Have your cake and eat it: Reconciling AI and Privacy in Deutsche Telekom's "Hallo Magenta" Digital Assistant

Harald Störrle
QAware GmbH, München

November 26, 2022

1 Introduction

Applications using statistical machine learning algorithms (SMLA) modeled after neural networks ("artificial intelligence", AI) are all the rage. In reality, AI Systems really are conventional IT systems that have one or more functional cores using AI techniques, today often of the deep learning variety. What this means is that all conventional wisdom on creating, testing, and running "AI Systems" is no different than creating, testing, and running conventional IT systems. Conventional wisdom still applies [Scu+15] — and is still not universally applied.

Now, on top of those conventional challenges, AI-infused IT systems pose genuinely new challenges regarding data management. The root cause is that supervised learning hinges on the availability of large amounts of high quality data. But where to get such data? We may be able to synthesize test data, but are they really representative of consumer behavior? And failing that, are test outcomes really meaningful?

So, using "real" data for testing is appealing. Also, it is readily available, and manually generating high-quality synthetic data can be very expensive. However, such data often originates with consumers, so they are the *property* of the consumers. Using such data without consumer consent outside defined and legitimate purposes violates a key tenet of the General Data Protection Regulation (GDPR).

When it is about processing data to provide a given service, user consent is implied simply by using the system. However, improving a system, e.g. by training an AI functional core, is not part of providing a service based on said AI. Thus, a separate item of consent is required for exploiting consumer data for system improvement. And that means, that only data from consenting clients may be used for training. Taking care not to confuse the consent level of data amounts to a set of cross-cutting requirements, which are notoriously difficult to implement.

To unpack this problem, we'll look at AI, data usage, and GDPR in turn. First, though, we introduce a case study — Deutsche Telekom's "Hallo Magenta" digital voice assistant, part of the Telekom Voicification Suite (TVS).

2 Case Study

Starting in 2015 with Amazon's Alexa, digital voice assistants ("voice bots"), have become one of the most prominent example of AI-infused systems. In 2019, Deutsche Telekom launched the "Hallo Magenta" voice bot. Figure 1 below explains the main functional flow of a voice bot using a weather forecast example.

- **Wake-Up-Word recognition** detects a certain phoneme sequence marking the start of a command addressed at the voice bot, such as "Alexa", "Hey Siri", or "Magenta". Only utterances following a wake-up-word are intended for machine processing, and getting this wrong is obviously a source of consequential errors.
- **Automatic Speech Recognition** transforms audio data into text. Under ideal circumstances, this works fairly well. But with background noise, dialects, concurrent voices from TV or radio sets, in-sentence language switches degrade ASR quality.
- **Natural Language Understanding** attempts to extract meaning from the audio transcript, yielding an Intent (the likely intention of the human speaker), and associated Entities (the grammatical objects of a sentence). In the above example, the intent would be to activate the "play" skill of the "Spotify" domain where the Entity would be "Taylor Swift" with type "artist".
- **Text-to-Speech** synthesizes an auditive output from a textual skill response. Given that we have full controll over the skill response, this is typically a well-behaved problem from an deep learning point of view.

Observe, that the system employs multiple deep learning systems as functional cores for distinct and very well-delineated tasks. These cores are embedded in a major conventional cloud-based IT system. Creating and operating such a system is a task for hundreds of experts over the course of years – a formidable challenge.

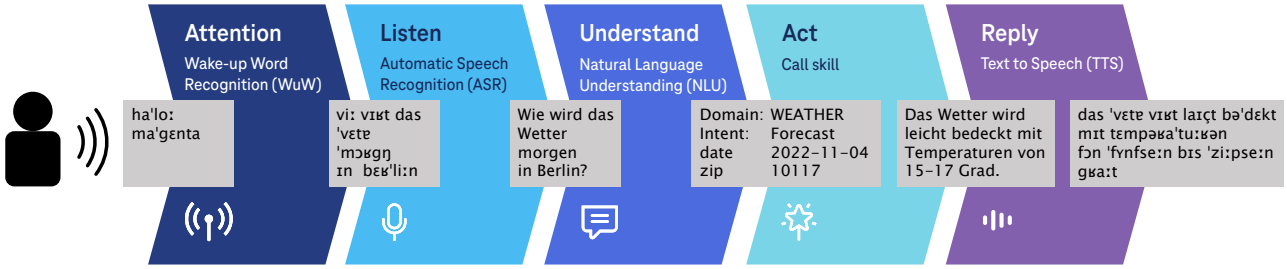

Figure 1: The main flow of a digital voice assistant: detect the Wake-Up-Word, transcribe speech to text, classify text as intents and recognize entities, execute some service, and generate speech output as a response.

