratory work of combining Explainable Artificial Intelligence (XAI), federated learning and imitation learning. It is hoped that the combination of these technologies can contribute to future applications in homecare and healthcare robotics.

II. Literature Review

In this section, the supporting literature is explored, concluding with the rationale behind applying FL to EHIL. Within this, we explore explainability within the field of AI and some specific implementations in robotics before selecting EHIL as a suitable example for this project. From here, privacy preservation techniques are considered based on their compatibility with the EHIL approach. In section II-A, explainability

in ML is explored and the requirements for explainability in robotics are outlined as they are understood from the literature and legal publications. The section rationalises the decision to employ EHIL as an example explainability framework and continues to outline the concept of EHIL, where the characteristics making it suitable for this project are mapped to their respective sections of this paper. Section II-B builds on the research already conducted in section I to contextualise the need for privacy solutions in care robotics systems while exploring the available techniques. Here, each technique is considered for its merits and drawbacks, with a rationale as to why it was not suitable for the purposes of this project. Section II-B ends with the introduction of FL as a suitable solution, before exploring FL in-depth in section II-C. Here, the concept of FL is rationalised as the appropriate solution to privacy preservation in care robotics AI training. Within this, criticisms of FL are acknowledged and explained in terms of their relevance to the overall aims of this project, while also considering similar implementations of FL, where a blending of techniques is seen to yield positive results.

A. *Explainability in Robotics*

Explainability is required in ML applications for the purposes of trust, causality, transferability, informativeness, fair and ethical decision making, accountability, making adjustments and proxy functionality. Fostering these qualities is necessary in ML due to the paradigm of estimation rather than complete logic as is seen in other algorithms [9]. The difficulty, however, with explaining current ML solutions is that "the models are complicated because the problem is complex and it is almost impossible to explain what exactly the models are doing and why they are doing it" [9].

At this point, we mention the difference between explainability and interpretability. Although there appears to be some conceptual overlap in how these terms are presented in the literature and in the media, this report aligns with Burkhart and Huber, where "interpretability is used in terms of comprehending how the prediction model works as a whole. Explainability, in contrast, is often used when explanations are given by prediction models that are incomprehensible themselves" [9]. The manner in which these explanations should be articulated remains a subject of ongoing research within the XAI community. They aim to create tools and techniques that unveil the inner workings of Black-Box AI solutions by generating human-comprehensible, insightful, and transparent explanations of AI decisions. [18].

Explainability of AI in fields such as healthcare is of particular importance due to the proxy functionality explainability affords in such systems. Once a system can provide explanations for its decisions, it can be examined based on other criteria such as safety, nondiscrimination, privacy, robustness, reliability, usability, fairness, verification and causality [21]. This in turn enables system engineers to properly address areas of concern, aiding in the overall ML-system development process [6]. The difficulty here is that there is no common or legal consensus on how explainability should be conducted, with the data protection authorities in Europe noting that automated processes "should find simple ways to tell the data subject about the rationale behind or the criteria relied on in reaching the decision," but not "a complex explanation of the algorithms used or the disclosure of the full algorithm. The information provided should, however, be sufficiently comprehensive [...] to understand the reasons for the decision" [3]. Such a requirement leaves the implementation of explanations open to debate, while article 22 of GDPR [10] exemplifies a "right of explanation", granting individuals the right to obtain an explanation of the outcomes automatically generated by an AI solution, as well as to challenge and evoke a pertaining reference, particularly when it may adversely impact a human legally, financially, physiologically, or mentally [18].

From a patient's perspective, the key benefits of an ML-system imbued with explainability are accountability on the part of the system owner and an increased trust in the system. Accountability becomes possible when the decisions of a system can be justified and defended. This results in relationships between AI-system owners and users where the developers/owners can be held legally liable for the actions of their machines based on the explanations they provide. Because of this, patients and users in the general case can build trust in these systems knowing that there is a strong motivation for the system owners to hold the patients' safety in high regard. While this area is still in development, the EU has already been using their GPDR regulation to resolve disputes related to the poor inference of ML models [18], because of this, it is vital for the ML community to develop and deploy explainable systems to protect both patients and commercial interests.

With regards to explainability in robotics, practitioners and academics continue to propose novel methods. Often, the explainability process becomes associated with a human teaching the robot how to perform a given task. One popular way to achieve this is through the use of Augmented Reality (AR) to convey information to the user regarding the robot's behaviour. In [47] and [67], Microsoft Hololens is used to convey the robot's intentions as both an explainability mechanism and a debugging tool. While these techniques show promise, it must be recognised that most patients will not have access to expensive technology such as the Microsoft Hololens.

Furthermore, other techniques allow the use of multimodal data when generating human-readable explanations. In [26], the authors are able to generate natural language explanations from robots. This falls into the category of question and answer scenario such as Q: "Why did you turn left there?" A: "I noticed someone at the end of the corridor.". Indeed, visual and question answering explanations currently form a large portion of robotics explanations [18]. Here, this is achieved by a natural language generator, which employs natural language templates to answer questions from human users about the robot's decisions and actions.

Another way we can achieve a system capable of answering questions about its current actions is through the creation of a knowledge graph. Knowledge graphs convert different types

of data into a uniform graph format that can be used as an integrated knowledge base. These can be used to expose human-like explanation through the recognition of objects and the associated action with this recognition [18]. It is this framework of explainability that the EHIL [77] approach conveniently aligns to, providing an explanation framework typical for robotics and suitable for providing human-like explanations seen to be desired by the research community and regulatory bodies discussed in this section.

It is for these reasons that the EHIL approach was chosen for this project, as it enables the creation of a logical graph for explanation to the user, and the system can be trained using only visual data and teleoperation. This, along with the generalisability of the approach and the use case of robotic pouring, directly align with the required features of AI-governed robotics solutions mentioned in section I. Given the cultural move towards explainability requirements (sections I-A, I-B and II-A), the legal ambiguity of these requirements (sections I-B and II-A) and the evolving categories of explainability in AI specifically relating to robotics (section II-A), the EHIL approach is a natural starting point for developing robotics systems capable of both explainability and data privacy inherent in their design.

Explainable Hierarchical Imitation Learning (EHIL) is an XAI framework employing Imitation Learning techniques to robotics applications. Furthermore, EHIL can be applied to any robotics task that can be broken down into a series of sequential steps through its knowledge graph representation which functions as the explainability vehicle. Within Imitation Learning, behavioural cloning is the specific technique applied within the EHIL framework. Behavioural cloning works analogously to supervised learning techniques, in that a collection of data - value pairs are provided to an ML model, which finds patterns in the data examples to map them to the desired value. In the EHIL approach, a robot is teleoperated by a human expert to perform a specific task, where state - action pairs are recorded as training data for the EHIL model, building on the following literature regarding demonstration-based imitation learning [71][69][2] and the following imitation learning methods [28][14][15]. In the case of EHIL, robotic pouring is used as the example task to be completed, and the state - action pairs correspond to the video recording of the pouring session and the angle of pour conducted by the robotic arm. The individual frames of the video are paired with the angle of pour at the same time step to create the training data for the model to "clone" the behaviour of the human demonstrator.

The aforementioned cloning behaviour forms the decision-making module of the EHIL framework. On its own, this module is not explainable, as there is no contextual information as to why the action was decided. To imbue the system with explainability, two image classifiers are added to the model which classify the video frames from the cloning module into contextual phases and transitional states of the overall task. Through this classification of video frames, a logical graph can be constructed of the sequence of steps involved in the completion of a task, enabling users to conduct safety checks before task execution or reason the failures after task execution, making the decision making process intuitive and explainable. [77]

Overall, the system is composed of three ML models referred to as hierarchies within the EHIL system, described as H1, H2 and H3. H1 and H2 are classifiers forming a logical graph, while H3 conducts the decision-making process through outputting the action value of the robot for each image frame it receives. While the proposed EHIL framework is novel, it builds upon the following hierarchical learning literature [59][27][76][75]. Specific information regarding the model architecture is further outlined in section III.

B. Data privacy in model training

As described in section I-B, data privacy is of paramount concern within the field of healthcare, with the main types of scurity-related issues in ML being evasion attacks during model inference and poisoning attacks during model training [5]. In this area, distributed ML model training plays a key role in system development due to the agility and security this affords the institutions concerned. Healthcare, like many other industries, is vulnerable to data breaches (as discussed in section I-B), which are a major problem for organisations with centralised data [11].

In the context of sensitive medical data, primary concerns revolve around re-identification attacks, which involve identifying individual data sources even when data anonymization techniques have been applied based on other information in the datasets. Additionally, concerns include dataset reconstruction attacks and tracing attacks, also known as membership inference, which infer the inclusion of a specific individual in a dataset [74].

Distributed model training in commercial practice is the process of training an ML model across various locations, often using slightly different data gathered at those locations. This has parallelisation benefits, enabling organisations to take advantage of their existing hardware rather than outsourcing compute resources to third party organisations. Inherent within this, is the benefit of reduced latency and bandwidth savings, where the volume of data communicated across an organisation's network can be vastly reduced. Complimentary to both of these benefits is the resultant increase in security with regards to sensitive data, as raw data is communicated less frequently [74]. With the context of this report in mind, distributed model training appears as a timely ally to those seeking to implement and adopt AI-governed systems in the field of healthcare. However, there are still significant concerns regarding the security of data and privacy of users in distributed ML solutions. As such, the rest of this section will investigate the main solutions employed to improve privacy of such systems and use this research to justify the decision to apply Federated Learning to the EHIL framework.

1) Secure Multi-Party Computation: Secure Multi-Party Computation (SMPC) is a cryptographic technique within the realm of secure computation that enables multiple parties to collaboratively compute a desired function on their private

inputs, while preserving the confidentiality of individual inputs. Unlike traditional computation methods that require data sharing, SMPC ensures that no single participant gains access to the complete input data of others, thereby safeguarding sensitive information. Through the utilization of advanced cryptographic protocols, such as homomorphic encryption, secret sharing, and oblivious transfer, SMPC guarantees that computations are carried out in a distributed and privacy-preserving manner. This field finds applications in scenarios where mutually distrusting entities seek to collectively analyze data while upholding the principles of confidentiality and privacy. The study of SMPC encompasses the exploration of secure protocol design, cryptographic primitives, and the theoretical foundations of privacy-preserving computation [31].

In summary, SMPC is considered secure when the involved parties learn only the final result, and no other information regarding the computation they have assisted in calculating [11]. An illustrative real-world application of SMPC, as described by Byrd and Polychroniadou, involves a scenario where a group of employees seeks to calculate the average salary without disclosing their individual salary information to one another. This task can be accomplished through MPC, where the only information exposed is the computed result, specifically the average salary. Each pair of employees holds a sizable shared number, with one employee adding it to their salary and the other subtracting it. Despite this manipulation, the outcome of the calculation remains unchanged, while the actual salary of any individual employee remains undisclosed. [11].

