scripted

Volume 17, Issue 2, August 2020

The Ghost in the Machine – Emotionally Intelligent Conversational Agents and the Failure to Regulate 'Deception by Design'

Pauline Kuss and Ronald Leenes***

DOI: 10.2966/scrip.170220.320

Abstract

Google's Duplex illustrates the great strides made in AI to provide synthetic agents the capabilities to intuitive and seemingly natural human-machine interaction, fostering a growing acceptance of AI systems as social actors. Following BJ Fogg's captology framework, we analyse the persuasive and potentially manipulative power of emotionally intelligent conversational agents (EICAs). By definition, human-sounding conversational agents are 'designed to deceive'. They do so on the basis of vast amounts of information about the individual they are interacting with. We argue that although the current data protection and privacy framework in the EU offers some protection against manipulative conversational agents, the real upcoming issues are not acknowledged in regulation yet.

Keywords

Google Duplex; conversational agent; persuasion; manipulation; regulatory failure

* LL.M./Analyst, hy GmbH, Berlin, Germany, paulinekuss@gmx.net. This paper is based on Pauline Kuss, *Deception by Design for the Goal of Social Gracefulness: Ethical and Legal Concerns of Humanlike Conversational Agents*, Tilburg, 2019.

** Professor in Regulation by Technology, Tilburg Institute for Law, Technology, and Society, Tilburg, the Netherlands, r.e.leenes@tilburguniversity.edu

1 Introduction

In May 2018, a crowd of software engineers cheered at Google's I/O conference after the demonstration of Duplex, an intelligent voice agent that fits in your pocket sized smartphone and is able to make calls on behalf of its user in a deceivingly human-sounding voice. Intended to take care of cumbersome tasks such as the booking of appointments at hairdressers or restaurants, the previewed feature of Google Assistant convincingly mimics human behaviour by integrating speech disfluencies like 'hmmm' and 'ums' into its conversation – leading to a result applauded as a significant design achievement by some[1] and criticized as "Uncanny AI Tech"[2] by many others.

The development of conversational agents like Google's Duplex imbeds artificial intelligence into systems that are designed to deceive humans about their synthetic nature. However, we seem to have moved beyond the Uncanny Valley and no longer feel uneasy by this close to human vocal behaviour. Very soon we could find ourselves in a world where discerning whether we are talking to a human or an intelligent system on the other end of the communication channel becomes challenging. In particular the potential combination with other, currently unrelated, developments in voice AI which allow for the realistic imitation of a person's voice based on only a snippet of a recording,[3] suggests worrying scenarios of deliberate deception and fraud including cases of voice phishing or politically-motivated manipulation.

---

1 Joshua Montgomery, "Congratulations to Google Duplex! What's Next?" (2018), available at https://mycroft.ai/blog/congrats-on-google-duplex-whats-next/ (accessed 12 September 2018).

2 Mark Bergen, "Google Grapples With 'Horrifying' Reaction to Uncanny AI Tech" (*Bloomberg*, 10 May 2018), available at https://www.bloomberg.com/news/articles/2018-05-10/google-grapples-with-horrifying-reaction-to-uncanny-ai-tech (accessed 12 September 2018).

3 See for example: 'Lyrebird AI' part of 'Descript', https://descript.com/

In an era in which the term 'fake-news' has become a household word, the hazardous potential of digital technology as a facilitator of distributing deceptive messages is nothing new. However, the possibility of deceivingly accurate voice imitation, its potential integration into regular communication channels and the possibly unavoidable power of emotional associations attached to the sound of a familiar voice, suggest yet another, efficiently scalable and – possibly most worryingly – highly personalisable tool for actors with malicious intentions. But even in cases of conversational AI which discloses its synthetic identity upfront, an ethical consideration of the manipulative potential embedded in interactive, trust-generating and seemingly human intelligent systems appears appropriate.

From a legal perspective, the concept of 'manipulation' is difficult to grasp – where do we draw the line between manipulative and merely persuasive interventions?[4] Manipulation involves the intentional misuse of another's weaknesses – a skill which emotionally intelligent conversational agents (EICAs) can be expected to master with near perfection given their ability to access and process a vast amount of data and to adapt their behaviour accordingly.

The development of deceivingly human voice AI reflects a general trend towards increasingly seamless human-machine interaction. This is highly desirable from the perspective of technology developers because it supports convenience, thereby increasing users' enjoyment of and willingness to engage with respective systems. The concealment of machine-operated interaction, however, necessarily leads to a growing disguise of the presence of intelligent systems in people's surroundings. Additionally, cloud and fog computing accelerate a decoupling of devices' outer appearance from their ability to record,

---

[4] Cass R. Sunstein, "Fifty Shades of Manipulation" (2015) 1(3-4) *Journal of Behavioral Marketing* 213-244.

store and process data as their real processing power is no longer contained in their enclosures.[5]

Duplex marks the beginning of a development that promises to seamlessly embed a growing number of intelligent systems in our physical surroundings *and* in our emotional and social spaces. What does it mean when the sphere of human interaction becomes increasingly interwoven with the input of intelligent systems – systems that appear much better equipped to convincingly represent interests than 'normal' human beings? Given conversational agents' continuous processing of what their conversation partner is saying in order to facilitate adequate responses, how long will it take until such systems integrate in-depth analysis of *how* things are said into their response-engineering algorithm? Identifying personality traits of the interacting data subjects based on their choice of words,[6] or detecting a predisposition for psychosis[7] and Parkinson disease[8] based on non-verbal cues – where do we draw the line for what information intelligent conversational agents may derive from their counterpart's voice?

While the abilities of Google's Duplex remain quite restricted at this point, the complexity of legal and ethical concerns related to humanlike conversational AI is evident already. We can expect the Duplex feature of Google Assistant to spread to Europe. The question then arises in how far such concerns are addressed by the current European legal frameworks for data protection and

---

5 Flavio Bonomi et al., "Fog Computing and Its Role in the Internet of Things" [2012] *Proceedings of the first edition of the MCC workshop on Mobile Cloud Computing* 13.

6 See generally James W. Pennebaker and Anna Graybeal, "Patterns of Natural Language Use: Disclosure, Personality, and Social Integration" (2001) 10 *Current Directions in Psychological Science* 90-93.

7 Gillinder Bedi et al., "Automated Analysis of Free Speech Predicts Psychosis Onset in High-Risk Youths" (2015) 1 *npj Schizophrenia* 1-7.

8 Athanasios Tsanas et al., "Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's Disease" (2012) 59 *IEEE Transactions on Biomedical Engineering* 1264-1271.

consumer protection. Given the power of lock-in effects, which might lock-in unfortunate, (privacy-harming) original design choices into subsequent versions or follow-up products of a new technology,[9] the moment to consider what conversational AI shall look like, is now.

Besides describing the causes of and concerns related to the manipulative potential of humanlike conversational agents, this paper assesses some of the legal concerns in view of current European legislation. A focus on data protection and privacy law is chosen, motivated by the observation that, first, the risk of manipulative systems naturally implies the possibility of an infringement of individuals' decisional and intellectual privacy. Secondly, the extent of data processing involved directly affects the manipulative potential of a conversational agent: regulations on the permitted type of data processed, the employed processing techniques as well as on the required level of transparency and data subject control can thus be suggested as implicitly addressing the concern of manipulative systems. The legal analysis therefore considers the General Data Protection Regulation (EU) 2016/679 (GDPR) and the Privacy and Electronic Communications Directive 2002/58/EC (ePrivacy Directive) as well as the proposed ePrivacy Regulation replacing the latter.