3 AI and Privacy

In the 2010s, the decades-long stagnation in artificial neural networks was overcome by "deep learning" convolutional networks. Three factors were instrumental for their success: (1) deeply layered neural nets with reinforcement learning made possible by better convergence criteria for training, (2) greatly increased computing power delivered by massively parallel GPUs, and (3) the availability of vasts amount of human-labeled data for training and testing. [BLH21]

In order to be suitable for training, raw data needs to be processed in a laborious manual annotation process, yielding "Ground Truths" (GT), or data sets that are considered 100% correct. GT data sets are often split in half, where the first half is used for training, and the second half is used to test the trained model to assess its performance level. What exactly happens in annotation depends on the kind of data and the kind of model data extracted from the GT. For training an ASR, input utterances must listened to and transcribed correctly. Biometric data such as audio recordings necessarily identify speakers, invoking the highest level of data protection.

Today, privacy is a universal human right on a par with freedom of speech.[1] In 2018, the European Union introduced the General Data Privacy Regulation[2] to implement this basic right. The GDPR now applies to all data originating from any EU resident, irrespective of their citizenship, and, crucially, irrespective of who is processing their data or where they reside: any company processing data of EU residents needs to comply with GDPR. This includes companies offering websites, apps, social media, etc. to EU residents from outside of the EU, such as US based tech companies. Also, being the first of its kind, the GDPR heavily influenced subsequent regulations in California (CCPA), Japan (PIPA), and Korea (APPI), respectively. In a nutshell, the GDPR is effectively a global regulation– and not living up to it may incur stiff fines.[3]

[1] See §12 of the UN Universal Declaration of Human Rights.

[2] The GDPR text is readily available and surprisingly easy to read, see `https://advisera.com/eugdpracademy/gdpr`.

[3] Fines may reach 2% of global revenue, and up to 4% in cases of repeat offenses. While the data protection agencies exercised considerable restraint in awarding substantial fines at first, more and more fines have been awarded since. The "Enforcement Tracker" `https://www.enforcementtracker.com` lists a total of almost 1.300 fines totaling more than 2,000m€ up to October 2022.

4 Testing Deep Learning components

Testing an AI component in isolation looks like a unit test: present test data and assess the quality of the result. It is important, though, that test inputs be representative in order to give reliable test results. Using actual user data, representative sampling requires extensive analysis of actual user data, an analysis to be conducted continuously as user expectations and behavior, as well as system capabilities, are fluent.

Also, in order to be effective, data must be annotated manually to create a Ground Truth. Clearly, manual curation is a slow, laborious and expensive task (unless re-captcha users do it for free), even before trying to achieve GDPR-compliance. Thus, resorting to unsupervised learning seems like a good alternative at first: user data could be used directly as input to train AI components. Of course, user consent must be secured, and only data from users with explicit feedback can be used. However, this opens the path for manipulated inputs resulting in quick degradation, as the MS Tay scenario has shown.

Regarding test results, when a conventional test case fails, we know there is a defect. In AI systems, classifications are always likelihoods turned into categorical results only by (arbitrary) thresholds. Also, adding learning inputs (or even just permuting them) may affect the training result in unpredictable ways. Furthermore, many AI have a (relatively small) practical limit of the number of categories they can accommodate. Approaching it yields unexpected behavior.

Unlike conventional testing, we are only ever considering the performance of many test cases together. AI-testing is more about overall error density then about detecting results of individual test cases. In that respect, testing AI systems is more like statistical quality control known from manufacturing than software testing.

