2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

The subtle line between personalization and user manipulation in a European regulatory perspective. A proposal for a technology-assessment methodology for Artificial Intelligence Systems.*

Andrea Bertolini

I. Introduction

Much of HRI research focuses on personalizing robots in order to ease societal acceptance, and favour their uptake. Companion robots are indeed conceived as a potential solution to numerous societal concerns, among which aging population, and individuals' isolation. In such a perspective, personalization is indeed key, for it ensures individuals feel comfortable using robots in their daily lives and environments. This also depends upon the so called cultural competences the machine possesses. In fact, how humans behave largely depends upon their heritage, and overall understanding of the environment. An identical reaction, posture or expression might, indeed, be perceived very differently according to the culture of the person exposed to it.

At the same time, however, studies in these domains focus on how humans perceive robots to be different from any other object they interact with. Animacy is a particular criterion that summarizes the reasons why individuals consider robots as something peculiar, of its own nature; typically in-between an animal or living thing, and a mere object. Such conclusion often times rests upon the perception of – a higher degree of – autonomy, anthropomorphic appearance, and the display of behavior that appears to bear an emotional content with it.

While this is actively pursued for the above referred reasons of easing technological adoption, and therethrough addressing potential societal concerns, it raises the issue of the ethical viability and legal admissibility of said approaches. Indeed, those researches are based on an induced misperception of reality on the side of the user, who might be aware and informed of the artificial nature of the machine, and yet be prone to attributing it with those characteristics and emotions it actually does not objectively possess.

A debate arises among social scientists to draw much needed lines. The more fundamental question being whether ontology matters or not. While it is easily demonstrated that machines are not alive, do not perceive the existence of the individual before them, and their alterity, some question whether this should matter at all, and be of concern. To some, only perception matters, thence what humans feel whenever they are exposed to that interaction.

A second theoretical question entails the admissibility of deception. Indeed, those design techniques aimed at inducing humans to misperceive the artificial nature of the machine, and attribute it some degree of animacy may be deemed deceptive. However, many objections may be raised with respect to the ethical admissibility of deception, that are dependent upon different theoretical frameworks. The purpose of deception, the intention of the programmer, as well as the consequences and effects for the user are often brought into consideration to conclude that some forms of deception may be allowed, and only some may be deemed problematic.

A third perspective of analysis is, however, entirely legal, yet rests upon both technical, and ethical considerations, and requires an understanding of the psychological implications for the users of technological applications. Indeed, the legal system is not indifferent to manipulation. Some behaviors traditionally amount to fraud or undue influence, and thence are criminally relevant, some others are instead sufficient to avoid a contract, some more are fully licit. When it comes to Artificial Intelligence Systems (henceforth AIS) among which robots may be enumerated (as per the proposals of the European Commission on the regulation of AI, namely the so called AI Act, henceforth AIA)[1], the European Commission (henceforth EC) has demonstrated that the issue is taken seriously. Indeed, article 5 AIA identifies so called prohibited practices, and reserves two hypotheses to different forms of user manipulation, identifying the primary criterion for their unlawfulness in the use of subliminal techniques (let. A), or the circumstance that frail individuals are specifically targeted (let. B), respectively. Other principles may then apply, providing relevant guidance in specific cases, despite with a much more limited degree of ex ante predictability, considering the absence of a clearly defined norm [2]

While the existing regulatory framework is incomplete, and inadequate to encompass the variety of AIS that could give rise to different kinds of user manipulation – in fact AI as a single uniform class of applications does not exist [3] –, it certainly shows both the relevance of the matter, the sensitivity of European policy-makers for the topic, and the possibility that said concerns will have to be addressed early

*Resrach supported by ABC Foundation.

Andrea Bertolini, LL.M. (Yale), Ph.D., is Associate Professor of Private Law at the DIRPOLIS department of Sant'Anna School of Advanced Studies (SSSA) in Pisa, and director of the Centre of Excellence on the Regulation of Robotics and Artificial Intelligence EURA (www.eura.santannapisa.it), andrea.bertolini@santannapisa.it.