In ML, the advantage of SMPC is the evaluation of one party's model using another party's private data, without seeing the data [41]. The primary drawback, as highlighted by Wang, pertains to the computational burdens associated with Multi-Party Computation (MPC) protocols. In machine learning algorithms, these protocols introduce a substantial computational overhead, increasing costs by a factor of 32 and necessitating an additional round of broadcasting among MPC servers. Furthermore, operations that are computationally trivial in plaintext, such as Softmax, ReLU, and other non-linear operations, become notably expensive due to the added communication overhead. These increased computational demands render the deployment of MPC in real-time machine learning inference frameworks less favorable [70]. While research has been conducted into the reduction of this computational overhead [70], this still falls short of what's required for the real-time inference in a robotics task-completion setting. It is for this reason that this technique could not be chosen for this project.

2) Homomorphic Encryption: Homomorphic encryption is a cryptographic technique that enables computation on encrypted data without the need for decryption. Unlike traditional encryption methods that require data to be decrypted before performing operations, homomorphic encryption allows mathematical operations to be conducted directly on the encrypted data. This unique property preserves the confidentiality of sensitive information while still permitting computations to be performed on it [29].

This presents a pivotal advantage within ML scenarios by enabling secure computations on encrypted data. This cryptographic technique ensures the confidentiality of sensitive data throughout the entirety of the analytical process, including data preprocessing, model training, and inference. By permitting mathematical operations to be conducted on encrypted data without the requirement for decryption, homomorphic encryption safeguards against unauthorized data exposure. This advantage addresses the previously mentioned concerns regarding privacy breaches in distributed model training where multiple parties can contribute data without revealing their individual datasets.

Unfortunately, the nature of homomorphic encryption limits the range of computations that can be completed, and therefore also limits the kinds of statistics and machine learning algorithms which can be implemented [4]. Until recently, the application of homomorphic encryption has been limited to non-standard ML models which have not proven accurate with practical and advanced datasets. Even with modern innovations, which apply homomorphic encryption to ResNet-20, there is a direct trade-off between security and inference time, making the adoption of homomorphic encryption for this project untenable. In [43], the proposed model achieves state of the art accuracy, but with a four hour inference time on one image. This inference time is with 98-bit security, the minimum required to be considered secure, and would require even more time if the security were to be increased to the standard 128-bit encryption. Furthermore, homomorphic encryption (like SMPC) can increase the size of the data it encrypts to the degree that it would significantly impact the practicality of transferring the quantities of data required to train the EHIL solution in a distributed manner. Encrypting an integer in homomorphic encryption may explode it's size from 4 bytes to more than 20 kilobytes [32][12]. Because of the reasons discussed here, it was deemed impractical to consider homomorphic encryption as a suitable solution to combine with EHIL.

3) Differential Privacy: Differential privacy (DP), as outlined in the work of Dwork [22] and McSherry [53], is a mathematical concept aimed at ensuring the statistical indistinguishability of individual inputs through the introduction of value perturbations. The adoption of differentially-private machine learning algorithms in centralized settings has garnered significant attention in the literature and has been implemented by major companies, such as Apple, which employs it in web search auto-completion. The incorporation of differential privacy introduces an element of randomness, making it so that even adversaries armed with additional information are left with uncertainty regarding the original values. It's important to note a clear trade-off: while the addition of randomness to collected data safeguards user privacy, it does come at the expense of accuracy. When properly applied, differential privacy strikes a balance, allowing for the extraction of meaningful insights from aggregated data [11]. On the point of the privacy/performance trade off, the noise added to the gradient

of an ML model is a significant barrier to optimization, resulting in significantly reduced performance compared to standard training [20][40][42].

Recent advances in applying DP to ML has seen significant improvements, gaining state of the art results on image classification tasks employing techniques similar to that of EHIL by replacing batch normalization layers with group normalization layers (depending on less statistics), using weight standardisation in convolutional layers, and applying parameter averaging techniques to smooth out the noise added by DP and reduce the variance of the final model [17]. The reason this technique has not been chosen for this project is because of the trade-off inherent in its design. When applying privacy budgets to visual data, it is not entirely clear at what point privacy has been maintained, and what users constitute as private. While visual data of a patient's written details may be obscured, the overall image of a vulnerable person receiving care may be preserved to the extent that the individual could be identified from the frame. It is because of this ambiguity that other solutions were explored for this project.

4) How Federated Learning is Different: Unlike the other techniques described in section II-B, FL circumvents the need to obfuscate the contents of training data by training models locally at the location the data was gathered. In this way, no centralised storage location is used and therefore cannot be vulnerable to attack in the same way. Xianjia et al. promote FL as "the" solution to using ML at the edge while preserving data privacy [74]. This is because a recent survey concluded that FL is more susceptible to security risks than it is to privacy risks, and these security risks are typical of any ML system requiring the communication of multiple participants where there are known mitigation methods implemented in practice [54]. This being said, while FL inherently deals with data ownership and governance issues, it cannot guarantee privacy and security by itself. Integration of other techniques is required in practice to produce a robust FL framework. In the next section, FL is considered in detail, investigating the core concept, criticisms, and existing implementations similar to that conducted in this project while acknowledging the techniques currently employed to bolster the security of FL systems.

C. Federated learning

1) Concept: Federated learning is a decentralized approach to machine learning that enables collaborative model training across a network of devices or nodes (called clients) while keeping data localized. Instead of centralizing data in a single location, FL allows models to be trained directly on individual devices, preserving data privacy and security. These local models are then aggregated into a global model through a central server while incorporating updates from each client using the federated averaging algorithm, taking a weighted average of the client model weights. This technique has been found to be robust to non-IID data and fast to train in decentralised settings. Compared to the work it builds on in synchronised stochastic gradient descent, FL reduces the communication rounds required to train a model in a distributed fashion by 10-100x. [52]. The main research directions for FL are currently deployment of FL in resource-constrained embedded systems, communication-efficient FL, energy-efficient FL, and privacy-preserving federated edge learning with the aim to improve the learning performance in networks where the general assumption is that resources are inherently at the edge [68][74].

Before implementation of FL to EHIL is conducted, it is recognised that EHIL is not a perfect candidate for FL. This is due to the nature of the data labelling process, whereby humans must first demonstrate the action of the robot via teleoperation and then label the image data gathered into the necessary classes to train the model. Ideally, FL is applied to situations where the labels for data can be gathered from contextual analysis carried out automatically on the client device. In this project, the techniques employed are unsuitable to be trained further using only patient interaction, and instead would require a professional to handle the teleoperation of the robot. This limitation considered, it would still be possible to implement robotics solutions using EHIL in controlled settings where professionals have access to the robotic equipment for training. This limitation is primarily related to scalability of the approach upon implementation and is explored further in section VI.

2) Privacy and Security: In reference to the context of this project, FL is a privacy-first approach to distributed model training. The experiments demonstrated in [52] are created in response to a White House report on the privacy of consumer data [66], and are formulated in such a way as to represent real-world settings where natural-language and computer-vision models can be trained using sensitive data gathered from the use of personal devices. It is noted in the paper that this kind of data greatly improves the usability of models trained using it, as the data is a true representation of the domain's data distribution (or a much closer approximation of it than is possible to obtain in controlled environments). By using a decentralised approach, privacy and security is bolstered by limiting the attack surface to only the device where the sensitive data is held as opposed to the cloud.

However, if a copy of the global model were to be obtained by malicious actors, it may still be possible to infer details of the training data from the trained model weights or parameters as demonstrated in [57] and [62]. Because of this, it has been argued that FL by itself does not provide the levels of security and robustness required by today's standards in distributed autonomous systems and must be combined with other techniques [74].

Differential privacy (section II-B3) is widely utilized in FL to enhance privacy by adding noise to model updates during aggregation. This mechanism prevents the extraction of private client information from the aggregated model updates. Byrd and Polychroniadou noted the prevalence of this approach, highlighting that most FL systems use DP to introduce noise to the parameters [11]. DP provides a robust defense against potential privacy breaches caused by malicious actors, with

the familiar trade-off between privacy and performance as explained in II-B3, where [44] has directly applied DP to FL in a brain tumour segmentation task.

SMPC (section II-B1) is also used in the FL setting, often in conjunction with DP to protect against extreme forms of attack [11]. When these techniques are combined, it becomes much more difficult to reverse-engineer data and what can be gathered is never the exact data of any user, which can be seen implemented in [8][36]. While these are promising innovations, the drawbacks of these techniques discussed in section II-B are still prevalent in the FL setting. The dynamic landscape of FL necessitates ongoing research into advanced security techniques that strike the right balance between privacy and utility, where novel solutions are being presented such as dispersed FL in order to cope with scenarios where the central server is compromised [39], Byzantine-robust FL to protect against model poisoning attacks [25] and the integration of blockchain technology in communication protocols [74].

3) Dataset availability issues: While the preserved privacy of patient data is considered a beneficial quality of FL overall, its drawbacks require mentioning. In implementation of ML systems, the data fed to the model is of great importance and directly relates to the model's performance. It is for this reason that FL is required, so that the model can access the sensitive but valuable patient data. The issue with this paradigm is that system engineers cannot analyse the data for outliers that are inhibiting the model's performance. Kairouz and colleagues bring attention to a significant concern, emphasizing that experienced modelers frequently engage in direct data subset inspection for various tasks. These tasks encompass fundamental sanity checks, debugging misclassifications, identifying outliers, manually labeling examples, and uncovering bias within the training dataset. The challenge lies in devising privacy-preserving methods capable of addressing these inquiries in decentralized data settings, and it remains a substantial open problem within the field [38].

4) Comparable implementations of Federated Learning: The decision to apply FL to EHIL is not solely based on the ability of FL to circumvent many of the privacy concerns regarding distributed data gathering/model training outlined in this report. In recent works, FL has shown promising performance in handling visual data for both classification and regression scenarios. In [60], FL is demonstrated to improve the performance of an image classifier when applied to breast density by 6.3%. Similarly, another imitation learning application shows promise in [46], where heterogeneous data is used to train a model to perform an autonomous driving task. Together, these examples cover the functionality EHIL demonstrates (as described in section II-A) and support the research hypothesis outlined in section I-B.

Furthermore, FL can be seen applied to medical data as early as 2019 in [61] due to it's privacy-preserving qualities, and in robotics with multi-agent trajectory forecasting in [49]. However, none of the papers mentioned in this research have been able to combine FL with explainable Imitation Learning techniques. While the research referenced here points to FL working well with the composite parts of EHIL, this project conducts the necessary work to observe the effects of FL on EHIL from the perspective of barriers to adoption. By combining the composite parts of EHIL in an FL framework and observing the effects from an alternative perspective, this project gains a deeper understanding of federated imitation learning as it relates to the political and ethical climate at the time of writing.

III. Methodology

A. Problem Formulation

At the heart of this project lies the innovative fusion of two pivotal paradigms—EHIL and FL. The primary motivation behind this integration emanates from the imperatives of contemporary Artificial Intelligence, specifically within the realm of robotics. The primary contribution of this endeavor centers on harnessing the power of EHIL and FL to bolster privacy within XAI robotics systems. In the current landscape, where concerns over data privacy loom large, the integration of these methodologies presents a timely solution to bridge the gap between effective machine learning and ethical considerations.