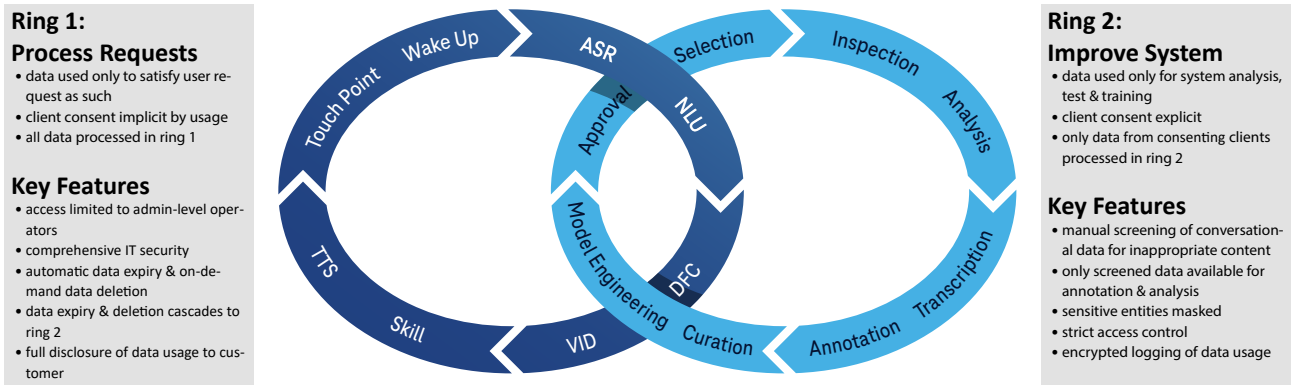

Figure 2: The domain level architecture of the Telekom Voicification Suite (TVS) is structured into two domains (or rings) with for two different classes of use cases. Only together can an AI-based system be fully operational, and only separated can GDPR-compliance be implemented.

5 Testing Deep Learning systems

Testing AI-infused systems at the whole-system level also offers some new challenges – though research so far has focused mainly on testing machine learning models, i.e., the machine learning components as such, as discussed above. For instance, a recent mapping study [Ric+20] found that two thirds of all publications on testing in machine learning deal with this type of test, while only 27% looked at system level testing, and a mere 1.5% specifically addressed the integration of ML based systems. We posit that, from an industrial perspective, system-level tests are indispensable. Any market-going product undergoes intensive system-level testing, including a massive battery of integration and end-to-end tests, usability and user acceptance tests, and follow-up market uptake and adoption evaluations. It appears that academic research is not very much concerned with the needs of industry, as far as testing of machine learning systems is concerned.

Turning back to our case study, we would like to highlight the fact that it is a consumer product and as such needs a good deal of end-to-end testing. Unlike end-to-end testing for purely digital products (think of a web application, say), a consumer product including hardware like a digital voice assistant require a physical setup to run E2E tests – not altogether unlike testing an embedded systems rigged up in an electro-mechanical test rig. So, an end-to-end test for a digital assistant sets up a physical speaker device in front of a device capable of producing and recording audios (in our case a Raspberry Pi equipped with the requisite speakers and microphones). For this set up, test data must be produced or procured (with all the issues mentioned above), and similarity between audio file must be determined to check actual vs. expected test results. While other types of AI-infused IT systems may not need the exact same setup, they share the problems with data privacy at the system level, too.

As a specialty of complex *chains* of AI components, deficiencies of one link in the chain may be compensated for by subsequent steps. For instance, incorrect ASR transcriptions may yield the correct NLU interpretation anyway. To achieve this, wrong or faulty transcripts with correct interpretations are added to the NLU Ground Truth. That is to say, individual items of test and/or training data may be incorrect at face value, yet contribute to a correct outcome at the system level. So, how *does* Magenta implement GDPR compliance?

6 How Magenta keeps users' privacy

Deutsche Telekom went to great lengths to keep customers' data private. While comprehensive IT security is indispensable (e.g. secure communication protocols, reliable encryption), in this article, we focus on functionality and design to create a system that fully respects customers' data privacy, yet allow for using their data for training and testing.

6.1 Architecture

DT implements privacy by design based on the overall domain architecture. We split the platform into two domains (or "rings") that operate independent from each other fig. 2. The first ring (fig. 2, left) contains all components to execute on a customer request. Using the platform amounts to consenting to *exactly* this amount of processing. In contrast, the second ring (fig. 2, right) contains all components to improve the platform, i.e., training and testing AI components. Usage in this ring requires explicit customer consent to this use case. The TVS isolates the two rings from each other. The only connection between them is a set of ETL jobs moving data from the first to the second ring, if and only if the flag for customer consent is set. So, only data from consenting customers ever gets used for anything but delivering the requested service.