While protection from manipulative systems might also be found in other legal fields such as contract and consumer protection regulations this requires information about specific operational settings – which is absent given the prospective nature of the developments sketched in this paper –, as well as a focus on one or more specific jurisdictions. Instead, the assessment of data protection and privacy law allows for a focus on those specific attributes of intelligent agents that form the basis of the particular manipulative potential of

[9] Woodrow Hartzog, *Privacy's Blueprint* (Harvard University Press, 2018).

such systems: their ability to 'know' a lot about the interacting individual – be it through real-time data processing or accessibility to other sources of data and customer profiles – and their capacity to adjust their behaviour accordingly in a statistically optimised fashion.

This paper is organised as follows. First, in section two, we present the context of our analysis, conversational agents. Next, section three explores the persuasive and manipulative aspects of these agents. Section four provides an analysis of the use of (deceptive) conversational agents from the perspective of the General Data Protection Regulation (GDPR) and the e-Privacy framework. Section five concludes the paper with a call to action.

2 Conversational Agents

The development of human-like machines capable of naturally conversing with people has been a long-standing goal for researchers in the field of human-computer interaction.[10] Increasingly, conversational agents, described as "dialogue systems often endowed with 'humanlike' behaviour",[11] emerge as common human-computer interfaces causing a "rise of conversation as platform"[12] as illustrated by intelligent voice assistants like Apple's Siri and Microsoft's Cortana.

Technology developers are keen on designing intelligent conversational agents that leave the user with an impression of *merely a human interaction*,

---

[10] Yaniv Leviathan and Matias Yossi, "Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone" (2018), available at https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html (accessed 12 September 2018).

[11] Giorgio Vassallo et al., "Phrase Coherence in Conceptual Spaces for Conversational Agents" in PCY Sheu et al. (eds.), *Semantic Computing* (Wiley, 2010).

[12] Ewa Luger and Gilad Rosner, "Considering the Privacy Design Issues Arising from Conversation as Platform" in R.E. Leenes et al. (eds.), *Data Protection and Privacy – The age of Intelligent Machines* (Oxford: Hart Publishing, 2018), pp. 193-212.

optimizing the agents' responses according to a counterpart's emotional and mental state or personality for the sake of user acceptance.

Considering Google's Duplex as an illustrative example, current developments in the field of humanlike conversational AI are driven by a combination of recurrent neural networks, automatic speech recognition technology and sophisticated text to speech engines which not only include speech disfluencies but also match the speed of their responses to the latency expectations of their conversation partner.[13]

The dynamic adaptation of speech latency to match the counterparty's expectations – thereby designing a conversation that is perceived as natural not only on the level of voice-quality but also with respect to the responsive behaviour of the intelligent agent – illustrates the sophistication of possibilities available to technology developers intending to design convincingly human-sounding AI agents. While dynamic adaptation of speech latency is merely one example of the greater research field of emotional speech synthesis,[14] it shows how the behaviour of intelligent systems can be dynamically adjusted, optimized to personally match individual conversation partners. What remains is the question regarding the pursuit of which interests and goals the responses of such system are optimized.

Without intending to pose allegations, it should be considered that there is a fine line between convincing or persuading people (e.g. into believing they are talking to a human being) and nudging or manipulating people. Although technology-induced power imbalances are far from novel, the level of

---

13 Leviathan and Yossi (*supra*, n. 10).

14 See generally Marc Schröder, "Emotional Speech Synthesis: A Review" *Seventh European Conference on Speech Communication and Technology* (2001).

sophistication with which they might be implemented in the context of conversational agents deserves particular attention.

3 Conversational Agents and Manipulation

Before providing an analysis of the specific characteristic of a deceivingly human voice and behaviour which endow EICAs with particularly powerful and thus potentially particularly concerning manipulative capacities, the persuasive and possibly manipulative nature of conversational agents must be explored.

3.1 Conversational agents as persuasive technology

According to Fogg, computers can 'persuade' – that is change people's behaviour or attitude – by appearing either as a *tool*, a *medium* or as a *social actor*.[15] He claims that computers' capacity to change people's behaviour and attitudes in their function as a social actor essentially depends on individuals' tendency to form relationships with technology. Supported by this human tendency, computers can exhibit persuasive effects through three distinct *persuasive affordances* when appearing in the role of a social actor:

- Establishment of social norms
- Invocation of social protocols
- Provision of social support and sanctioning

For the context of conversational agents, in particular the second and third affordances appear of importance: conversational agents can leverage social protocols to influence user behaviour such as the invocation of politeness norms, turn taking or reciprocity through the intentional expression of respective social

[15] B.J. Fogg, "Persuasive Computers: Perspectives and Research Directions" (1998) *CHI* 226-232.

cues. Likewise, the conscious provision of social support or sanctioning in the form of praise or criticism – a frequently observed dynamic in human-human interactions – can be easily used by conversational agents to affect individuals' conduct.[16]

Both of these persuasive affordances build on the human tendency to behave socially vis-à-vis computers, echoing the 'Computers are Social Actors' (CASA) paradigm developed by Reeves and Nass.[17] They suggest that anthropomorphism is driven by *mindless* user behaviour, which can be intentionally triggered through the provision of respective contextual cues – most notably through the expression of human features and characteristics.[18] It can therefore be assumed that intelligent systems appearing in the role of a social actor are more persuasive the more accurately they mimic human behaviour.[19]

Extending Fogg's model we propose two additional persuasive affordances that computers can use to persuade: *leveraging of situational or personal features* and *leveraging associations of existing relationships.* These capture, first, the power of data resources and processing capacities for fine-tuned personalisation and (real-time) adaptation of a system's behaviour and, second, the particular ability to communicate through a deceivingly accurate human voice. We suggest that these two additional categories will be of increasing visibility and relevance in light of human-sounding conversational agents.

---

16 B.J. Fogg, Gregory Cuellar, and David Danielson, *Motivating, Influencing, And Persuading Users: An Introduction to Captology* (CRC Press, 2009), p. 140.

17 Byron Reeves and Clifford Nass, *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places* (Cambridge: CUP, 1996).

18 Clifford Nass and Youngme Moon, "Machines and Mindlessness: Social Responses to Computers" (2000) 56 *Journal of Social Issues* 81-103.

19 Reservations to this might be implied by the uncanny valley effect.

3.2 Conversational agents as intentional actors

In order to define something as persuasive it is not enough that it simply influences human behaviour: although the Summer sun is a reason for people to put on sunscreen, we would be reluctant to talk about the sun as a persuasive actor. Fogg notes that since

> machines do not have intentions, a computer qualifies as a persuasive technology only when those who create, distribute, or adopt the technology do so with an intent to affect human attitudes or behaviours.[20,21]

The question to what extent EICAs have to be considered a persuasive technology therefore necessitates the identification of intentions involved – taking into account both the intentions embedded into the system by its creators as well as the interests of the user operating the system for a particular purpose.

Google promotes Duplex and its deceivingly human voice as offering a convenient tool that relieves customers from cumbersome phoning tasks while

---

[20] Fogg (*supra* n. 15), p. 226.