The core objective of this endeavor is to chart a path toward establishing the viability and efficacy of the FEHIL framework within the care robotics industry. In an era marked by the emergence of robotics in caregiving contexts, the integration of FL and EHIL presents a promising avenue for enhancing task execution while ensuring the highest standards of privacy. By exploring the intersection of these methodologies within the care robotics sector, this study ventures beyond the theoretical and seeks practical applications to ground this study's contributions to the field.

Central to this research's empirical foundation is the comprehensive comparative analysis of EHIL and FEHIL. Drawing inspiration from evaluation methodologies established in [77], this study aims to quantify the impact of privacy enhancements induced by FL on model performance. By applying similar evaluation metrics, the study endeavors to unravel the extent to which the fusion of EHIL and FL elevates both predictive proficiency and explainability, thereby aligning with the overarching themes of this project—effectiveness and transparency.

In pursuit of a seamless connection between previous work and current research, the chosen use case for experimentation is robotic pouring. This use case serves as an illuminating bridge between EHIL's foundations and the unique privacy-enhanced framework established by this study. By delving into a familiar domain, the study ensures a cohesive comparison while unearthing insights that hold relevance within broader care robotics applications.

B. Federated Learning

1) Introduction to Federated Learning: Federated Learning [52] stands as a pivotal innovation within the realm of machine learning, infusing systems with a paramount aspect: privacy. This methodology operates on the principle of decentralised collaboration, enabling disparate devices or servers to

collectively refine models while preserving the integrity of localised data. Beyond deployment, FL empowers robotics systems to perpetually evolve through continuous training post-deployment. With the current landscape marked by ethical quandaries regarding privacy and legal responsibility, this confluence of FL and EHIL holds the promise of bestowing clarity upon a field grappling with ambiguity, thereby fostering its ethical growth and robust development.

2) Key components of Federated Learning:

1) **Central Server:** At the heart of Federated Learning lies the central server, which orchestrates the collaborative training process. This server initiates the training by providing an initial model to the participating local clients. It aggregates and processes the model updates received from these clients, ensuring the gradual refinement of the global model. The central server's role is critical in harmonising the contributions of individual devices into a cohesive and improved model.
2) **Local Clients (or Edge Devices):** The local clients, often referred to as edge devices, encompass a range of devices distributed across the network. These could be smartphones, IoT devices, or servers at the edge of a network. Each local client possesses its unique dataset, typically representing a specific slice of the overall data distribution. Local clients download the global model, perform local model training using their private data, and subsequently generate updated models containing their learning insights. These models are then forwarded to the central server for aggregation.
3) **Communication Protocol:** The communication protocol serves as the conduit for exchanging model updates between the central server and the local clients. This protocol encompasses the methodologies and algorithms that manage the transfer of information. It must strike a balance between ensuring data privacy while enabling efficient and reliable communication.
4) **Iteration and Aggregation:** Federated Learning operates iteratively, spanning multiple rounds of model updates. In each round, local clients independently perform training on their data, generating improved models. These client models are then aggregated at the central server, typically through techniques like weighted averaging or more sophisticated aggregation strategies. The aggregated model, enriched with insights from diverse data sources, becomes the foundation for the subsequent round of training.
5) **Privacy and Data Security Mechanisms:** Integral to Federated Learning is its emphasis on privacy. By design, raw data remains confined to local clients, significantly reducing the risk of sensitive information exposure. Model updates, which are shared, are carefully engineered to protect data privacy. In practice, techniques such as differential privacy or secure aggregation further bolster data security.

C. Explainable Hierarchical Imitation Learning

1) Introduction to EHIL: Explainable Hierarchical Imitation Learning (EHIL) [77] is an approach that sheds light on how intelligent agents make decisions. EHIL's strength lies in its ability to simplify complex behaviours. In this study, EHIL aligns with our focus on care robotics, where transparency in robot behaviours is crucial. By breaking tasks into manageable steps, EHIL supports caregiving efficiency and collaboration. As we combine EHIL with Federated Learning to address privacy, we aim to advance AI in care robotics while preserving its human-centric essence.

2) EHIL Composition and Hierarchical Structure: The EHIL model is designed to combine explainability with imitation learning in a hierarchical framework. It consists of three hierarchies, each with a specific role, working in synergy to enable both transparent decision-making and accurate task imitation.

Hierarchical Model Architecture: While the EHIL framework is flexible and does not require a specific model architecture, this project uses the same example as is used in the EHIL paper. Because of this, the architecture of this example along with the specifics of implementing a solution to robotic pouring are used as aids in articulating the EHIL framework. With this in mind, the EHIL model leverages deep residual neural networks to expedite training and enhance performance. This architecture consists of three hierarchies denoted as Hi in algorithm 1 (Denoting the FEHIL training process), with shared image feature extraction modules (figure 3). The process begins with input images, which pass through convolutional and max pooling layers, multiple residual blocks, and an average pooling layer. This yields a 1-D feature vector for subsequent analysis.

1) **Hierarchy 1: Task Phase Determination:** The first hierarchy focuses on categorizing tasks into a few distinct phases (in the case of pouring these are; Approaching, Pouring, Slowing down, and Leaving). These phases are crucial for understanding the robot's action behavior. Model H1 is trained using labeled image data from a demonstration database to predict these phases accurately.
2) **Hierarchy 2: Mid-Action State Determination:** The second hierarchy estimates more nuanced information about the specific action currently being undertaken. This state determination is complex and necessitates more fully-connected layers than Hierarchy 1. Model H2 is trained with the assistance of parameters from H1 to expedite learning and improve classification accuracy.
3) **Hierarchy 3: Action Determination:** The third hierarchy bridges the gap between visual observations and the robot's actions. It maps image arrays to the robot's end-effector angular velocity, enabling precise behavior replication. H3 is constructed with multiple fully-connected layers and inherits a shared sub-model from H2.

Logical Graph Construction: A key feature of EHIL is the creation of a logical graph, which reveals the sequence of logical states and transitions inherent in a task. This graph aids in interpreting the robot's behavior, tracing errors, and enhancing user trust. It demonstrates the logical progression from beginning to finishing a given task.

Algorithm 1
Federated Explainable Hierarchical Imitation Learning

Input:
- Dataset of demonstration images
- Hierarchy architectures (a_1, a_2, a_3)
for $i \leftarrow 1$ **to** 3 **do**
  Initialize model H_i with architecture a_i.
  Initialize clients with H_i and individual datasets.
  if $i \neq 1$ **then**
    Inherit weights from the previous hierarchy H_{i-1}
  end if
  training $\leftarrow$ true
  while training **do**
    Distribute H_i to clients.
    Initialize an empty list H_i_weights
    for each client **do**
      Compute the loss of H_i on local_data
      Train for one epoch
      Append weights to H_i_weights
    end for
    Aggregate H_i_weights:
    $H_i \leftarrow$ average(H_i_weights)
    if stopping_condition_met **then**
      training $\leftarrow$ false
    end if
  end while
end for
Output: Trained federated models for each hierarchy (H_1, H_2, H_3)

D. Combining EHIL with FL

This section outlines the methodology for integrating EHIL with FL to create FEHIL. Building upon insights from the literature review, which established the applicability of FL to supervised learning tasks with similarities to the individual hierarchies of EHIL, the integration of these two frameworks unfolds through a piecemeal, sequential approach. Algorithm 1 outlines the high level concepts acting in the proposed method.

Establishing Benchmark Results: Using the code and data employed in [77], the three hierarchies were trained to establish benchmark results. This facilitated direct comparison between the EHIL model before and after FL was applied.

Sequential Hierarchical Training: The integration began by recognizing that FL could be applied to the hierarchies of EHIL in a sequential manner. This necessitated training the hierarchies one after the other, allowing subsequent hierarchies to benefit from the knowledge encoded in the weights of preceding ones, mimicking the behavior seen in [77] to facilitate direct comparison. It is worth noting that this arrangement may not be possible to replicate in real-world scenarios due to the parallel usage of the hierarchies.

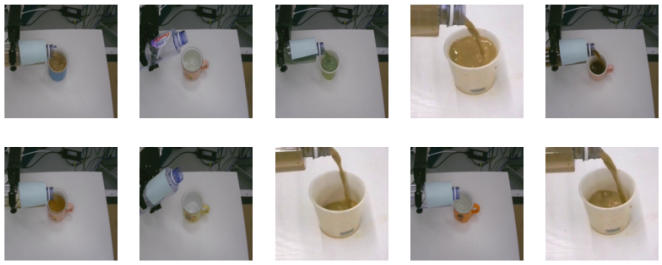

Fig. 1: Dataset examples

Data Partitioning and Distribution: Data had to be partitioned into separate clients, ensuring that there was no overlap between client datasets. The distribution of data across these clients was crucial and required careful examination to ensure representative learning.

Simulated FL Environment: Given the project's scope, the FL process was simulated and executed on a single machine. Communication frameworks and advanced security measures lay beyond the project's purview.

Client Setup and Architecture: Simulated clients were established, each equipped with a copy of the hierarchical model architecture undergoing training.

Local Training Protocol: A tailored protocol was designed for training locally on each simulated client. This protocol aligned with the overarching EHIL framework while incorporating FL mechanisms.

Aggregation Protocol: A protocol for aggregating model weights from individual client models was established, integrating insights from FL techniques.

Performance Metric Evaluation: Performance metrics were recorded on local datasets, along with evaluation on a global test set. This comprehensive assessment enabled monitoring of training on each client for signs of catastrophic forgetting. The metrics used for evaluation were specific to the client, with accuracy used for H1 and H2, and Mean Squared Error for H3.

Scope and Considerations: Given the project's specific goals, certain aspects of FL implementation were excluded from consideration. These included communication efficiency, device heterogeneity, additional privacy preservation techniques, data drift, concept drift, scaling, system architecture, and user engagement over time. These aspects were acknowledged as beyond the scope of the present study.