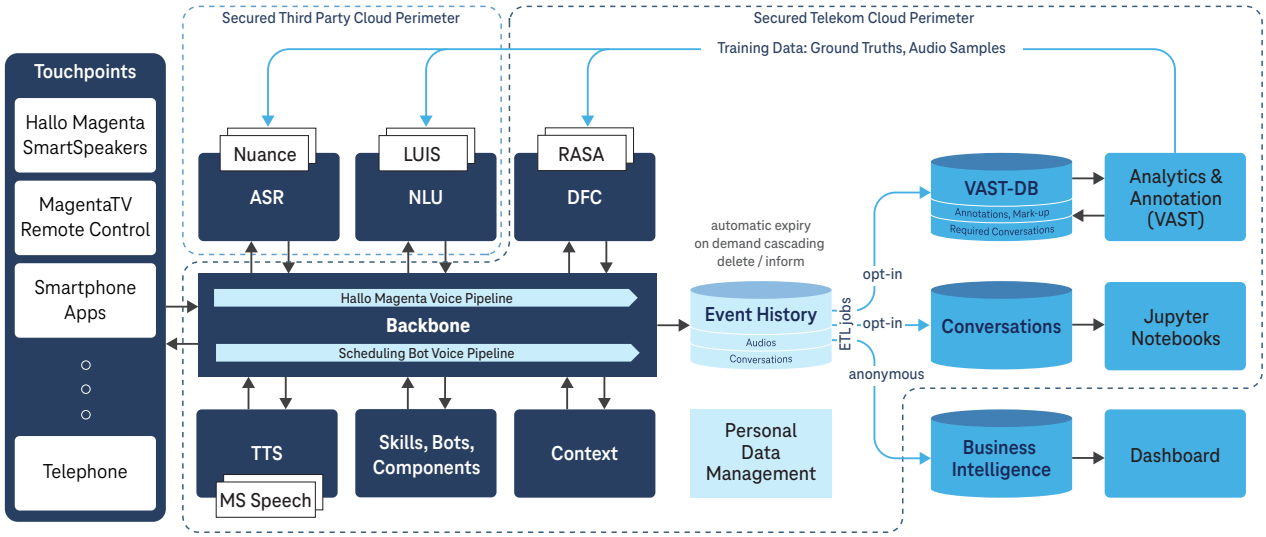

Figure 3: Technical architecture of the Telekom Voicification Suite (TVS)

See fig. 3 for a more technical view of the components of the TVS.

6.2 Screening and Masking

Once moved into the second ring, data items have a sensitivity level attached to them. Only "cleared" and "insensitive" data items may be used for any subsequent activities, including inspection, analysis, and training. Data items declared "sensitive" are hidden to all regular users and usages of the platform immediately. Turning data with "unknown" sensitivity into "cleared" requires a manual declaration which only a small contingent of specially equipped and trained employees of Deutsche Telekom may do. One of the restrictions is that they only do their job on DT premises using DT equipment. During the pandemic, their home offices had to be hardened appropriately.

Independent of manual screening, some sensitive details can be hidden automatically. This is particularly relevant for uses cases such as using a "Hallo Magenta" SmartSpeaker as a DECT device. In this use case, e.g., calling contacts or missed calls are announced to consumers. Consequently, such data will appear in the transcript, a class of data subject to telecommunication secrecy ("Fernmeldegeheimnis"), a regulation even more stringent than GDPR. Therefore, such data will be redacted, i.e., concrete data replaced by asterisks. This mechanism can also be used to protect DT employees profanity and inappropriate language.

6.3 Capability Control

Complementing the technical provisions to safeguard privacy, there are organizational rules. For instance, the spectrum of capabilities is split into 14 disjoint packages that may be awarded independently of each other to allow a fine grained control over visibility of data. The process of awarding capabilities grows increasingly difficult for increasing impact of said capabilities. The process of awarding capabilities is complemented by an off-boarding process. For instance, high-impact capabilities expire after a month and must be renewed manually. Additionally, the capabilities awarded are constantly monitored, to prevent capability-hogging.

6.4 Data Deletion

In order to satisfy the strict regulations on storage duration limitations, we take advantage of a feature of the underlying data storage solution (CosmosDB), which is capable of recording a time-to-live for each data entry. Upon creation of user utterance data packets, their time-to-live is set in accordance with the storage duration limitation in effect at the given time. When the expiry date is met, the data gets deleted automatically. The deletion then cascades to all secondary systems that made use of the data by way of the same ETL jobs that copied them to the respective secondary systems in the first place. This is achieved by a UUID for every data item. The same mechanism is used to propagate individual deletion requests, i.e., when clients delete all or some of their conversation data from the system by issuing such a request from the companion smartphone app coupled to their speaker devices.