[21] Given the development of intelligent systems since the writing of this sentence in 1998, one could wonder whether self-learning machines might not one day be regarded as actors holding intentions themselves. Indeed, considering a scenario in which an intelligent phone assistant, after being informed that the originally desired timeslot was unavailable, requests whether an appointment could be possible anytime later that day. Does such request still fall within the user-dictated intention of booking an appointment or does it exceed it, making the question an expressed intention of the system itself? Obviously, the exact phrasing of the user's instruction – did she ask the system to book 'an appointment at 5pm' or did she additionally mentioned 'or if that's unavailable, any time later would also be fine' – would already impact the outcome of such analysis. It seems illogical though, that (if intentionality is considered a question of *ability*) the same machine could in some instances be regarded as intentional actor while being disregarded of such intentionality in other situations. This paper remains conservative with regards to personal interests of machines and understands the intentionality of a technology as equivalent to the intentions of its creators and employing users – reflecting what Fogg calls a computer's *endogenous* and *autogenous* intent respectively Fogg (*supra* n. 15), p. 226.

allowing for natural and intuitive human-machine interaction.[22] Besides user satisfaction and a general strive for AI success stories, additional motives can be assumed. For instance Duplex might serve the company's interest in attention-capturing technological novelty or the stimulation of user engagement. And, of course, Duplex will also generate and collect valuable conversation and customer data that can be leveraged for further improvements, subsequent products or premium price tags for advertisement deals. Users employing the calling assistant are likely to be motivated by the expectation of time-savings, convenience or the general enjoyment of playing with the newest feature of their phone.[23]

Also malicious and illegal user intentions are conceivable, including scenarios of intentional deception and voice phishing, a form of auditory identity fraud, with the ultimate goal of economic exploitation or political manipulation.

The concept of *manipulation* can be described as neighbouring the concept of *persuasion* on a *Spectrum of Influence*.[24] Manipulation is slightly more controlling than persuasion albeit not as incontrovertibly controlling as coercion, which makes a precise definition of manipulation more complex and elusive. Anne Barnhill offers a definition of manipulation that is useful for our purposes:

> Manipulation is intentionally directly influencing someone's beliefs, desires, or emotions such that she falls short of (the manipulator's) ideals for belief,

---

22 Leviathan and Yossi (*supra* n. 10).

23 Once Duplex-like systems escape the current limits of only operating in the niche contexts of booking restaurant tables or hairdresser appointments, further user intentions can be expected such as handing over uncomfortable social interactions to the intelligent assistant. Similarly, users could pretend to be their personal assistant by introducing themselves as such, intending to escape the full responsibility of their statements in a given conversation.

24 Ruth Faden and Tom Beauchamp, *A History and Theory of Informed Consent* (OUP, 1986).

> desire, or emotion in ways typically not in her self- interest or ways that are likely not to be her self-interest in the present context[25]

This suggests a consequentialist perspective as it takes the outcome *contrary to the self-interest* of the manipulated individual as one defining element. Complementing this first theoretical notion of manipulation, she offers a second, more intuitive definition following the thoughts of Joel Rudinow[26] that further emphasizes this situational relevance through a focus on situational weaknesses:

> Manipulation is intentionally making someone succumb to weakness or a contextual weakness, or altering the situation to create a contextual weakness and then making her succumb to it.[27]

Given this definition of manipulation, we can now illustrate how intelligent conversational agents can be used to persuade or manipulate individuals through the affordances described by Fogg and extended by us (Table 1).

---

[25] Anne Barnhill, "You're Too Smart to Be Manipulated By This Paper" (2010), available at https://vdocuments.mx/1-youre-too-smart-to-be-manipulated-by-this-paper-anne-barnhill-.html (accessed 21 July 2020), p. 22.

[26] Joel Rudinow, "Manipulation" (1978) 88 *Ethics* 338-347.

[27] Barnhill (*supra* n. 25), p. 24.

Persuasive Affordance	Persuasion Example	Manipulation Example
Establishment of social norms	*Intent:* Increase social acceptance of interacting with EICAs	
	Intervention: Win users' acceptance with rational arguments for the desirability of interacting with conversational agents (e.g. convenience) and the possibility to opt-out of interactions Priming of **target's (perceived) interest** while ultimate **choice remains with target**	*Intervention:* *Simply establish AI agent as given without revealing its identity; make an opt-out impossible or difficult; make alternatives to the interaction tedious, time-consuming or costly* Give targets **no choice** or artificially/unnecessarily **increase the cost of the alternative** to intended choice
Proposed Extension for the Context of Conversational AI:		
Leveraging associations of existing relationships	*Intent:* Trigger trust within a target by capitalizing on emotional associations of existing personal relationships	
	Intervention: Reveal the synthetic nature of the conversational agent through an introduction as personal assistant of a close friend in order to achieve a target's willingness to share their agenda for the purpose of finding a suitable date for a joint night out No pretence of own personality but **identification as intelligent assistant** and **explicit reference to the social relationship** involved in respective associations	*Intervention:* Employment of a voice imitation algorithm to simulate the voice of a person (closely) known to the target in order to leverage respective person's reputation, friendship or authority for malicious interests such as economic fraud or political manipulation **Pretence of own personhood** by the artificial agent; employing **identity fraud through voice phishing** to leverage the trust of existing personal relationships and social contexts for malicious purposes

Table 1: Examples of persuasion and manipulation by conversational agents leveraging the persuasive affordances of technologies appearing in the social actor functionality.[28]

[28] Due to space constraints we have only included two of the five affordances in the table.

The identification of interests and thus intentionality embedded within EICAs supports their denomination as potentially manipulative technology. Evidently, an assessment of Duplex's intentionality constitutes a challenging task, depending in its outcome on the particularities of future technical developments as well as on potential economic interdependencies between this and other Google products. Visible plurality of the interests involved suggests that the target population of the respective intentions might be equally multi-layered, including not only the direct user of the system but also the individual who will eventually interact with the EICA on the other end of the (phone) line, as well as potential misusers of the technology. For the sake of clarity, we will refer to respective individual as the *passive recipient* of a communication, describing the person interacting with the conversational agent without being the one actively initiating the human-machine interaction.[29]

[29] It is conceivable that an individual calls a restaurant that employs a conversational agent at their phone line. While in such scenario it would have been the individual who practically initiated the interaction, we nevertheless consider her the *passive recipient* as she intended to communicate with the human receptionist at restaurant rather than consciously choosing to involve an AI, resulting in a human-machine interaction.

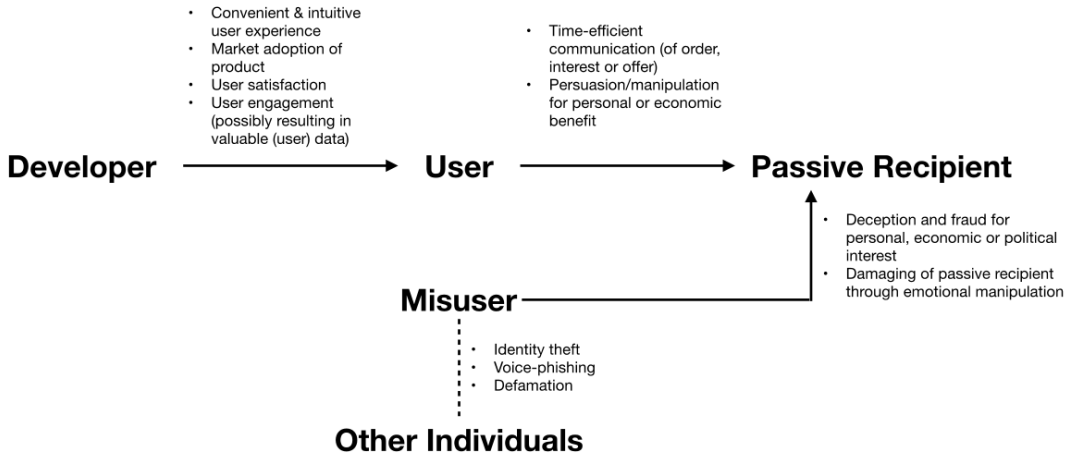

Figure 1: Network of possible intentions of developers, users, misusers and passive users of conversational agents and respective targets of persuasive or manipulative interventions.