In the pursuit of optimizing the FEHIL model's performance, it is essential to acknowledge that hyperparameter tuning remains a promising avenue for enhancement. While the current implementation serves as a foundational exploration of the FEHIL framework's potential, future research endeavors could benefit from a comprehensive investigation into hyperparameter configurations. This includes fine-tuning learning rates, batch sizes, regularization techniques, and client-specific

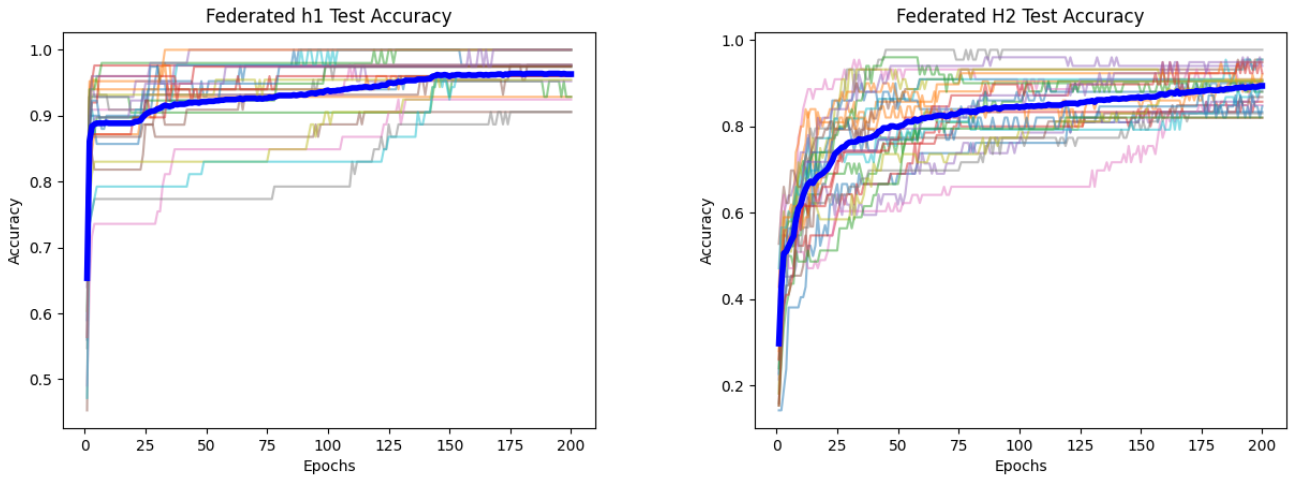

Fig. 2: The federated training of H1 and H2 using 25 clients over 200 training rounds measured by accuracy. Each of the faint lines represents the performance of the global model on each clients' test sets after each round, while the bold blue line represents the performance of the global model on the global test set.

epoch tuning among others. These adjustments, if judiciously applied, may lead to improved convergence rates and overall model efficacy.

Furthermore, the choice of model architecture is pivotal in determining the FEHIL model's capabilities. While this study leverages a ResNet/CNN-based architecture, it is noteworthy that the field of AI is marked by continual advancements. As such, exploring alternative model architectures, such as transformer-based models, could hold promise for further elevating the FEHIL model's performance. The utilization of state-of-the-art architectures, informed by ongoing developments in the AI community, may unlock new avenues for enhancing both predictive accuracy and explainability within the FEHIL framework.

IV. Experimental Setup

A. Datasets

The experimental design of this study leveraged a dataset sourced from the EHIL experiments, specifically focusing on a robotic drink pouring task. This dataset, employed as the foundation for our machine learning investigations, was collected using a consistent setup.

The task itself involved the precise execution of a robotic drink pouring process. This included a UR-16E robotic arm, an RGB camera for visual data capture, a gripper-equipped source container, and a designated target container.

Data collection for this study involved the teleoperated execution of drink pouring demonstrations. A human-controlled master manipulator was employed to guide the actions of the robotic arm, enabling the collection of diverse instances of the drink pouring task. The dataset amalgamates various sensory inputs, predominantly RGB images obtained from the camera (as seen in figure 1). These images were initially recorded at a resolution of 640x480 pixels and subsequently resized to dimensions of 50x50 pixels for standardised processing. Additionally, the dataset encompasses gripper velocities, extracted through the utilisation of arm kinematics. These velocities are articulated as 6-dimensional vectors, encompassing both angular and linear components. However, for the purpose of the experiments conducted here, a simplified representation focusing solely on the angle of pour was used, this was presented as a single continuous output.

The pouring action is primarily modulated by manipulating the wrist joint's rolling angle, which consequently tilts the source container and regulates the pouring process. To facilitate accurate data synchronisation, the dataset was collected at a frequency of 25 Hz. This synchronisation facilitated the alignment of image and action data, enabling the creation of precisely matched image-action pairs. Through this process, ground-truth trajectories were established for subsequent analysis.

Importantly, the dataset is segmented into 25 distinct subsets, each representing a unique client within a federated learning framework. This segmentation reflects the federated adaptation of the EHIL experiments. Each client's dataset encapsulates the intricacies of the drink pouring task within their specific context, contributing to the dataset's overall diversity. Across these 25 clients, the dataset sizes per-client exhibit variability, spanning from 195 frames to 265 frames. These client-specific datasets were then individually split into train/test sets at an 80/20 rate. The amalgamation of these individual test sets formed the measurement for the performance of the global model.

In total, the dataset comprises 5690 frames, encompassing a comprehensive array of scenarios within the drink pouring task. This dataset, with its inherent variability and granularity,

forms the bedrock of our machine learning experimentation, enabling us to delve into the realm of explainable hierarchical imitation learning with a robust and versatile foundation.

B. Experimental Design

The experimental design for this study encompassed several phases, each contributing to a comprehensive evaluation of the FEHIL framework. The design was carefully structured to ensure accurate benchmark establishment, incremental expansion of scope, and effective convergence assessment.

To initiate the experimentation process, a preliminary step involved an in-depth comprehension of the codebase associated with the EHIL paper. This familiarization process was essential to grasp the foundational concepts and technical intricacies of the EHIL framework.

Following this initial phase, the experiment design proceeded to establish benchmark results that would serve as reference points for future evaluations. Instead of directly adopting results from the EHIL paper, a pragmatic approach was adopted. EHIL experiments were conducted locally, enabling the verification of code functionality and data interaction. These preliminary experiments were executed with a cap of 200 epochs and a suitable early-stopping criteria, providing a foundational benchmark for model performance.

Across all experiments, whether EHIL or within the federated learning paradigm, a consistent batch size of 20 was employed. In the context of federated learning, each client underwent one epoch of training per training round.

Building on the EHIL results, the experiment then transitioned into the realm of federated learning. Initial federated learning trials were conducted using a subset of the data, comprising 10 clients, spanning 200 rounds, facilitating a preliminary assessment of the federated approach's efficacy. Encouragingly, these initial trials demonstrated promising outcomes, with comparable results observed across all three hierarchies.

Buoyed by the initial positive outcomes, the scope of the experiment was broadened to encompass the complete dataset, featuring the full complement of 25 clients. However, it was observed that the third hierarchy failed to converge satisfactorily within the anticipated maximum training rounds. To address this, a strategic extension of training was undertaken. The third hierarchy's training duration was prolonged to a maximum of 1000 epochs, with the incorporation of an early stopping mechanism. This mechanism dictated that if the mean squared error (MSE) failed to improve by more than 0.001 for ten consecutive epochs, the training loop would be prematurely terminated (matching the early stopping mechanism for the other two hierarchies, only represented as a minimum improvement for accuracy). Ultimately, this extended training regimen enabled the third hierarchy to converge effectively after 402 training rounds.

C. Model Hierarchy Architectures

This section describes the specific parameters pertaining to the architecture depictions seen in figures 3 and 4. Figure 3 (left) depicts a residual block, while figure 3 (right) uses the residual block as a building block for a residual neural network architecture. Figure 4 takes the shared architecture from figure 3 and uses this as the building block from which to illustrate the unique individual properties of the three hierarchies. Through the use of these figures and the descriptions given in this section, the architectures of the three hierarchies are clarified.

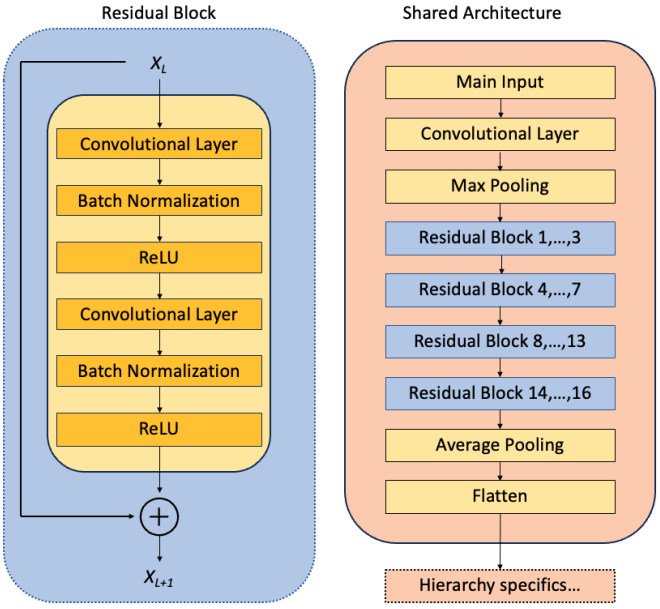

Fig. 3: The residual block (Left) and shared architecture (Right) components of the FEHIL model. The shared architecture component is also used for model initialisation from the learned weights of the previous hierarchy. Architectural design originated from EHIL [77].

1) Hierarchy 1: Hierarchy 1 takes input images of size 50x50 pixels with 3 color channels. The model architecture consists of convolutional layers interspersed with layer normalization and activation functions. The model comprises a total of 11,245,700 trainable parameters.

The architecture begins with an input layer which receives the input images. This is followed by a convolutional layer with 64 filters of size 3x3 and a stride of 2, resulting in an output shape of 25x25x64. A layer normalization layer is applied to the convolutional outputs, followed by an activation function.

A max-pooling layer with a pool size of 2x2 and stride of 2 reduces the spatial dimensions of the data, resulting in an output shape of 13x13x64. This is followed by another convolutional layer with 64 filters of size 3x3. Similar to before, layer normalization and activation are applied.

The subsequent layers follow a similar pattern: pairs of convolutional layers with intermediate layer normalization and activation functions. These layers are followed by a residual connection using an element-wise addition operation that combines the output of the preceding layer and the output

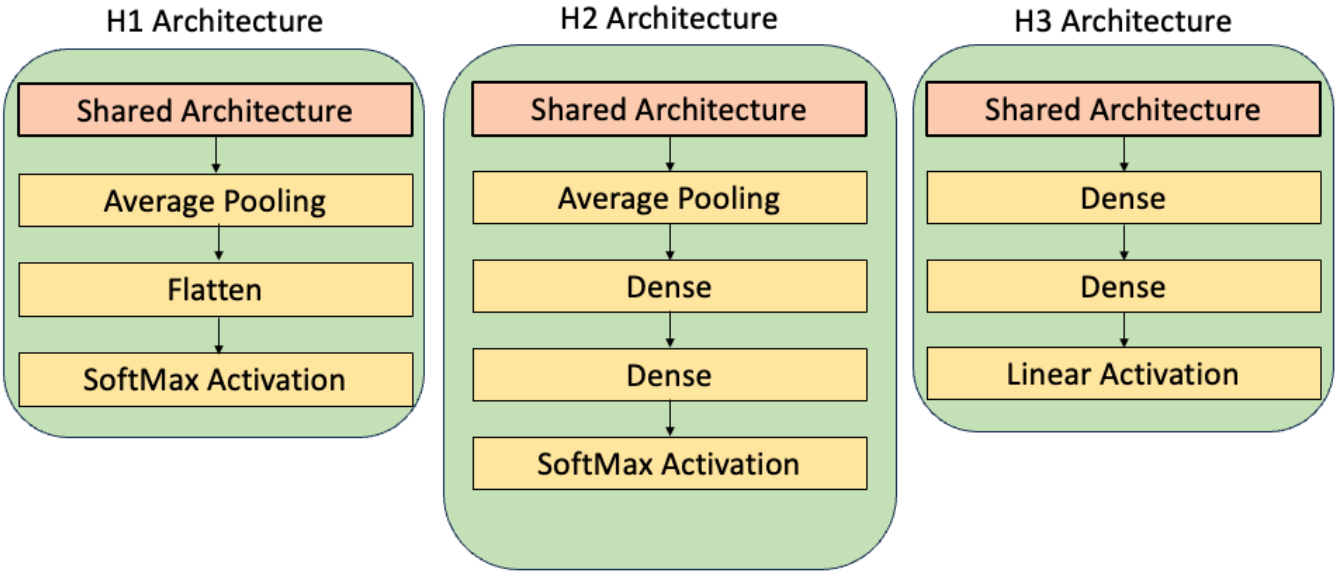

Fig. 4: The unique characteristics of the model architectures pertaining to the three hierarchies in the FEHIL model. The Shared Architecture relates to the architecture described in Figure 3.

of the earlier max-pooling layer. This is illustrated by the Residual Block in figure 3 (Left) and the first block in figure 3 (Right). Subsequent blocks replace inputs from the max-pooling layer with the outputs of the previous residual block.