[1] 1. European Commission: Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts. COM/2021/206 final. European Commission (2021)

on, in the design phase of all devices – including and not limited to companion robots or social robots – so as to ensure that their deployment and use will subsequently be permitted.

The absence of a clearly defined regulatory framework certainly leaves a greater degree of uncertainty. However, on the one hand, rules and criteria may be identified that are applicable already today. On the other hand, the technical complexity of the subject matter – dependent upon its intrinsic multidisciplinary nature – suggests a bottom-up, technology-specific approach ought to be maintained, in particular in a regulatory perspective [3].

It seems thence both viable and advisable to begin today subjecting AIS, including social robots, to a specific technology-assessment focusing on potential risks of user manipulation. This is not to be intended as something hampering innovation or aimed at introducing unreasonable limitations to experimentation. Quite the contrary, it will provide necessary information both to researchers and policy-makers about what constraints they need to comply with, and when ad-hoc regulatory intervention is needed, respectively. This will also minimize the risk of the adoption of too-broad and vague all-encompassing norms that fail to provide much needed protection, and yet burden producers and smaller enterprises disproportionally.

Building upon the methodology developed with the H2020 project CONBOTS[2] for the ELSE assessment of robotic (AI-based) technologies, and further specifying it through the research conducted within the project PERSEO[3], the paper proposes a model for the technological-assessment of all potentially manipulative and deceptive AIS. The purpose of the assessment is multifold, namely: (i) help researchers understand the potential manipulative characteristics of the AIS they aim at developing, by identifying potential concerns and risks they might otherwise be unaware of; (ii) identify those elements that could be of legal relevance in excluding, limiting or conditioning the use of the AIS; (iii) provide information about the risks associated to the general public and (iv) data for policymakers to intervene with ad-hoc effective regulation when needed.

To further clarify all these aspects, the paper will first briefly discuss HRI research, highlighting aspects of potential concern in an ELSE perspective (section II); it will then stress the distinction between reality and appearance in the philosophical debate, that constitutes also the basis to then draw legal considerations (section III); it will then provide a brief overview of the regulatory framework (section IV). Based on all such considerations it will present a questionnaire each researcher should complete when developing an AIS with potential manipulative capacities (section V), and will then discuss some policy implications and suggestions (Conclusions).

[2] Home | CONBOTS

[3] PERSEO – European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 955778

II. HRI Research and Personalization

The purpose of this section is to highlight those aspects of HRI research and personalization that could be of relevance and potentially of concern in an ELSE perspective, and in particular with a focus on user manipulation.

The considerations drawn do not entail stating that all of HRI research is manipulative, nor that it should be prohibited. Much less it intends to maintain that the purpose of researchers working in HRI is that of manipulating users, and take advantage of them in any specific way.

Instead, it aims at undergoing a multidisciplinary analysis that helps – first and foremost roboticists – understand the broader implications of a complex and multifaceted phenomenon, which will most certainly play a role in our daily lives, and in society as a whole.

Such an effort will benefit HRI and social-robotics research in particular, by helping develop machines that are useful and agreeable, empowering humans with new capabilities, opportunities and potential[4], while maintaining a human-centric approach, thence fostering individual rights and fundamental rights above all. At the same time, and in a more practical perspective, it intends to ensure that the outputs of research are not later found to be problematic in a legal and regulatory perspective, by fostering a dialogue, already in the design phase, between social scientists – and lawyers in particular – and roboticists and AI experts.

A. HRI Research and animacy

The major concern in social robot design is possibly that of making sure humans perceive robots to be agreeable, not just useful, like an industrial robots, a cobot, a prosthetic limb or exoskeleton. This, indeed, is the peculiar trait that distinguish this very branch of robotics from all others [5].

Summarizing this entire field in a paragraph is impossible, and irrelevant a task. It shall suffice to stress that going beyond the minimal objective of avoiding the "uncanny valley" [6] [7], which some robotic applications may still elicit, research pursues animacy.