Deletion does not actively propagate to dormant backups. Instead, backups are deleted after a while if they are not used. Backups that come into usage are subjected to the same deletion policy as live data is, that is to say, expired data is identified as being expired on loading, and discarded right away. Similarly, data that has been marked for deletion in the past by their owners will be deleted on load.

6.5 Transparency

A central aspiration of the GDPR is to provide consumers with transparency of and control over their data, as they are processed. This is implemented in three tiers. First, data expires automatically after a pre-set time, a feature of the underlying data store. Before expiry, consumers may delete parts or all of their transactional data from the platform, and may request a report on their data that is currently stored. Obviously, these rights extend to all parts of the platform, including components third-party suppliers. To ensure this, we have created a closed platform where user data never leaves the platform, no matter for what use case.

It is only by virtue of this closed platform that we could implemented the Personal Data Management component that ensures reliable and comprehensive deletion of or information about consumer's data. In the TVS, it takes considerable criminal energy to leak data personal, or "forget" about data stored on personal equipment, i.e., the proverbial infamous thumb drive with "just a few" conversations on it.

7 Summary

In summary, testing of AI components and AI-based systems exhibits some notable differences to conventional testing: As AI based system require constant training and testing, and the two activities fuse together; and operational data may fall under the GDPR requiring substantial provisions to ensure compliance. Data that is obviously "wrong" may still be "correct" from a E2E perspective.

Our case study shows that it is possible to use customer data for AI training & testing and yet be fully compliant to GDPR. In other words: you can have your cake and eat it, too, although it takes a major effort. However, this is assuming that compliance is a pivotal goal right from the start. Adding compliance to an existing system is a much, much more difficult task: as with all cross-cutting requirements, retrofitting them is notoriously difficult up to the point of being unpractical. Thus, platforms created pre-GDPR are at a substantial disadvantage, notably Amazon's Alexa.

8 Outlook

Critics may argue that digital speech assistants are just gimmicks without real benefit. Our concern with privacy is a waste of time, such critics may argue: if somebody values their privacy, they simply shouldn't use this class of device. Such criticism misses two points however: First, the "gimmick" is readily adopted by a great number of people in some application scenarios such as remote controls for TV and SmartHome applications. Second, The ability to conduct complex interactions by voice is a game changer in many service industries and the public administration. Where existing interactive voice systems force users to follow a numbing decision tree traversal, which rarely achieve more than channeling requests to some human operator, the ability to automate simple conversations allows an unprecedented degree of flexibility. In terms of usability, dialog based systems like the one presented here are to conventional interactive voice systems what WIMP-style direct interaction graphical user interfaces are to 3270-terminals.

Imagine you need to make an appointment with your family doctor. Traditionally, one would call in, or use a web or smartphone application to book a date. In order to take load off of doctors' front desk staff, Deutsche Telekom has launched Terminfinder, an application to arrange dates via the phone, based on the Telekom Voicification Environment described above. The system is in production for the first clients, and in piloting for several more expected to go live in early 2023.

Compared to web or smartphone apps, using a mere phone call does not rely on internet access and availability of suitable devices while keeping the same economic benefits, particularly, 24/7 availability, no waiting on busy lines, and lower operating cost by relieving staff of menial tasks. Also, it increases accessibility for people with poor eyesight, lack of digital literacy, or, in fact, literacy (which is a surprisingly high portion of the population). Apart from booking a date, the human voice is a biometric feature that is well suited for authentication, non-invasive health screenings and many other applications of great economic potential. In each of those cases, however, it is instrumental for business success that clients' privacy is respected. So, counter-intuitively, the privacy compliance problem is less of a consumer issue than a business impediment. A business that wishes to offer a voicification service is obliged to ensure GDPR compliance to its customers, a formidable challenge even for a Telecom giant like DT, and utterly insurmountable for any small or medium business wanting to reduce work load of their employees on the phone.

References

[BLH21] Yoshua Bengio, Yann Lecun, and Geoffrey Hinton. "Deep learning for AI". In: *Com. ACM* 64.7 (2021), pp. 58–65. DOI: doi.org/10.1145/344825.

[Ric+20] Vincenzo Riccio et al. "Testing machine learning based systems: a systematic mapping". In: *Empirical Software Engineering* 25.6 (2020), pp. 5193–5254.

[Scu+15] D. Sculley et al. "Hidden Technical Debt in Machine Learning Systems". In: *Proc. Advances in Neural Information Processing Systems (NIPS)*. Ed. by C. Cortes et al. Vol. 28. Curran Associates, Inc., 2015.