Of note is that the recipient is the only actor unable to influence the intentionality attributable to the intelligent agent, as she is not involved in defining its *endogenous* or *autogenous*[30] intent. At the same time, the recipient is the target of both users' and misusers' interests and thus the subject of potentially related persuasive or manipulative intentions. Furthermore, compared to developers, users and misusers, the recipient is likely to be least knowledgeable about the system's technical nature, presence and capacities, suggesting an imbalance of power and calling into question the autonomy and rationality of the recipient's choice when agreeing to respective interaction – granted she is asked in the first place. The recipient therefore has to be regarded as the actor most in need of protection against the system's manipulative potential.

[30] Fogg (*supra* n. 15), p. 226.

3.3 The concerning power of persuasive conversational agents

If we accept that conversational agents are to be regarded as persuasive technology, we can explore their powers and the concerns they raise if adopted in conversations between a machine (initiator) and a natural person (recipient), for instance through a robocall. This section argues that EICAs are particularly powerful tools of manipulation due to their particular ability to trigger anthropomorphic user behaviour and their capacity for conversational engineering resulting in a personalisation according to mind, emotion and context.

3.3.1 Anthropomorphism and user expectation

Following the aforementioned idea that certain social cues can trigger mindless behaviour on the side of the human actor in human-machine interactions, the ability of EICAs to create an intuitive and deceptively accurate impression of everyday human-to-human interaction can be expected to support anthropomorphism and to trigger the expression of inappropriate social behaviour by concerned individuals towards the machine.

Elaborating on the concerns of intelligent systems imitating human behaviour in the commercial context, Kerr suggests that anthropomorphism is concerning from a consumer protection perspective, as people erroneously assume intelligent online assistants to be neutral or even customer-serving in their interests, overlooking the assistant's likely economic partiality.[31] [32] Kerr's

---

[31] Ian R. Kerr and Marcus Bornfreund, "Buddy Bots: How Turing's Fast Friends Are Undermining Consumer Privacy" (2005) 14 *Presence: Teleoperators and Virtual Environments* 647-655.

[32] Kerr also raises the point that the intentional design of intelligent systems aimed at triggering anthropomorphic behaviour appears intuitively repulsive from a moral point of view as it deludes individuals' into 'friendships' with artificial entities and the illusion of a mutually shared experience.

argumentation points towards the important link between designing deceptively accurate human-like AI, anthropomorphism, user expectation and consequential, potentially worrisome user behaviour. When picking up the phone, hearing a human voice on the other end of the line, people expect a social encounter between two human beings. Without a reason to challenge this assumption, they will implicitly expect their human-sounding conversation partner to also exhibit other human characteristics. They will thus *not* expect their counterpart to have access to a vast amount of data and processing power, enabling the same to sophisticatedly analyse the subtleties of conducted interaction and optimise its responses through statistical computations and profiling techniques.[33] Not expecting the actual (processing) capacities of the other party, individuals are unable to reasonably judge the potential consequences of their behaviour in given circumstances – reflecting what Luger describes as a missing "grammar of interaction".[34] Individuals will thus not be given any reason to adequately adapt their own behaviour.[35]

While the inappropriate anthropomorphism of intelligent systems might appear only mildly worrisome to some, the potentially accompanying erosion of people's agency to make informed, sovereign choices raises serious concerns regarding individuals' autonomy, dignity and privacy. Respective concerns are particularly obvious in cases where an AI does not identify itself as an artificial

---

[33] See for a similar account in the context of cochlear and retinal implants, Bert-Jaap Koops and Ronald Leenes, "Cheating with implants: Implications of the hidden information advantage of bionic ears and eyes" in M.N. Gasson, E. Kosta, and D.M. Bowman (eds.), *Human ICT Implants: Technical, Legal and Ethical Considerations* (TMC Asser, 2012) p. 113-134.

[34] Luger (*supra* n. 12).

[35] It is left to the reader to think of remarks that might slip your tongue carelessly in a casual conversation, which you might re-consider twice if you knew your counterpart to be a data-infused profiling machine.

agent, thereby intentionally deluding the expectation of the interacting person.[36] Capitalizing on this human tendency to *treat human what appears human*, the design of interactive systems imitating human behaviour with deceiving accuracy appears to imply a concealment of the system's mathematical capacities, underlying data resources and potentially involved stakeholder interest – be it intentionally or as an unintended side-effect.

It may be noted that this human tendency to interact socially with machines exhibiting human characteristics holds even in cases where the individual is well aware of the synthetic nature of their counterpart, as suggested by Weizenbaum's findings with ELIZA.[37] Moreover, respective discussion is nothing new: already in 1944 the Heider-Simmel illusion showcased a human willingness to attribute motives and character traits to inanimate objects as un-human as moving geometrical figures.[38] Also the ethical issue of deception through autonomous agents has already been discussed by existing scholarship such as Schafer's analysis of the use of autonomous agents for online police operations.[39] However, what is new with EICAs addressed by this article is – besides their formerly unknown sophistication – the broad market reach of respective technology and the ubiquity of their employment enabled through cloud infrastructure. These developments merit the here presented discussion as

---

[36] However, Weizenbaum's findings with ELIZA suggest that the human tendency to interact socially with machines exhibiting human characteristics holds even in cases where the individual is well aware of the synthetic nature of their counterpart (Joseph Weizenbaum, *Computer Power and Human Reason* (WH Freeman and Company, 1976)).

[37] *Ibid.*

[38] Fritz Heider and Marianne Simmel, "An Experimental Study of Apparent Behaviour" (1944) 57(2) *American Journal of Psychology* 243-259.

[39] Burkhard Schafer, "The taming of the sleuth – problems and potential of autonomous agents in crime investigation and prosecuting" (2006) 20 *International Review of Law, Computers & Technology* 63-76.

they imply the decentralisation and uncontrolled scalability of arising concerns discussed in the following.

3.3.2 Power imbalances and conversational engineering

The (intentional) concealment of the actual capacities of an intelligent agent, leading to a respective ignorance on the side of interacting individuals, threatens to introduce considerable power imbalances into the sphere of social interactions. Arguably, in most social encounters power imbalances always exist to some extent due to information asymmetries, resources inequality or motivational intransparency. However, respective concerns are multiplied exponentially with the introduction of socially engaging intelligent systems that vastly exceed their human counterparts in their capacity for data-driven communication design.