Animacy may be defined as the perception of that life-like behavior machines may display, that induces human to consider them to be different from any other object, and thence animate. It is often assessed together with "intelligence"[5] as another – independent – variable [11].

[4] These are the criteria identified by the European Commission in its policy statement on the European approach to artificial intelligence from April 2018, see 4. European Commission: Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. Artificial Intelligence for Europe. European Commission (2018)

[5] Intelligence is in between brackets because we could discuss how that ought to be defined, and at least distinguishing between general and special intelligence – on which 8. Floridi, L.: Philosophy and Computing: an Introduction. Routledge, London and New York (1999) –, but even more debating whether that may again be purely based on human perception and manipulation - 9. Turing, A.: Computing Machinery and Intelligence. Mind 49, 433-460 (1950) – or whether it requires some degree of self-perception and awareness - 10. Gutmann, M., Rathgeber, B., Syed, T.:

Animacy is thence very much based on perception, and makes that prevail over reality [12]. It is indeed hard, if not at all impossible, to question that machines are not alive, do not possess feelings, do not perceive the existence of the human being facing them, are not interested in their well-being, emotions and desires [13]. Yet they might display all of that [2], and – to some extent, at least – this appears to be necessary to make the interaction desirable for the human, thence agreeable.

It is often observed that the perception of animacy is dependent upon a number of factors including (i) the ability to move autonomously (or to appear to do so)[6]; (ii) the ability to answer in a fashion that is appropriate to the context; (iii) the ability to display emotions [12] .

At the same time, other elements may cause a robot to be perceived as living, that are very much rooted in both the human tendency to anthropomorphize, as well as their empathy, even towards what they perceive to be pain, and/or difficulty in completing a task on the side of the machine [14]. In such a perspective, technological limitations may prove a favorable element, easing the acceptance by certain kind of users, often frail ones.

B. The focus on frail individuals: children and the elderly

Indeed, much HRI research involves frail individuals, primarily children, and the elderly. The reasons for such a choice are mainly twofold: social robots are considered having relevant potential in the care of those very categories [15], and people with a more limited rationality – such as children and adults with some form of mental impairment or dementia – are more inclined to attributing animacy, in particular children under the age of seven [12].

While social robots may certainly prove beneficial in treating many conditions [16], the circumstance that those people are the primary target of research, as well as of the potential uses of developed technology certainly plays a relevant role in a regulatory perspective. *Ex ante*, when conceiving an experiment, more restrictions will inevitably apply, including with respect to informed consent and its acquisition. More relevantly, however, when assessing the legitimacy, and thence usability of a given technology or service, it will lead to the application of specific criteria and – in some cases at least – prohibitions, such as those put forth by the AIA (see §IV below).

Finally, it shall be stressed that the regulatory framework also encompasses broad and general principles – including fundamental rights – whose application is less automatic and necessitated in the possible outcome it achieves. The circumstance that frail individuals are involved will also become of relevance in the balancing of opposing rights, which is typical of those domains (see §IV below). To exemplify, it may be argued that "care" intended in its intrinsically relational dimension, may only be provided by humans, not machines. Thence, whenever the right to care – and not merely healthcare services – is ascertained, that might not be primarily proffered through robotics applications, for that might entail a violation of the fundamental rights of the individual entitled to receiving it [17].

III. ONTOLOGY AND APPEARANCE

One element that appears to be of central relevance in a regulatory perspective is the circumstance that robots are machines [10] [13] [18], thence inanimate objects, and do not possess those qualities they may pretend to have.

While some philosophical models of agency claim that they may in fact be deemed moral agents[7] [20, 21], a regulatory framework could only attribute personality – and subsequently responsibility – to a machine on the grounds of functional (e.g. law and economics) considerations [22]. Said otherwise, only when there are technical legal grounds – such as in the case of a corporation – the legal ordering might decide to attribute rights and obligations to inanimate entities such as robots and other technologically advanced applications. Those choices, however, would not lead to the attribution of personal or fundamental rights to the device, and thence would not alter the framing of their display of animacy as potentially manipulative for users.