The architecture continues to stack residual blocks, each time doubling the number of filters in the convolutional layers. Specifically, the number of filters is increased to 128, then to 256, and finally to 512. The spatial dimensions of the data are reduced through max-pooling or strided convolutions after each doubling of filter count.

After the final doubling of filters to 512, an average pooling layer is applied with a pool size that reduces the spatial dimensions to 1x1x512, followed by flattening the data to a vector of size 512. This is represented in figure 3 (Right) where the residual blocks end. The output of this is then fed through an average pooling layer, a fully connected layer with 128 units, followed by another fully connected Dense layer with SoftMax activation and 4 units, which corresponds to the number of classes in the phases of robotic pouring. This is denoted in figure 4 (Left).

2) Hierarchy 2: Hierarchy 2 follows a self-initialization strategy by inheriting the weights from Hierarchy 1, up to the point where the shared hierarchy ends (as denoted in figure 3). The initial layers of Hierarchy 2 are therefore identical to those of Hierarchy 1 up to the end of the shared hierarchy. Additionally, the architecture of Hierarchy 2 introduces new layers for further processing (figure 4 Middle).

Following the Hierarchy 1 inheritance, a global average pooling layer is applied, which takes the output from the Hierarchy 1 section and performs average pooling across the spatial dimensions. This results in a tensor of shape (None, 512), effectively collapsing the spatial information while retaining the depth.

Subsequently, two fully connected layers are added to Hierarchy 2. The first dense layer has 128 units, and the second dense layer contains 64 units. These layers contribute to the feature extraction and dimensionality reduction process.

The final layer in Hierarchy 2 is a dense layer with SoftMax activation, which consists of 10 units. This layer is used for classification purposes, where each unit corresponds to a different class label. The weights in this section are initialized from scratch. The total number of trainable parameters in Hierarchy 2 is 11,254,090, and this model is designed for classification tasks involving 10 classes, representing the state transitions of a logical graph.

3) Hierarchy 3: Hierarchy 3 is constructed based on the weights inherited from Hierarchy 2. The architecture of Hierarchy 3 further extends the capabilities of the model and introduces new layers for specific tasks (figure 4 Right).

Following the Hierarchy 2 inheritance, Hierarchy 3 incorporates additional layers for specific tasks. A dense layer with 128 units is added, contributing to feature extraction and representation learning. Subsequently, another dense layer with 64 units is introduced, continuing the process of dimensionality reduction and feature refinement.

The final layer in Hierarchy 3 is a dense layer with regression activation consisting of a single unit. This layer is dedicated to regression tasks, where the goal is to predict a continuous numerical value. The weights of this section are initialized independently for the new task. The total number of trainable parameters in Hierarchy 3 is 11,263,819, and the model is designed for predicting the angle of pour for the robotic arm.

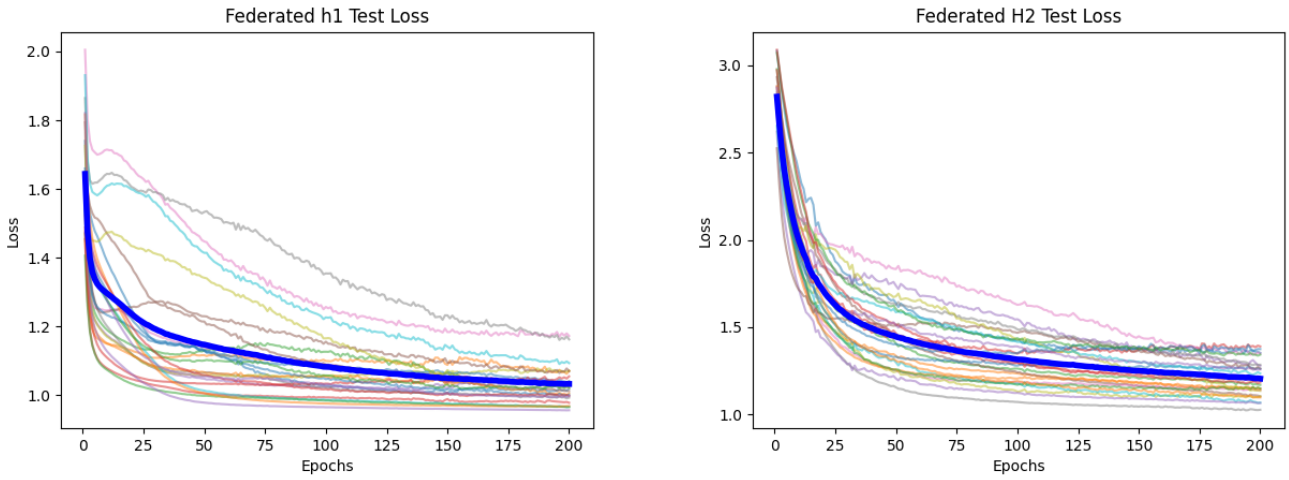

Fig. 5: The federated training of H1 and H2 using 25 clients over 200 training rounds measured by categorical cross entropy. Each of the faint lines represents the performance of the global model on each clients' test sets after each round, while the bold blue line represents the performance of the global model on the global test set.

D. Implementation Details

This section provides a factual overview of the technical aspects involved in combining FL with EHIL. The implementation specifics are outlined, including the programming tools, frameworks, and environments employed to ensure the reproducibility of our experiments.

1) Programming Languages, Libraries, and Frameworks: The original EHIL codebase was developed using the Keras library as the foundational framework. For creating a controlled training environment, Windows Subsystem for Linux (WSL) along with Conda was utilized to conduct the benchmark EHIL experiments. Given the integration goals of EHIL with FL, TensorFlow solutions were explored. TensorFlow Federated (TFF) was chosen to simulate FL due to its alignment with the Google coding ecosystem and compatibility with Keras. Simulating FL enabled faster development compared with the manual distribution of training across different machines. In this way, the data could still be separated and trained privately, resulting in the same behaviours as far as model training is concerned. Testing across different platforms, including M1 Mac, Windows, and WSL, proved challenging due to the fragility of the TFF package and dependency resolution issues. Testing was ultimately accomplished using Google Colab notebooks designed for TFF, resolving dependency challenges.

2) Data Preprocessing and Integration with TFF: Adapting data for TensorFlow Federated (TFF) involved a multi-step process. Initially, raw image data underwent resizing and organization into a structured numpy array. This transformed data was then encapsulated within a MapDataset, aligned with TFF's client-centric approach. Leveraging TFF's functions, tff.simulation.datasets.ClientData.from_clients_and_tf_fn() converted the numpy array into ClientData objects. Subsequently, federated_data.create_tf_dataset_for_client() utilized these objects to generate tailored datasets for individual clients.

3) TFF Compilation and Strong Typing: TFF offers a runtime environment for FL scenarios. TFF's language structure uses decorators like @tff.tf_computation and @tff.federated_computation to designate functions within its typed environment. This streamlines computations, optimizing training speed in simulated FL setups. The @tff.tf_computation decorator focuses on TensorFlow's computational graph, while @tff.federated_computation enhances distributed execution across nodes, aligning with FL's principles. By utilizing these decorators, TFF's runtime environment maximizes parallelism and TensorFlow optimization, pivotal for efficient training in Federated Learning simulations [64].

4) Hierarchical Model Training: TFF's stringent typing framework necessitated distinct training of hierarchy levels. Architectural disparities prevented unified FL code reuse across all three levels. Consequently, hierarchical training occurred separately for each level. After training, dedicated model hierarchy files were preserved. For inter-hierarchy weight transfer, file operations facilitated seamless data exchange.

5) Metrics Tracking and Model Persistence: Throughout the training process, metrics were captured using summary writers and csv files. Post-experiment analysis was performed by visualizing these metrics using TensorBoard. After each experiment, the trained models were saved to files for reproducibility and further analysis. Csv files were used for the later combination of EHIL and FEHIL results into the same graph for direct comparison.

V. Results and Analysis

In the following section, a breakdown of the attained insights and comparative evaluations will be presented. The quantitative analysis, as detailed in Section V-A, sheds light

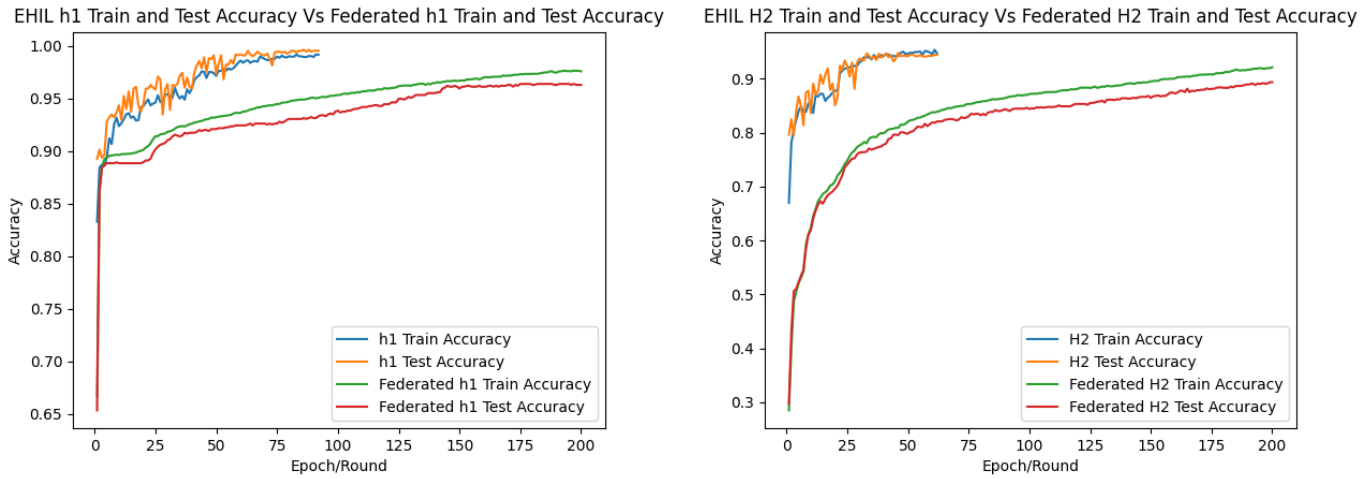

Fig. 6: Comparing the accuracy of the EHIL with FEHIL for H1 and H2. In both cases, EHIL reached its stopping criteria before the end of its training budget, while the federated approach ran for the full 200 rounds without reaching this criteria.

on the dynamics of the FL approach, capturing the intricacies of FEHIL's performance compared to the EHIL benchmark. Through this, convergences and divergences in training trajectories and outcomes are revealed. Subsequently, the focus pivots to a comparative analysis, delving into the deeper implications of FL's mechanisms and its distinctive trajectory when matched against the established benchmark.