While human communicators are bound to learn from their own experience (or individual study), an artificial agent can hardly be seen as a single actor, but rather constitutes one instance of a bigger system that cumulatively gathers learning-relevant experiences, enabling each instance to feed on an abundance of data and models stored on its servers. Fogg describes several advantages of computers over humans with respect to their persuasive capacity, including computers' *persistence*; *ability to store, access and manipulate great volumes of data*; *scalability* and *ubiquity*.[40] Its access to a great amount of data, which can be leveraged as argument within as well as for the strategic optimisation of a persuasive agenda, grants conversational AI a significant advantage over humans in shaping an interaction and its outcome. Systems' potential capacity of real-time profiling to support optimized adaptation of an agent's behaviour or its fundamental characteristics raises a type of concern that might be referred to

[40] Fogg (*supra* n. 15).

as *conversational engineering*. The imbalance of power implied by (intransparent) conversational engineering appears morally worrisome as it favours intelligent conversational agents in their ability to steer an interaction for persuasive or even manipulative intentions while undermining persons' capacity to accurately judge the dynamics of the social encounter they find themselves in. While similar imbalances and its manipulative consequences might already be visible in existing applications of data-based decision making or profiling techniques,[41] we propose that they are particular prominent in the context of deceptively human, interactive EICAs due to their outstanding social character and how embedded they can become into every-day social encounters.

3.3.3 Personalisation according to mind, emotion and context

The idea of conversational engineering illustrates the ability of EICAs to personalize their behaviour with respect to their conversation partner, furthering the system's persuasive power. At the point of writing, no details on the exact scope of the data processing activities involved in the Duplex system have been released by Google.[42] The idea of an intelligent system which elaborately analyses your choice of words for potentially manipulative intentions or interprets your timbre and tone of voice for profiling purposes which go beyond the goal of presenting you with a pleasant interaction, might thus remain merely a hypothetical thought for now. However, a search for context relevant patents held by Google suggests that within Google work is done to develop intelligent

---

41 E.g. the dynamic pricing schemes of airlines which, based on their model and several data points known about an individual, personalize the ticket prices offered to respective customers with the intention of maximizing the overall profit by balancing premium prices against the risk of being left with empty airplane seats.

42 In existing publications Google states to use context parameters, conversation histories "and more" (Leviathan and Yossi (*supra* n. 10). which appears to be a conveniently broad notion neither including nor excluding any type of data really.

systems capable of adapting their behaviour according to the personality and current emotional state of an interacting individual, as well as their contextual and environmental surrounding.[43]

Google is surely not the only one developing intelligent systems capable of adapting their behaviour to the mental and emotional state of the interacting individual. Amazon recently patented an updated version of its virtual assistant Alexa that would analyse users' speech and other signals of emotion or illness, enabling the suggestions of activities suitable for a user's emotional state and the proactive offer to purchase medicine.[44] [45] Amazon's recent purchase of PillPack, a US-wide operating online seller of prescription drugs,[46] offers one explanation for the patent's focus on the medical market, illustrating the relevance of transparently assessing the web of interests that possibly affect the behaviour of intelligent assistants, as such systems are likely to be less objective than the general user might expect.

---

43 For instance William Zancho et al., "Determination of Emotional and Physiological States of a Recipient of a Communication", available at https://patentimages.storage.googleapis.com/e6/d6/c8/04858db5fb697b/US7874983.pdf (accessed 21 July 2020); Bryan Horling et al., "Forming Chatbot Output Based on User State" https://patents.google.com/patent/US9947319B1/en (accessed 21 July 2020).

44 Huafeng Jin and Shuo Wang, "Voice-Based Determination of Physical and Emotional Characteristics of Users", available at https://patents.google.com/patent/US10096319B1/en (accessed 21 July 2020).

45 Highly interesting in terms of its dubiousness is also the included patent claim for targeting advertisements to match the detected mood of a user, offering advertisers the possibility to pay for emotionally targeted placement of their products – a promising marketing strategy given the significant correlation between impulsive buying and customer features such as personality profiles (Bas Verplanken and Astrid Herabadi, "Individual Differences in Impulse Buying Tendency: Feeling and No Thinking" (2001) 15 *European Journal of Personality* S71) or current emotional state (Peter Weinberg and Wolfgang Gottwald, "Impulsive Consumer Buying as a Result of Emotions" (1982) 10 *Journal of Business Research* 43-57).

46 Margi Murphy, "Amazon Sends Pharmacy Stocks Tunbling after Snapping up Online Chemist" (*The Telegraph*, 2018), available at https://www.telegraph.co.uk/technology/2018/06/28/amazon-sends-pharmacy-stocks-tumbling-snapping-online-chemist/ (accessed 19 October 2018).

3.4 Dual Use and the Weaponization of Conversational Agents

As for most technologies, intelligent systems have the potential for dual use and thus carry the risk of weaponization.[47] While the use of intelligent calling agents in the context of armed conflict might appear as an unrealistic scenario at first sight, the threat of serious misuse of such systems in contexts such as political campaigning or electoral fraud is actually highly concerning. Considering the already discussed persuasive potential of human-like voice AI, emerging systems combining conversational abilities with (already existing) voice imitation algorithms[48] intensify such worries. How unlikely are scenarios of employing such system for mass callings – possibly using the voice of popular political figures – intended to influence political dynamics in a particular country?

The potential risk of technology as a tool for (political) manipulation is surely not new arising only with the advent of intelligent conversational agents.[49] And yet, intelligent conversational agents display two characteristics that suggest them as a particularly potent instrument for potential manipulation: first, as the calls are conducted automatically without the need for human intervention, communicating (manipulative) messages through conversational agents is highly scalable. Not even the precise wording of the intended conversation would have to be humanly designed. Secondly, while scalability might also be seen for the spread of digital video footage or nudge-intending social media content, the channel of a phone call gives conversational agents a

---

47 Goncalo Carrico, "The EU and Artificial Intelligence: A Human-Centered Perspective" (2018) 17 *European View* 29-36.

48 See for example: 'Lyrebird' (*supra* n. 3).

49 For illustration of existing possibilities one may consider the supposed engagement of Cambridge Analytica in the 2016 US presidential election or popularly discussed examples of visual deep-fakes involving well-known politicians.

much more personal character. A phone call is explicitly directed at one single person and constitutes a social interaction quite familiar to most people. Consequentially, the message conveyed can be highly individualized to optimize the impact of the intended nudge. Additionally, recipients might be less sceptical towards messages received through *personal* interaction, as the possibility of dangerously authentic fake-calls is less prominent within the public awareness compared to by now better-known examples of visual deep-fakes.

4 Existing Legal Framework

Now that we have an understanding of the potential of emotionally intelligent conversational agents that produce increasingly natural conversation bringing to bear knowledge about persuasion and manipulation, connected to information about the state of mind of the recipient and their emotions, as well as information from the vast trove of the recipient's onli*f*e, we can explore what this entails from the perspective of the law, in particular, data protection and privacy regulation (in the EU). In this context, the data protection (GDPR) and e-Privacy frameworks are most prominent.

4.1 GDPR

The General Data Protection Regulation 2016/679 (GDPR) regulates the processing of personal data which is defined as "[1] any information [2] relating to an [3] identified or identifiable [4] natural person".[50] Personal data is a very

[50] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (hereinafter 'GDPR') Art. 1(1).

broad notion.[51] The Art. 29 Data Protection Working Party notes that the term includes *any* information regardless of its nature, content or format.[52] Acoustic information, including voice recordings are explicitly listed as personal data[53] and additionally referred to as an example of *biometric data*, which come with the particularity of providing both content about an individual as well as a link between the same and some piece of information.[54] Voice recordings are thus to be regarded as identifiers of natural persons, implying fulfilment of the definitional elements [3] and [4] above. With respect to information relating to a natural person derived from voice recordings, the element of 'identified or identifiable' is satisfied when respective data can be linked to a natural person through any "means reasonably likely to be used [...] by the controller or by another person".[55] The status of information as personal data is thus dynamic, depending on context and advances of re-identification technologies,[56] which suggests considering information derived from individuals' voices as personal data until effective irreversible anonymisation can be assured. The use of respective information for the personalization of an agent's behaviour suggests that a link between the data and an individual can be assumed.[57] Also technical information such as smartphone identifiers, IP addresses or phone numbers are linked to the person addressed by the conversational agent, contributing to

---

51 Article 29 Data Protection Working Party, "Opinion 4/2007 on the Concept of Personal Data" (WP136, 2007).

52 *Ibid.*, p. 6.