Some researchers, instead, believe the contraposition between reality and appearance to be questionable [23] since they are not perceived as such by humans[24]. Moreover, humans also simulate, that way easing most relationships; ultimately, intentions matter, for they allow to differentiate those forms of manipulation that may be deemed morally wrong from those, instead, typically accepted [25] [26]. Many forms of deception are, in fact, merely white lies of little or no relevance, and if machines were capable of reproducing those kinds of behavior that would most certainly favor their uptake.

Truth is that the form of manipulation social robots might put into place is profoundly different from that of humans, and focus should be placed on the effect it produces on the observer. Indeed, it is primarily based on what people are led to think about and to feel for the machine they are interacting with. A social robot is designed and programmed so as to elicit something that is already an inner component of the human counterpart, more than to produce something new [27]. In such a perspective, intentions do not matter [28], also because machines do not possess them, and those of their creators and/or programmers are mostly irrelevant, unless openly mischievous. This approach to deception may be defined objective, for it is not based on the internal states of

Organic Computing: Metaphor or Model? Organic Computing—A Paradigm Shift for Complex Systems, pp. 111-125. Springer (2011) –.

[6] In fact, the robot may be remotely controlled, and thence not at all autonomous and yet provide the participat in the experiment the sensation that it is, indeed, capable of moving by itself, independently of any direct human control or supervision.

[7] This, however, presupposes deconstructing the traditional model of agency used in the philosophical and legal debate, based on self-awareness and intentionality – 19. Himma, K.E.: Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? Ethics and Information Technology 11, 19-29 (2009) –, to accept one whereby the circumstance that the effect produced by the apparent functioning of the device is of moral relevance (e.g. the harming of a huma being) suffices to establish agency.

the machine or of the human behind it, creating, using, or deploying them for business purposes, such as offering a good or service for profit.

IV. The Regulatory Framework

The applicable regulatory framework, including its most recent epiphany – the AIA –, appears coherent with such a theoretical backbone. While the subjective mental state plays a primary role in criminal law – for crimes are primarily pursued when scienter may be established or, residually, negligence, never on purely objective grounds – prohibitions – more broadly intended – and consequences in the civil and administrative domain, may well – and often are –justified on purely objective grounds, such as the effects produced.

For this very reason, we may in fact radically exclude the possibility of holding a machine criminally liable – unlike maintained by Hallevy [29] – for the intent is certainly lacking and, eventually, ought to be established upon the human agent who decided to use, create or deploy the specific application.

The ill-intent of the programmer, creator, user, deployer would indeed matter – in most cases – to establish his criminal responsibility. That, however, would amount to a distinct and additional sanction, certainly related to and dependent upon a – user or – technology that is deemed illicit, eventually because it displays manipulative or deceptive capacities. Yet, even in all those cases where such a criminal proclivity was absent, if the application met those objective requirements (see §A below) that caused it to be considered prohibited, its use would be banned.

It is therefore necessary to briefly consider those criteria that the European Commission identifies as relevant when discussing users manipulation via AIS, among which social robots may certainly be enumerated.

A. The AIA

The AIA was proposed by the European Commission the 21st April 2021 and approved in its final formulation on March 13th 2024 [1]. It represents the first attempt to regulate AI, being clearly more legally oriented than ethically based. In doing that, it adopts a risk-based approach which identifies three levels of risk – unacceptable, high, medium, or low – based on the possibility that the sector and the intended use of different classes of technologies may affect safety, psyche and fundamental rights of people involved.

Article 5 deals with devices that have to be considered prohibited by default and includes: (1)(a) those which use subliminal techniques able to manipulate the user, (1)(b) those which exploits the vulnerabilities of specific groups, (1)(c) those which have a general purpose of social scoring, (1)(d) real-time remote biometric identification systems which allow surveillance for law enforcement aims.

For the purposes of the analysis conducted here, only the first two classes – art. 5(1)(a) and art. 5(1)(b) – will be further commented upon.