A. *Quantitative analysis*

Overall, the FEHIL approach was able to achieve comparable results to EHIL, but required far more training. This can be observed in figures 6 and 9. These results inconclusively suggest that FL requires more training than the traditional approach where the model has access to all the data. This is inconclusive because there has been insufficient opportunity to experiment with hyperparameters within the limits of this project.

We notice similarities in the final results for accuracy and MSE through the hierarchies where these metrics are used in tables I, II and III. While these results took much longer to obtain, this demonstrates that FEHIL can be used to achieve comparative results to the EHIL framework trained in the traditional manner. With this in mind, it is important to note that the results do not equal the performance of the benchmark and in the case of H2 perform with a 5% accuracy reduction.

In figure 7 we can observe that the categorical cross entropy of the H1 and H2 hierarchies performs much worse than the benchmark throughout training. This means that the measured difference between the predicted probability distribution and the true distribution was highly dissimilar while still making the correct prediction in most scenarios (accuracy). This could be due to the diversity in local datasets causing the optimisation landscape to be more rugged.

With each round representing an instance where the local client would upload their fine-tuned model to the central server, the performance of the newly aggregated model on their local dataset demonstrates the presence or absence of catastrophic forgetting at the beginning of each round. We can observe this where the accuracy reduces in the various clients through training in Figure 2. In this case, the results seem encouraging in that consistent catastrophic forgetting cannot be observed through training, and clients could expect to receive a global model that performs comparatively on their local dataset. Analogously, the loss shows comparable trends in the training of H2, but the early training of H1 shows some consistent lack of improvement in the earlier rounds of training. This could be due to differences in the training data where the features learned by other client models are incompatible with the features demonstrated in these clients' datasets.

B. *Comparative Analysis*

Through the previous section (V-A), it was observed that while FEHIL eventually achieved similar results to EHIL, these took much longer to obtain. A possible reason for this is the inherent inaccuracy in FL and the indirect learning that is achieved through model weights averaging. Especially in complex model architectures such as ResNet, the individual parameters can expect to find different gradients as they optimise a different optimisation landscape. While averaging these weights is seen to gain some knowledge and improve the performance of the global model iteratively, the optimisation landscape isn't learned in the same way that a model with access to all the data might. Through this, a messier approximation of the desired output is achieved that functions as a suitable agent in this scenario. This messy approximation, while achieving high accuracy, can result in a high loss as the merging of model weights introduces noise into the probability distribution generated by the model in order to make a prediction. It seems that through the training of H1

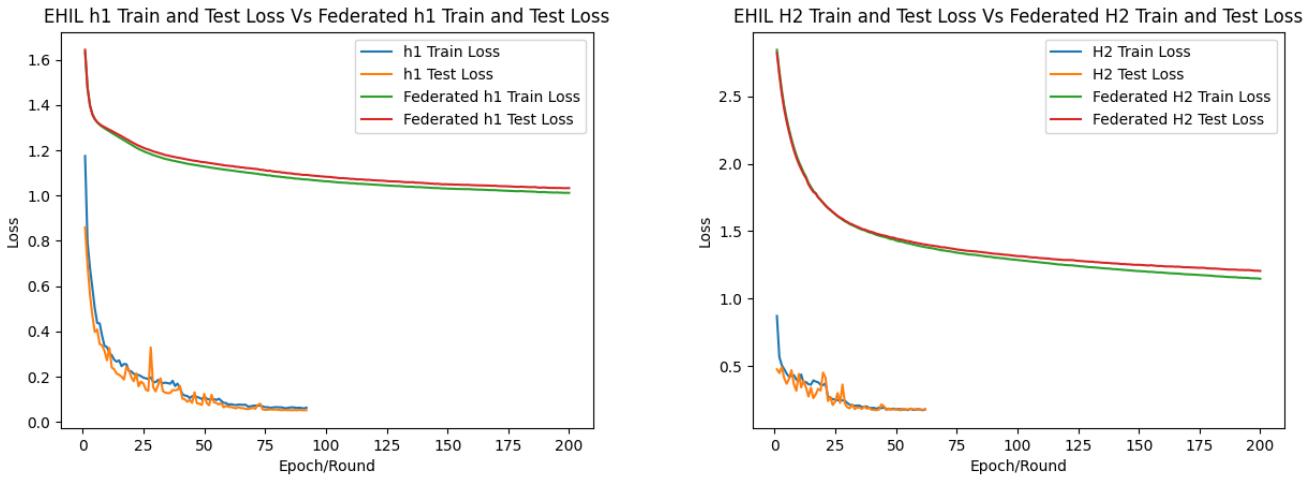

Fig. 7: Comparing the categorical cross entropy of EHIL with FEHIL for H1 and H2. In both cases, the EHIL session reached its stopping criteria before the end of its training budget, while the federated approach ran for the full 200 rounds without reaching this criteria.

and H2, we can observe that this effect is persistent throughout the separate classification scenarios. Interestingly, H2 sees this translate to a more significant reduction in classification accuracy than H1. This could be a symptom of the greater number of available classes in this hierarchy and the noisy predictions directly disrupting the classification where each possible class in the distribution has a greater opportunity to be affected by noise.

Compared to the classifier hierarchies, the federated third hierarchy shows a much slower convergence at nearly 10 times the number of training rounds/epochs required for EHIL. This result is demonstrative of the differences between classification and regression in the benchmark case and again when FL is applied. In EHIL experiments, the regression model converges faster than the classifiers at epoch 54 compared to 91 and 61 for H1 and H2 respectively. While it is not clear why this convergence is faster, it can be observed that the opposite occurs in the federated case. The curvature of the training curve seen in figure 6 clearly shows a gradual leveling off at the cutoff 200 rounds limit, while the same location in figure 9 denotes a still-falling loss value that will not reach it's stopping criteria for another 200 training rounds. From this, we can see that the nature of averaging the model weights varies by the specific optimisation problem the model is presented with.

VI. Discussion

Data Distribution: As shown in Figure 1, the synthetic data used in this project is unlikely to reflect real-world settings, and so cannot verify the robustness of FL to new, user-generated data. While we can observe some differences in the data, the composite parts of the EHIL framework are generalisable and would require considerably more examples to verify performance across the settings real world implementations of the technique would expect.

In order to demonstrate robustness to new scenarios, a better dataset would have to be gathered where the client data deliberately varies in terms of the visual features that might be used to conclude the same actions from the EHIL model. With such data, experiments could be run to analyse the impact this has on local and global model performance. From here, experiments into mitigation techniques could be better explored and the effects of more personalised FL techniques might be more clearly demonstrated.

Catastrophic Forgetting: A key barrier to the continued participation in FL training by participating parties is the issue if catastrophic forgetting, whereby the next round of the global model performs worse for a particular client than the model that client recently fine-tuned (as seen in the up-ticks in loss through training rounds most prominently in figure 5 Left). In this way, participants in the training scheme would be discouraged from contributing model updates and the value of the user data would be lost. Mitigation techniques such as lifelong learning are appropriate to explore here, with a variety of available techniques designed to prevent such events. While this is a developing field and highly implementation dependant, even simple measures such as testing the global model before reassigning weights at the beginning of an FL round would enable clients the option to reject the new global model. Furthermore, conducting this kind of analysis can enable system developers to spot outliers and begin to address data distribution concerns without actually accessing the data.

The EHIL Approach: To make the best use of user data in distributed model training, the data would come from the environment in which the care-robot is actually used. This would be difficult using EHIL, because of the need to label data manually by a human expert. While EHIL represents a positive step in developing XAI, other techniques may be better placed to address the motivations of this project mentioned in section

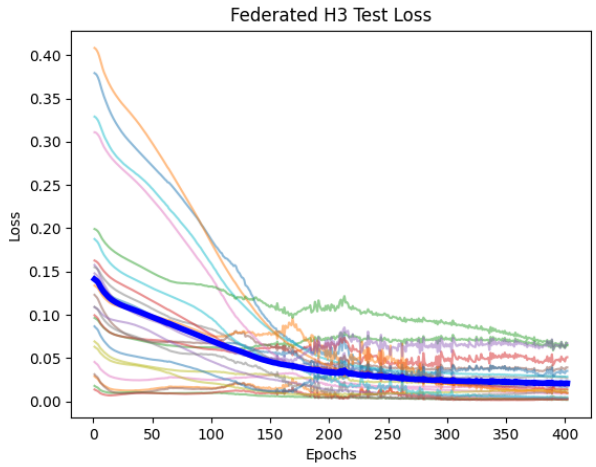

Fig. 8: The loss of all clients through the training of hierarchy 3. Each line represents the loss of the global model on the test set of each client, with the bold blue line denoting the loss of this model on the global test set.

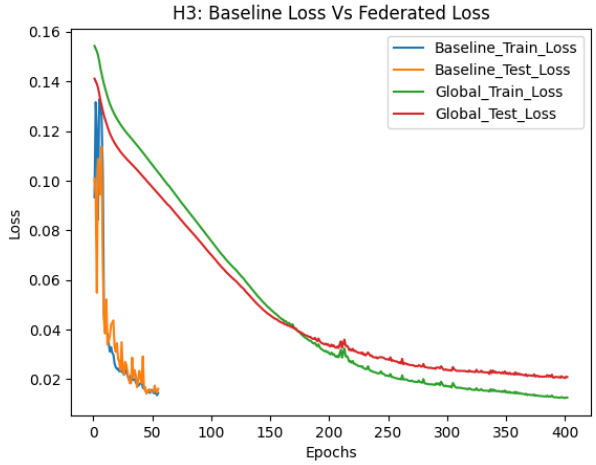

Fig. 9: Comparing the loss of the EHIL appraoch with the FEHIL approach through their respective epochs/rounds. Both approaches eventually met their stopping criteria, which is a minimum MSE improvement of 0.001 over 10 rounds/epochs.

I. Discussed further in section VII, such techniques would require some method of circumventing or automating the labelling process. Labelling, being both resource intensive and human-expert dependant, confines the potential environments in which data can be gathered to those where these human experts are available. In terms of implementation, the ability to provide patients with the power to demonstrate tasks to their robotic carers would dramatically increase the power of these AI-governance systems while reducing the cost of system development.