53 *Ibid.*, p. 7.

54 *Ibid.*, p. 8.

55 GDPR, Recital 26.

56 Nadezhda Purtova, "The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection Law" (2018) 10 *Law, Innovation and Technology* 40-81, p. 47.

57 Article 29 Data Protection Working Party, "Opinion 05/2014 on Anonymisation Techniques" (WP216, 2014), p. 7.

making this individual an identifiable person,[58] following the "standard of the reasonable likelihood of identification".[59]

'Relating to' a natural person [element 4], again, has a broad scope. Such relation can be either in content, purpose or result.[60] Relating through 'content' is rather straightforward. It refers to information *about* a person, which in the current context would include personal phone numbers, but also personality traits or mental and emotional states of an individual should such information be derived through voice analysis. If the information collected through the conversation is used or likely to be used to "evaluate, treat in a certain way or influence the status or behaviour of an individual",[61] it relates to this person per 'purpose'. Data that relates to a person as 'result' if "their use is likely to have an impact on a certain person's rights and interests".[62] Such a result is present irrespective of the gravity of the impact – the different treatment of one person from another suffices.[63] EICAs adjust their behaviour according to individual interactions and perceived environments, what Hildebrandt refers to as "data-driven agency".[64] In such a context, "any information can relate to a person by reason of purpose, and all information relates to a person by reason of impact."[65] It follows that whatever information is processed by a EICAs for the purpose or

---

58 European Commission, "What Is Personal Data?", available at https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-personal-data_en (accessed 21 July 2020).

59 Purtova (*supra* n. 56), p. 47.

60 Art. 29 Working Party (*supra* n. 51), p. 10.

61 *Ibid.*, p. 10.

62 *Ibid.*, p. 11.

63 *Ibid.*, p. 11.

64 Mireille Hildebrandt, "Law as Information in the Era of Data-Driven Agency" (2016) 79 *The Modern Law Review* 1-30.

65 Purtova (*supra* n. 56), p. 55.

with the result of (accidentally[66]) treating one individual different than another has to be considered personal data triggering protection under the GDPR.[67]

The data obtained from the recipient by the conversational agent, either through voice, or additional sources, can only be processed if the controller has a legitimate ground for such processing (Art. 6 GDPR). Considering that no contractual relationship exists between the individual interacting with the EICAs and the agent's provider, that the latter has no legal obligation to process the conversational data and neither does a public interest exist in respective processing, paragraphs 6(a) data subject consent and 6(f) necessity for the purpose of a controller's or third party's legitimate interest appear the only grounds reasonably available to legitimize the processing of personal data in the context of EICAs under Art. 6 GDPR. Importantly, Art. 6(f) requires a balancing test of the interests involved, clarifying that the legitimate interest of a controller or third party constitutes no legitimizing ground for processing where it is overridden by the interests or fundamental rights of the data subject concerned.

Recital 47 elaborates on the concept of 'legitimate interests', noting that "reasonable expectations of data subjects based on their relationship with the controller" should be taken into account, as legitimate interests might for example exist in cases where a client or service relationship is present between the data subject and the controller.[68] The processing of personal data occurring through the employment of a EICAs by individuals for the purpose of placing a restaurant reservation or, reversely, the use of such system by a restaurant for the

---

66 *Ibid.*, p. 56.

67 One could challenge whether the customization of agents' voice, choice of words or pace of speech constitutes sufficiently different treatment. However, the Art. 29 Working Party explicitly established a very low threshold of impact, implying that such customization are to be regarded as 'relating to' an individual by purpose and/or impact.

68 GDPR, Recital 47.

answering of customer-calls can therefore be expected to find justification under Art. 6(f) granted the balancing test is passed. Recital 47 furthermore states that "the processing of personal data for direct marketing purposes may be regarded as carried out for a legitimate interest", suggesting that the operation of EICAs for unsolicited marketing calls may equally be legitimized under the exception of legitimate interests if these are adequately balanced against the interests, rights and freedoms of the receiving individual.

The Art 29 WP holds that the requirement constitutes no "straightforward balancing test" but instead "requires full consideration of a number of factors",[69] including safeguards and measures in place such as easy-to-use opt-out tools.[70] The WP emphasizes the threshold of 'necessity' required by the concerned article and clarifies that in order to satisfy Art. 6(f) a 'legitimate interest' must be (a) lawful, (b) sufficiently specific and (c) not speculative.[71] The scale of data collection, lack of transparency about the logic underlying the processing, sophistication of profiling and tracking techniques employed as well as a resulting de facto (price) discrimination are factors that could negate Art. 6(f) as a valid basis of lawful processing.[72] According to the Working Party, the potentially negative impact on a data subject has to be considered in a broad sense, encompassing also emotional distress such as irritation or fear as well as chilling effects resulting from the impression of continuous monitoring.[73] Validity of Art. 6(f) in the context of EICAs thus depends on a case-to-case assessment of involved interests, including the consideration of inter alia the

---

69 Article 29 Data Protection Working Party, "Opinion 06/2014 on the Notion of Legitimate Interests of the Data Controller under Article 7 of Directive 95/46/EC" (WP217, 2014), p. 3.

70 *Ibid.*, p. 31.

71 *Ibid.*, p. 25.

72 *Ibid.*, p. 32.

73 *Ibid.*, p. 32.

nature of concerned data, the relationship between the data controller and data subject as well as the expectations of the latter with respect to data confidentiality. The processing of personal data for the purpose of operating a conversational agent that expresses (financially) discriminatory, deceptive or outright manipulative behaviour or which in any other way has a considerable negative impact on the interacting individual clearly cannot be justified on the ground of legitimate interest.[74]

As we have outlined above, voice analysis can offer highly sensitive insights relating for example to an individual's emotional or mental health. This would render the data processed by EICAs under 'special categories of personal data' in Art. 9 GDPR. Art. 9 excludes legitimate interest of the controller as a valid processing ground. Data subject consent on the other hand is a valid ground in Art. 9(2)(a).