First of all, it should be underlined that the word 'manipulation' identifies a broad concept, that can leave space of discretion if not further defined or circumscribed. This margin of ambiguity can be decisive for determining the possibility of circumventing, or not, the prohibition, making it practically ineffective [30].

On the contrary, the kind of vulnerabilities devices should take advantage of to be banned are explicitly mentioned and limited to: age, physical disability, and mental disability. This conception of human vulnerability appears reductive if we consider, for instance, the phenomenon of anthropomorphism, that certainly exposes users to the. Moreover, to be vulnerable – both physically and psychologically – is a characteristic inherent to human nature, that cannot be removed for it does not depend on external and objectively valuable conditions only [31]. This is the kind of 'fragility' that social robots and other possibly manipulative devices target. For this reason, the harmful effect does not simply impact certain categories of individuals, but could potentially interest every one of us.

Ambiguity rests on the conceptualization of the adjective "subliminal", while the individual's awareness of the artificial nature of the machine, as well as of its capacity to simulate emotions it does not possess, seems to have little impact on the ability of the device to manipulate him, and subsequently on the need for regulatory intervention. As far as social robotics is concerned, it has been demonstrated that the design of the machine can play a significant role in the perception users have of the device and in the emotional bond they establish. However, the physical appearance of the robot is an explicit, not covert, detail, as much as its artificial nature is immediately evident to – at least – any rational person. Nevertheless, this does not per se prevent the occurrence of damaging consequences, in particular when humans raise their expectations about the real capacities of a system and rely upon them to deal with an essential aspect of their lives[8] [28].

One of the merits of the AIA is certainly the textual reference to psychological damages that new technologies may cause, among which – one could argue – a progressive decrease in the ability to socialize with other humans, replaced by human-machine interactions may be counted. In those instances where machines are perceived to be a preferable form of interaction [32], because they affirm or anyway do not challenge the user – unlike another human possessing own preferences, will, character and opinions –, the temptation to replace a human counterpart with an artificial one would be real. If this occurred repeatedly, thanks to an effective approach to technological personalization, humans might lose an ability to naturally socialize and interact in a complex environment as that we live in and feel reassured in the artificial cocoon, technology could create.

However, the concrete application of this provision appears more complex and difficult to achieve than what the European Commission seems to have considered. Firstly,

[8]

since the prohibition is made dependent upon the repercussions on the individual psychological dimension, difficulties will emerge in observing and demonstrating such form of harm in an objective fashion and in an early stage of technological deployment [33]. That could be easier in the event of serious distortions of the decision-making process or of proper coercion. Nonetheless, even such cases where the manipulative effect does not necessarily end dramatically ,should not be underestimated – or excluded from an appropriate regulation [34]. Moreover, it is frequent that some forms of manipulation can barely appear as such – even for those who suffer them – if analysed as isolated and circumstantial manifestations. The actual impact is better appreciated if we consider manipulative practices in an aggregate form, as 'dynamics' that manifest themselves over time, rather than as 'instantaneous events'. It follows that the causal link between a device's manipulative methodology and a negative consequence on the user's psychological dimension could be missed or not easily reconstructed. This should not be, in and by itself, sufficient to deny the danger of the AI system considered.

Overall, while the formulation advanced by the AIA needs to be deemed partial, incomplete and overall unsatisfactory for the reasons very briefly sketched, it confirms both the relevance of the perspective here maintained, as well as demonstrates the need for an articulate regulatory intervention, balancing opposing interests and need in a sufficiently narrow-tailored fashion. Such an awareness clearly emerges from Recital 15 AIA, whereby it is stated that, despite other possibly beneficial outcomes – or beneficial intents – the regulation has the primary aim to reaffirm "the Union values of the respect for human dignity […] and Union fundamental rights", which represents the pillars of the internationally shared purpose of protecting human beings.

B. Human dignity as a limit to individual choices, even with respect to manipulation

Indeed, the principle of human dignity represent an objective and external criterion, capable of directing the regulation of new technologies, including those here considered. Being a fundamental right and a principle, it reconciles the flexibility of ethical instances with the binding and non-arbitrary nature of legal norms [35].