Hyperparameter Tuning: With more time, hyperparameter tuning would enable increased model performance and provide greater insight into the efficacy of FEHIL. Specifically, the number of training epochs conducted by each client per round of federated learning would have been of significant benefit to tune. The typical implementation of FL, where the model weights are naively aggregated through averaging, serves the purposes of understanding the core viability of FL in its ability to conduct distributed model optimisation in the scenarios investigated here. However, it is unclear whether these results reflect the true potential of FL. From the results observed here, it can be said that FL will hinder the speed of development of such systems which may cause entities to consider more efficient, less privacy preserving measures. If FL is found to be a genuine hindrance, this is of course a legitimate decision, but the results here should not be used as direct evidence with which to make such a decision. To better understand the potential impact of FL, it would be beneficial to run the same experiments with variable training epochs per training round of FL. Through this, the effects of more fine-tuning could be observed from multiple perspectives. Among these are; accuracy, loss, MSE, catastrophic forgetting and convergence rates.

Towards the adoption of privacy-preserving AI-governed care robots: Given the limitations of this project, it is important to note that we have demonstrated the ability of FL to achieve comparable results to the non-distributed setting when applied to hierarchical behavioural cloning. This result, however long it took to obtain through training, demonstrates that FL is a suitable training framework for AI-governed care robotics to make use of user-data once deployed. Furthermore, it should be considered by any party implementing this technology that the training cost is distributed amongst the clients and is therefore essentially free, with the cost coming down to just the communication and aggregation of model weights. As described in section I, the key legal grey-areas which must be navigated for robotics AI in healthcare are explainability and privacy. Through the demonstrated efficacy of FEHIL, we have shown a potential solution exists that allows entities to continue innovating and developing in the field.

VII. Future Work

A. Varying Number of Epochs Per Training Round

To continue the research conducted in this project, further experiments would be conducted with more epochs of training conducted by each client per training round. These would be structured as:

- **Low Epochs:** Since 1 epoch per round was conducted in this project, this would be bolstered with a 2 epoch experiment.
- **Moderate Epochs:** To attempt a balance between client fine-tuning and model aggregation, experiments with 5 and 10 epochs per training round would be conducted.
- **Excessive Epochs:** To ensure that the boundaries of performance are reached, the clients would train for 30 epochs per training round in one experiment, and then to full convergence in another.

Through this variation, the correct balance of epochs to training rounds can be found and the viability of FEHIL can be discovered through observing the effect on convergence rates.

B. More Data

To better explore the impact of applying FL in real-world scenarios, the real-world scenario must be simulated as closely as possible. Considering specifically the EHIL approach, this would require more unique containers, variable environments, variable liquids and unique human-labelers per client. Through this, the effects of outliers can be observed and the ability of FL to overcome data distribution challenges through tuning hyperparameters and applying subsequent techniques can be investigated.

C. Weighted Averaging

Should future experiments feature a larger dataset with more variability in available data per client, the model weight aggregation in each training round will need to be weighted in order to mitigate the adverse effects of the aggregation such as catastrophic forgetting.

D. Lifelong Learning

In addressing the challenge of catastrophic forgetting in FL, it is essential to ensure that the global model's performance does not hinder individual client performance. Building upon insights from personalized and lifelong FL techniques, as discussed in [74] and [45], future work should optimize the training process to eliminate any potential disincentives for individual clients. Techniques such as sense-checking the global model before adopting its weights would serve to maintain the performance of the client in any given training round. This could be balanced against the traditional approach with a minimal number of clients to observe the effects of lifelong learning techniques on overall model performance and convergence.

E. Alternative Learning Techniques

While the EHIL approach serves as an excellent example for imbuing a system with explainability from its design, it is not necessarily the best solution for harnessing user data after model deployment. As explained in section II-A, EHIL requires the demonstrator to also label the data. Because of this, the EHIL approach would not be suitable on its own to be used for robotics systems in the home (a motivation described in section I). To address this, research must be conducted into alternative learning and explainability techniques that do not depend on expert labelling. Potential solutions involve autolabeling, inverse reinforcement learning and video-based imitation learning with self-supervised learning, but it should be noted that this is an open problem.

F. Simulation and Sim-2-Real

While the results achieved in these experiments are encouraging, they are representative only of metrics with correlation to success when deployed in a robotic system. In order to reveal the true potential of this technique, the trained models must be combined and put together in a robotic simulation system. By simulating the process in a physics simulator, it is possible to evaluate whether the techniques are suitable to be deployed into a real-life robotics system. Without these steps, the research conducted here can only be used as encouraging evidence.

Metric	EHIL	FEHIL
Accuracy	0.996	0.963
Loss	0.052	1.033

TABLE I: Results for H1

Metric	EHIL	FEHIL
Accuracy	0.945	0.894
Loss	0.177	1.206

TABLE II: Results for H2

Metric	EHIL	FEHIL
MSE	0.016	0.021

TABLE III: Results for H3

VIII. Conclusion

The experiments conducted here have demonstrated that FL can be successfully applied to hierarchical learning using the EHIL approach, with robotic pouring as an example. The key findings from these experiments are that while FEHIL can achieve similar results to EHIL, the federated approach takes far longer to converge, which could be seen as a barrier to adoption.

A key limitation of this project has been the lack of consideration for the number of training epochs for each client per round of federated learning. Without rerunning the experiments in this project while tuning this parameter, it cannot be conclusively stated whether the aforementioned convergence speed limitations are inherent in the approach or addressable.

Overall, the project represents a positive step towards the adoption of AI-governed robotics in healthcare by demonstrating that it is possible to achieve state-of-the-art performance while preserving data privacy and explainability. Preserving these qualities directly addresses concerns raised by academics in the field as well as regulatory authorities, it is hoped that this work can be continued in the ways outlined in section VII to further improve perceptions of AI in the fields of robotics and healthcare.

References

[1] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. "Interpretable machine learning in healthcare". In: *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*. 2018, pp. 559–560.

[2] Brenna D. Argall et al. “A survey of robot learning from demonstration”. In: *Rob. Auton. Syst.* 57.5 (May 2009), pp. 469–483. ISSN: 0921-8890. DOI: 10.1016/j.robot.2008.10.024.

[3] *ARTICLE29 - Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 (wp251rev.01)*. [Online; accessed 21. Aug. 2023]. Aug. 2023. URL: https://ec.europa.eu/newsroom/article29/items/612053.

[4] Louis JM Aslett, Pedro M Esperança, and Chris C Holmes. “A review of homomorphic encryption and software tools for encrypted statistical machine learning”. In: *arXiv preprint arXiv:1508.06574* (2015).

[5] Ho Bae et al. “Security and privacy issues in deep learning”. In: *arXiv preprint arXiv:1807.11655* (2018).

[6] Xiao Bai et al. “Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments”. In: *Pattern Recognition* 120 (2021), p. 108102.

[7] BBC News. “NHS data sale ’an invasion of privacy’, campaigners say”. In: *BBC News* (June 2021). URL: https://www.bbc.co.uk/news/uk-england-somerset-57568711.

[8] Keith Bonawitz et al. “Practical Secure Aggregation for Privacy-Preserving Machine Learning”. In: *CCS ’17: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. New York, NY, USA: Association for Computing Machinery, Oct. 2017, pp. 1175–1191. ISBN: 978-1-45034946-8. DOI: 10.1145/3133956.3133982.

[9] Nadia Burkart and Marco F Huber. “A survey on the explainability of supervised machine learning”. In: *Journal of Artificial Intelligence Research* 70 (2021), pp. 245–317.

[10] Lee A. Bygrave. “Article 22 Automated individual decision-making, including profiling”. In: *OUP Academic* (Feb. 2020). DOI: 10.1093/oso/9780198826491.003.0055.

[11] David Byrd and Antigoni Polychroniadou. “Differentially Private Secure Multi-Party Computation for Federated Learning in Financial Applications”. In: *arXiv* (Oct. 2020). DOI: 10.48550/arXiv.2010.05867. eprint: 2010.05867.

[12] Leo de Castro et al. “Does fully homomorphic encryption need compute acceleration?” In: *arXiv preprint arXiv:2112.06396* (2021).

[13] Erika Check Hayden. “Privacy loophole found in genetic databases - Nature”. In: *Nature* (Jan. 2013). ISSN: 1476-4687. DOI: 10.1038/nature.2013.12237.

[14] Jianyu Chen, Bodi Yuan, and Masayoshi Tomizuka. “Deep Imitation Learning for Autonomous Driving in Generic Urban Scenarios with Enhanced Safety”. In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 03–08. DOI: 10.1109/IROS40897.2019.8968225.

[15] Junhong Chen et al. “Supervised Semi-Autonomous Control for Surgical Robot Based on Banoian Optimization”. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 2020–24. DOI: 10.1109/IROS45743.2020.9341383.

[16] Marcy Cuttler. “Transforming health care: How artificial intelligence is reshaping the medical landscape”. In: *CBC* (Apr. 2019). URL: https://www.cbc.ca/news/health/artificial-intelligence-health-care-1.5110892.

[17] Soham De et al. “Unlocking high-accuracy differentially private image classification through scale”. In: *arXiv preprint arXiv:2204.13650* (2022).

[18] Weiping Ding et al. “Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey”. In: *Inform. Sci.* 615 (Nov. 2022), pp. 238–292. ISSN: 0020-0255. DOI: 10.1016/j.ins.2022.10.013.

[19] Zrinjka Dolic, Milieu Consulting Rosa CASTRO, Milieu Consulting Andrei MOARCAS, et al. “Robots in healthcare: a solution or a problem?” In: (2019).

[20] Friedrich Dörmann et al. “Not all noise is accounted equally: How differentially private learning benefits from large sampling rates”. In: *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE. 2021, pp. 1–6.

[21] Finale Doshi-Velez and Been Kim. “Towards a rigorous science of interpretable machine learning”. In: *arXiv preprint arXiv:1702.08608* (2017).

[22] Cynthia Dwork et al. “Our Data, Ourselves: Privacy Via Distributed Noise Generation”. In: *Advances in Cryptology - EUROCRYPT 2006*. Berlin, Germany: Springer, 2006, pp. 486–503. DOI: 10.1007/11761679_29.

[23] Yaniv Erlich et al. “Identity inference of genomic data using long-range familial searches”. In: *Science* 362.6415 (2018), pp. 690–694.

[24] Melanie Evans. “Why Big Tech Wants Access to Your Medical Records”. In: *WSJ* (Jan. 2019). URL: https://www.wsj.com/articles/hospitals-give-tech-giants-access-to-detailed-medical-records-11579516200.

[25] Minghong Fang et al. “Local model poisoning attacks to {Byzantine-Robust} federated learning”. In: *29th USENIX security symposium (USENIX Security 20)*. 2020, pp. 1605–1622.

[26] Laura Fernández-Becerra et al. “Accountability and Explainability in Robotics: A Proof of Concept for ROS 2- And Nav2-Based Mobile Robots”. In: *International Joint Conference 16th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2023) 14th International Conference on EUropean Transnational Education (ICEUTE 2023)*. Cham, Switzerland: Springer, Aug. 2023, pp. 3–13. DOI: 10.1007/978-3-031-42519-6_1.