Suggesting an even stricter interpretation, it could be argued that also with respect to less sophisticated conversational agents, which do not involve the processing of special category data on first sight, data subject consent should be regarded the only valid basis for lawful processing. Considering that the content of a communication can, potentially, always include sensitive information concerning one of the conversation partners or another individual, the processing of special category data by systems which are restricted to the processing of conversational content only – a processing that is necessary to enable an agent to generate adequate responses – cannot be ruled out entirely. Moreover, also without an analysis of someone's voice, the choice of words, which are inevitably

[74] The purpose(s) for which data are being processed (art. 5(1)(b) GDPR) by the conversational agent are a significant issue to be discussed as well, but due to space constraints we leave this for another occasion.

processed by any type of conversational agent, can reveal sensitive insights concerning one's emotional state, cognitive complexity or personality.[75]

A precautionary approach would therefore proclaim the necessity to justify any processing involving conversational data by conversational agents under Art. 6/9 GDPR only on the basis of consent.[76]

4.1.1 Fairness of intelligent systems

The General Data Protection Regulation 2016/679 (GDPR) regulates the processing of personal data through a framework of principles set out in Art. 5,[77] listing of particular importance in respect to previously identified challenges of manipulative systems the principles of *fairness* and *transparency.*

The only available ground for a lawful processing of personal data in the context of conversational agents seems to be consent. Challenged by the principle of fairness, the legitimizing power vested in user consent stands in clear contrast to data subjects' limited ability to understand the complex technology behind intelligent systems – especially when such complexity is hidden behind the veil of apparently human-like familiarity. This holds particularly true for intelligent systems that process not only the verbal content of a conversation but also the voice features of interacting individuals. As the general data subjects' knowledge about the revealing nature of voice analysis can be expected to be marginal at most, the GDPR – in order to honour the principles of fairness – should be read

---

75 Y.R. Tausczik and James W. Pennebaker, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods" (2010) 29 *Journal of Language and Social Psychology* 24-54.

76 If processing is to be legitimised by consent, this raises a whole range of issues, because the consent must be informed, freely given, unambiguous etc. We leave these for another occasion.

77 A much more extensive treatment of the applicability of the GDPR and its requirements can be found in Pauline Kuss, *Deception by Design for the Goal of Social Gracefulness: Ethical and Legal Concerns of Humanlike Conversational Agents* (Tilburg, 2019).

as mandating comprehensive explanations aimed at supporting data subject's understanding of the nature and potential consequences of such processing. Moreover, (mis)using the insights derived from such voice analysis for the purpose of designing more *persuasive* – a.k.a. *manipulative* – systems appears to violate the principle of fairness, raising the question of where to draw the line between 'making a user experience more intuitive and pleasant' and 'designing a system that pushes all the right buttons to trigger users' sympathy and (inappropriate) trust'. Likewise, EICAs that fail to disclose their synthetic nature at the beginning of an interaction, thereby misusing their ability to authentically mimic human behaviour for the intended deception of interacting individuals, violate the principle of fairness.

4.1.2 Transparency of AI behaviour

The importance of disclosing a EICA's synthetic nature illustrates the association between the principles of fairness and *transparency*, and their respective relevance in the context of manipulative systems: a lack of transparency concerning the synthetic nature of the calling voice, the data and processing capacities available to or the interests of the same, result in an unfair imbalance of power that greatly disadvantages the called individual who finds itself an easy target for the potentially opaque intentions of the calling AI. One obvious difficulty arising in this context is the challenge of determining precisely where persuasion ends and manipulation begins. Another difficulty arises in respect to the detection and evidencing of manipulative behaviour of a conversational agent: if done well, individuals targeted by a manipulative system are likely not to notice the manipulation – let alone in cases where they are not even aware of the fact that they are interacting with an AI rather than an actual human being at the other end of the line. Respect for the principles of fairness and transparency is thus fundamental and a clarification of their exact meaning and related requirements

in the context of human-sounding voice AI would be essential.

4.2 Privacy law

In their "Typology of Privacy", Koops et al. describe privacy as a complex "set of related concepts that together constitute privacy"[78] and identify types of privacy, including *privacy of relations*,[79] *privacy of person*[80] and *privacy of personal data*. According to the authors privacy can imply both a *freedom from*, as well as a *freedom of* something.

As a freedom *of*, privacy's close association with the concept of 'autonomy' is apparent.[81] While privacy as a negative right appears more directly connected to data protection concerns, the understanding of privacy as a positive freedom highlights the strong link between privacy protection and the issues of manipulation and deception.

Referring to the eight primary types of privacy suggested by Koops et al., the context of EICAs most visibly gives rise to concerns with respect to individuals' communicational, intellectual,[82] decisional[83] and associational[84]

---

78 Bert-Jaap Koops et al., "A Typology of Privacy" (2017) 38 *University of Pennsylvania Journal of International Law* 483-575, p. 488.

79 Encompassing the protection of the establishment of social relationships and communication.

80 Encompassing the protection of thought and personal decision-making.

81 Koops et al., *supra* n. 78, p. 514.

82 The intentional design of systems meant to deceive people with respect to their synthetic nature challenges the privacy of persons' opinion and believes encompassed in this privacy type.

83 Decisional privacy appears generally challenged by persuasive and manipulative technologies and is equally at risk in the context of intelligent systems which conceal their synthetic nature as such undermine individuals' capacity to make self-serving privacy choices.

84 Describing individuals' freedom to choose whom to interact with, associational privacy is challenged in cases where adequate disclosure of the synthetic nature of a conversational agent is missing as this undermines individuals' informed choice concerning the interaction they decide to engage in.

privacy.[85] While the GDPR's broad protective scope appears to already safeguard communicational and informational privacy, it seems important that privacy law complements the respective legislation in particular through provisions emphasizing the importance of transparent disclosure of intelligent agents so as to ensure the protection of individuals' intellectual, decisional and associational privacy.

4.3 ePrivacy Directive

In contrast to the GDPR, the ePrivacy Directive[86] is not restricted to the protection of *personal* data itself but covers confidentiality of communication more broadly.

Of particular relevance in our context is Art. 13 of the ePrivacy Directive, which introduces the concept of "automated calling and communication systems without human intervention (automatic calling machines)" to refer to marketing calls "made by an automated dialling system that plays a recorded message".[87] While the technology of EICAs as discussed here did not exist at the time of the Directive's writing, its similarity with 'automatic calling machines' proposes applicability of Art. 13 by analogy. Similar to the unsolicited call by an automatic calling machine, the individual responding to the call of a EICA is likely not to have requested the interaction with the machine.[88]

---

[85] It could be argued that human-sounding conversational agents also threaten to compromise spatial privacy, as individuals' capacity to execute control over the actors they admit to the private space of their personal phone line would be undermined in cases where they are unable to know of the synthetic nature and thus of the computing capacities of the voice at the other end.

[86] European Parliament; Council of the European Union, "Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 Concerning the Processing of Personal Data and the Protection of Privacy in the Electronic Communications Sector (Directive on Privacy and Electronic Communications)" (2002) L 201 *Official Journal of the European Communities* 37.

[87] Information Commissioner's Office, *Guide to the Privacy and Electronic Communications Regulations* (2018), p. 16.

[88] On the other hand, in cases where a conversational agent is employed to place a reservation with a restaurant, it could be argued that the latter did request such call implicitly by stating

Relevance of Art. 13 for the context of conversational agents seems to depend on the provision's underlying intention: is the article meant merely as a protection from the nuisance of unrequested mass-calls or does it aim to safeguard individuals when interacting with automated communication systems more generally? Recital 40 of the Directive describes the provision as a safeguard against the intrusion of privacy caused by highly scalable automated calling machines. Considering the connotation of 'intrusion', it appears valid to suggest an analogy between the purpose of Art. 13 and the tort of trespassing.[89] Among many operational purposes, the employment of EICAs for automated marketing calls is indeed conceivable, suggesting unsolicited communication as an additional concern arising with autonomously operating calling agents, in the context of which Art. 13 ePrivacy Directive would be clearly applicable. However, it can be debated whether the intended protective scope of the provision also covers scenarios similar to those described, in which the recipient's interest in the call would not to be challenged if the caller was a human being. Clearly, in such case it is not the occurrence of the call itself, but rather the processing of personal data by and the persuasive potential of EICAs that might give rise to privacy concerns.

The "Typology of Privacy"[90] illustrates that privacy interests relate not only to spatial privacy – the type of privacy protected by the action of trespass –

---

an interest in being called for the purpose of reservations when offering a phone number to prospective customers. Also with respect to private communications, such as the scheduling of a personal meeting between two friends, the interacting individual might not have chosen to converse with a machine and yet, having an interest in seeing his friend, she can be expected to welcome the call.

89 Such analogy was made by the California Supreme Court in the context of unsolicited e-mails in *Intel Corp. v. Hamidi*, reasoning that the act of connecting oneself to the internet or buying a telephone cannot be considered an invitation to receive masses of unwanted e-mails and phone calls. See *Intel Corp. v. Hamidi* 30 Cal. 4th 1342 (2003).

90 Koops et al., *supra* n. 78.

but that they also, inter alia, include individuals' decisional and intellectual privacy. While it appears that Art. 13 was written with the protection of the former in mind, one can argue that the purpose of protecting individuals from an intrusion of their privacy should be interpreted more broadly as to acknowledge the concept of privacy in its complexity. Following such reasoning, we suggest to read Art. 13 as to safeguard individuals more generally when interacting with automated communication systems, considering the potential infringement of individuals' communicational, decisional and intellectual privacy through the data processing involved in and the persuasive potential of such systems. Irrespective of a recipient's general interest in the call, the recipient does have an interest in being protected – if not against the occurrence of the communication itself, then still against the potentially privacy-intrusive implications of interacting with an intelligent data-processing system.

Under Art. 13 user consent is required to allow for respective calls, implying that even if conversational AI did not involve the processing of communication data or personal information, the interacting individual would have to give prior agreement to a call they themselves did not initiate. However, since the article lists a "purpose of direct marketing" as explicit attribute of the automated calling systems covered by its application, it declares itself inapplicable for conversational agents employed in a non-marketing context. Besides posing the requirement of target consent, Art. 13(4) ePrivacy Directive explicitly prohibits "in any event (...) practice[s] which disguise or conceal the identity of the sender on whose behalf the communication is made". While this provision appears to offer a solution to the identified need to demand the transparent disclosure of intelligent systems, again its application is limited to practices with "the purpose of direct marketing".

4.4 The ePrivacy Regulation

While a first proposal text has been published in January 2017, work on the Regulation's draft continues at the point of writing, leaving the most recent proposal and comments published by the Council in September 2018 as the basis for the current analysis.

The ePrivacy Regulation[91] appears to fill the regulatory gap caused by the Directive's restricted definition of 'automated calling machines' by explicitly defining "automated calling and communication systems" (Art. 4(h)), leaving aside the necessary context of marketing purposes criticized previously. The respective definition refers to "systems capable of automatically initiating calls to one or more recipients in accordance with instructions set for that system, and transmitting sounds which are not life speech"[92] – a definition that seems to cover emotionally intelligent conversational agents. While paragraph (3)(f) of the same article lists such system as one of multiple technologies that can be used for the purpose of "direct marketing communications", the ePrivacy Regulation achieves a disjunction of this purpose and the definition of automated communication systems that improves the respective provision of the Directive. However, a stand-alone section elaborating on the risks, rights and requirements related to automated communication systems remains missing in the current Regulation draft. In fact, concerns such as the need to obtain recipients' consent prior to the interaction with an EICA or the requirement of identity disclosure are only raised with regard to unsolicited and direct marketing communications

---

91 European Commission, "Proposal for a Regulation of the European Parliament and of the Council Concerning the Respect for Private Life and the Protection of Personal Data in Electronic Communications and Repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communication", available at https://ec.europa.eu/digital-single-market/en/news/proposal-regulation-privacy-and-electronic-communications (accessed 21 July 2020).

92 *Ibid.*, art. 4(3)(h).

in Art. 16 of the Regulation. A consideration of scenarios in which automated calling systems could be used for purposes other than marketing, such as the scheduling of personal appointments, which nevertheless implies risks for the rights and freedoms of the interacting individuals due to the necessarily involved processing of their (conversational) data, thus remains absent. Similarly, while the Regulation demands revealing the identity of the natural or legal person behind the automated marketing communications system thereby suggesting a promising contribution to the protection of individuals' privacy interest, it lacks a general requirement to disclose the synthetic nature of deceivingly human-sounding EICAs in non-marketing contexts.

5 Conclusion

While human-sounding, emotionally intelligent conversational agents (EICAs) constitute a persuasive technology by nature – simply because they inherently persuade interacting users to treat them according to social protocols through their human-imitating behaviour – their designation as manipulative technology depends on a case-to-case assessment of the particular intentions embedded, their potential consequences as well as the pursued way of achieving the same. The degree of control exerted and the extent to which targets' capacity for autonomous decision-making is intentionally undermined should be considered markers to identify the presence of manipulative rather than merely persuasive interventions.

Convincing human AI agents are likely subject to anthropomorphism, resulting in mindless social behaviour of the interacting individuals who might easily misjudge the computing capacities and thus the overall power of the friendly voice on the other end of the phone line. With the general trend towards embedded and more seamless computing systems, computers' presence and the

potential consequences thereof become increasingly intransparent for individuals who nevertheless find themselves subjected to the techno-regulatory impact of such systems. Besides ethical concerns related to the affront to individual freedom, the danger of identity fraud and the justifiability of manipulation, respective opaqueness of the systems also deprives individuals of the informational basis needed to make sovereign choices with respect to the protection of their privacy and personal data. In a way, the strength of EICAs is also their greatest weakness: they are *purposefully designed to* appear human-like, to conceal their synthetic nature and computing capacities. Demanding transparency is thus antipodal to the engineers' efforts and the technological achievement of human-like AI, implying a conflict between regulatory and economic interests – a conflict in which the protection of fundamental rights should be watched particularly carefully.

We have argued that existing European legislation in principle does provide protection to data subjects regarding the processing of their personal data by intelligent conversational agents. While respective provisions are certainly of relevance in the context of potentially manipulative technologies, the particular concerns arising with humanoid EICAs such as inappropriate, anthropomorphism-triggered self-disclosure or people's growing inability to comprehend the synthetic nature and capacities of the computing systems surrounding them, are not addressed.

The identified limitation illustrates the currently changing nature of AI-powered (communication) products and suggests a lacking awareness thereof on the side of the legislator. EICAs are not experienced by consenting users nor are they restricted to the operation through actors with commercial interests. They can also be employed by individual people for their personal interests, resulting in a shift of implied (privacy) concerns to individuals that has not been considered by current privacy legislation and raising questions concerning the

desirable allocation of liabilities and responsibilities. This is not a matter of only data protection and privacy law, but also one of contract, consumer protection and liability law.

By definition, the setting of social interactions and relationships constitutes a core interest of the societies we live in – urging us to continuously consider the values we embed into those technologies that more and more casually enter our lives in the form of social actors. Further discussion should thus be opened on the extent to which we wish such integration to take place: besides pressuring for transparency and recipients' consent, should we regulate the sophistication of and the data that may be used for personalized human-machine interaction? Do we wish to prohibit systems that exploit individuals' (emotional) weaknesses and where do we draw the line between the design of a convenient user-experience and persons' intentional deception?

The trend towards more and more seamless human-machine interactions promises that those instances in which we consciously interact with, in which we are consciously aware of the presence of respective systems and able to prevent leaving behind a data trace by *simply being*, are likely to decline rapidly in the future.