[27] Roy Fox et al. “Multi-Task Hierarchical Imitation Learning for Home Automation”. In: *2019 IEEE 15th International Conference on Automation Science and*

Engineering (CASE). IEEE, pp. 22–26. DOI: 10.1109/COASE.2019.8843293.

[28] Xiaokuan Fu, Yong Liu, and Zhilei Wang. "Active Learning-Based Grasp for Accurate Industrial Manipulation". In: *IEEE Trans. Autom. Sci. Eng.* 16.4 (Feb. 2019), pp. 1610–1618. DOI: 10.1109/TASE.2019.2897791.

[29] Craig Gentry. "A fully homomorphic encryption scheme". PhD thesis. Stanford University, 2009.

[30] Malin Knutsen Glette, Karina Aase, and Siri Wiig. "The relationship between understaffing of nurses and patient safety in hospitals-A literature review with thematic analysis". In: (2017).

[31] O. Goldreich, S. Micali, and A. Wigderson. "How to play ANY mental game". In: *STOC '87: Proceedings of the nineteenth annual ACM symposium on Theory of computing*. New York, NY, USA: Association for Computing Machinery, Jan. 1987, pp. 218–229. ISBN: 978-089791221. DOI: 10.1145/28395.28420.

[32] Saransh Gupta, Rosario Cammarota, and Tajana Šimunić Rosing. "Memfhe: End-to-end computing with fully homomorphic encryption in memory". In: *ACM Transactions on Embedded Computing Systems* (2022).

[33] Melissa Gymrek et al. "Identifying personal genomes by surname inference". In: *Science* 339.6117 (2013), pp. 321–324.

[34] *Hospitals Selling Patient Data: Privacy Concerns Arise – excel-medical.com*. [Online; accessed 13. Aug. 2023]. Aug. 2023. URL: https://www.excel-medical.com/hospitals-selling-patient-data-privacy-concerns-arise.

[35] IFR International Federation of Robotics. *International Federation of Robotics*. [Online; accessed 11. Aug. 2023]. Aug. 2023. URL: https://ifr.org/service-robots.

[36] Bargav Jayaraman et al. "Distributed Learning without Distress: Privacy-Preserving Empirical Risk Minimization". In: *Advances in Neural Information Processing Systems* 31 (2018). URL: https://proceedings.neurips.cc/paper_files/paper/2018/hash/7221e5c8ec6b08ef6d3f9ff3ce6eb1d1-Abstract.html.

[37] Shouling Ji et al. "De-Health: all your online health information are belong to us". In: *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE. 2020, pp. 1609–1620.

[38] Peter Kairouz et al. "Advances and open problems in federated learning". In: *Foundations and Trends® in Machine Learning* 14.1–2 (2021), pp. 1–210.

[39] Latif U Khan et al. "Dispersed federated learning: Vision, taxonomy, and future directions". In: *IEEE Wireless Communications* 28.5 (2021), pp. 192–198.

[40] Helena Klause et al. "Differentially private training of residual networks with scale normalisation". In: *arXiv preprint arXiv:2203.00324* (2022).

[41] Brian Knott et al. "Crypten: Secure multi-party computation meets machine learning". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 4961–4973.

[42] Alexey Kurakin et al. "Toward training at imagenet scale with differential privacy". In: *arXiv preprint arXiv:2201.12328* (2022).

[43] In Lee. "Service robots: A systematic literature review". In: *Electronics* 10.21 (2021), p. 2658.

[44] Wenqi Li et al. "Privacy-Preserving Federated Brain Tumour Segmentation". In: *Machine Learning in Medical Imaging*. Cham, Switzerland: Springer, Oct. 2019, pp. 133–141. DOI: 10.1007/978-3-030-32692-0_16.

[45] Boyi Liu, Lujia Wang, and Ming Liu. "Lifelong federated reinforcement learning: a learning architecture for navigation in cloud robotic systems". In: *IEEE Robotics and Automation Letters* 4.4 (2019), pp. 4555–4562.

[46] Boyi Liu et al. "Federated Imitation Learning: A Novel Framework for Cloud Robotic Systems With Heterogeneous Sensor Data". In: *IEEE Rob. Autom. Lett.* 5.2 (Feb. 2020), pp. 3509–3516. ISSN: 2377-3766. DOI: 10.1109/LRA.2020.2976321.

[47] Hangxin Liu et al. "Interactive Robot Knowledge Patching Using Augmented Reality". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 21–25. DOI: 10.1109/ICRA.2018.8462837.

[48] Arne Maibaum et al. "A critique of robotics in health care". In: *AI & society* (2022), pp. 1–11.

[49] Nathalie Majcherczyk, Nishan Srishankar, and Carlo Pinciroli. "Flow-FL: Data-Driven Federated Learning for Spatio-Temporal Predictions in Multi-Robot Systems". In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2021, pp. 8836–8842. DOI: 10.1109/ICRA48506.2021.9560791.

[50] Jims Marchang et al. "Security and privacy in assistive robotics: cybersecurity challenges for healthcare". In: *EPSRC UK-RAS Network* (July 2023). URL: https://shura.shu.ac.uk/32231.

[51] Hironori Matsuzaki and Gesa Lindemann. "The autonomy-safety-paradox of service robotics in Europe and Japan: a comparative analysis". In: *AI & society* 31 (2016), pp. 501–517.

[52] Brendan McMahan et al. "Communication-Efficient Learning of Deep Networks from Decentralized Data". In: *Artificial Intelligence and Statistics*. PMLR, Apr. 2017, pp. 1273–1282. URL: http://proceedings.mlr.press/v54/mcmahan17a?ref=https://githubhelp.com.

[53] Frank McSherry and Kunal Talwar. "Mechanism Design via Differential Privacy". In: *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*. IEEE, pp. 21–23. DOI: 10.1109/FOCS.2007.66.

[54] Viraaji Mothukuri et al. "A survey on security and privacy of federated learning". In: *Future Generation Computer Systems* 115 (2021), pp. 619–640.

[55] Blake Murdoch. "Privacy and artificial intelligence: challenges for protecting health information in a new era". In: *BMC Medical Ethics* 22.1 (2021), pp. 1–5.

[56] Liangyuan Na et al. "Feasibility of reidentifying individuals in large national physical activity data sets from

which protected health information has been removed with use of machine learning". In: *JAMA network open* 1.8 (2018), e186040–e186040.

[57] Milad Nasr, Reza Shokri, and Amir Houmansadr. "Comprehensive privacy analysis of deep learning". In: *Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP)*. 2018, pp. 1–15.

[58] World Health Organization et al. *Active ageing: A policy framework*. Tech. rep. World Health Organization, 2002.

[59] Ronald Parr and Stuart Russell. "Reinforcement Learning with Hierarchies of Machines". In: *Advances in Neural Information Processing Systems* 10 (1997). URL: https://proceedings.neurips.cc/paper_files/paper/1997/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

[60] Holger R. Roth et al. "Federated Learning for Breast Density Classification: A Real-World Implementation". In: *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*. Cham, Switzerland: Springer, Sept. 2020, pp. 181–191. DOI: 10.1007/978-3-030-60548-3_18.

[61] Micah J. Sheller et al. "Multi-institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation". In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Cham, Switzerland: Springer, Jan. 2019, pp. 92–104. DOI: 10.1007/978-3-030-11723-8_9.

[62] Reza Shokri et al. "Membership inference attacks against machine learning models". In: *2017 IEEE symposium on security and privacy (SP)*. IEEE. 2017, pp. 3–18.

[63] Rohit Singla and Christopher Nguan. "Perspective Chapter: Service Robots in Healthcare Settings". In: *Trends in Assistive Technologies* (2022).

[64] *TensorFlow Federated*. [Online; accessed 31. Aug. 2023]. June 2022. URL: https://www.tensorflow.org/federated/get_started.

[65] *Understaffing in Nursing Homes | Risks and Consequences*. [Online; accessed 11. Aug. 2023]. Mar. 2023. URL: https://www.nursinghomeabuse.org/nursing-home-neglect/understaffing.

[66] *View of Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy*. [Online; accessed 22. Aug. 2023]. Aug. 2023. URL: https://journalprivacyconfidentiality.org/index.php/jpc/article/view/623/606.

[67] Chao Wang et al. "Explainable Human-Robot Training and Cooperation with Augmented Reality". In: *CHI EA '23: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–5. ISBN: 978-1-45039422-2. DOI: 10.1145/3544549.3583889.

[68] Shiqiang Wang et al. "Adaptive federated learning in resource constrained edge computing systems". In: *IEEE journal on selected areas in communications* 37.6 (2019), pp. 1205–1221.

[69] Weitian Wang et al. "Facilitating Human–Robot Collaborative Tasks by Teaching-Learning-Collaboration From Human Demonstrations". In: *IEEE Trans. Autom. Sci. Eng.* 16.2 (July 2018), pp. 640–653. DOI: 10.1109/TASE.2018.2840345.

[70] Yongqin Wang, Rachit Rajat, and Murali Annavaram. "MPC-Pipe: an Efficient Pipeline Scheme for Secure Multi-party Machine Learning Inference". In: *arXiv* (Sept. 2022). DOI: 10.48550/arXiv.2209.13643. eprint: 2209.13643.

[71] Yue Wang et al. "MASD: A Multimodal Assembly Skill Decoding System for Robot Programming by Demonstration". In: *IEEE Trans. Autom. Sci. Eng.* 15.4 (Jan. 2018), pp. 1722–1734. DOI: 10.1109/TASE.2017.2783342.

[72] Nicole Wetsman. "Hospitals are selling treasure troves of medical data — what could go wrong?" In: *Verge* (June 2021). URL: https://www.theverge.com/2021/6/23/22547397/medical-records-health-data-hospitals-research.

[73] *Workforce - Care Quality Commission*. [Online; accessed 11. Aug. 2023]. Aug. 2023. URL: https://www.cqc.org.uk/publication/state-care-202122/workforce.

[74] Yu Xianjia et al. "Federated Learning in Robotic and Autonomous Systems". In: *Procedia Comput. Sci.* 191 (Jan. 2021), pp. 135–142. ISSN: 1877-0509. DOI: 10.1016/j.procs.2021.07.041.

[75] Qian Xie et al. "Automatic Detection and Classification of Sewer Defects via Hierarchical Deep Learning". In: *IEEE Trans. Autom. Sci. Eng.* 16.4 (Mar. 2019), pp. 1836–1847. DOI: 10.1109/TASE.2019.2900170.

[76] Danfei Xu et al. "Neural Task Programming: Learning to Generalize Across Hierarchical Tasks". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 21–25. DOI: 10.1109/ICRA.2018.8460689.

[77] Dandan Zhang et al. "Explainable Hierarchical Imitation Learning for Robotic Drink Pouring". In: *arXiv* (May 2021). DOI: 10.48550/arXiv.2105.07348. eprint: 2105.07348.