This is because human dignity has already a legal connotation. It is the foundation of modern Constitutions and Charters of Rights all around the world – including the European Charter of Human Rights – and it is on the basis of preeminent judgments in the Courts of Justice [36]. As such, it is an inviolable principle, inherent in each individual by the mere fact of belonging to humanity.

For the same reason, it can represent an external limit to other competing rights, including the right of self-determination. Therefore, a practice – even if supported by personal consent – can be limited or impeded through the principle of human dignity, if it is considered detrimental or able to diminish what is proper and inborn of human beings [37] [38]. So understood, such a principle can allow to go beyond (i) self-interest – with this intending the one of private individuals, who may ignore possible repercussion of the use of a manipulative devices –, but even (ii) economic interests of manufacturers – who merely aim to bring their products or invention to the market [2].

Despite lacking a precise and universally recognized definition [39], its theoretical interpretation and judicial application allow us to draw some lines. In fact, it has already been proven to be effective in the protection of the core essence of humanity from many cases of severe infringement, such as totalitarianisms, and discrimination [40], marginalization and exploitation of minorities or human vulnerabilities; without it we would not even be able to explain what is wrong with slavery [41].

In the cases here considered, if a social robot was deemed to be impinging upon human dignity, that could suffice to prevent its use and application. A number of factors could contribute to determine such a conclusion, including those recalled above, namely the circumstance that it causes humans to isolate and prefer artificial interactions to real ones, eventually leading completely delusional existences. Similarly, if an elderly person was being cared for entirely or primarily through robotic applications, that could be deemed violating both his dignity and his right to care, which is separate and distinct from that to cure [42].

Not even utility, or a perceived – or even measured – beneficial nature of the device would per se suffice in trumping a judgment of violation of the user's dignity, as it is typical of fundamental rights, which always require to be balanced one against the other.

So interpreted, human dignity can be an essential and efficient balancing tool to both (i) guarantee the – physical and psychological – integrity of human beings and (ii) orient technological development towards applications which allow the promotion of fundamental rights and values.

Robots are gradually becoming part of our private life, to perform tasks side by side – or even in place – of human beings. As a consequence, HRI is becoming a central research topic, for an effective interaction is essential to allow both the diffusion of the technology in question and user satisfaction.

Most certainly the principle of human dignity being so broad, and flexible in its applications allows for a great degree of *ex ante* uncertainty, in particular for researchers and developers of technology, as well as *ex post* difformity in applications among courts and across countries, even simply European Member States.

Therefore, absent more detailed regulation, given the loopholes of the AIA, its role could prove crucial in this domain. However, it would be certainly preferable for the legislator to act in a more granular and detailed fashion, along those very lines they have already identified, so as to – attempt to effectively – govern risks of user manipulation via AIS, and social robots.

To this end, the proposed methodology ought to provide guidance already today, in the earliest stages of technological design and development, so as to consider those instances,

address them and possibly solve them. In all cases, the proposed methodology provides an essential mapping of associated risks, and increasing sensitivity and understanding.

V. A Proposal for a Methodology for Technological Assessment

The fast pace of technological development challenges regulation. However, legal systems are typically elastic and may accommodate evolution in particular through general principles and broad norms. In such a perspective, proceduralization is particularly useful, for it ensures sufficient guidance to those that are required to comply with it, and helps achieve greater degrees of transparency.

One way of proceduralizing assessments is through questionnaires, leading to the drafting of reports, sufficiently detailed and to the point as to avoid information overload.

A. The ELSE assessment of technologies

Different methodologies are today emerging for the assessment of the ELSE implications of AIS, including robotic devices.

The outcome of the EC's High Level Expert Group on AI, namely the Ethic Guidelines on a Trustworthy AI, was in fact concretized in a dedicated questionnaire. Other expert groups have adopted similar approaches, and technological applications developed within EU-funded projects have already been assessed pursuant to similar criteria, with respect to all their potential ELSE implications (e.g. also user isolation and potential dehumanization), drafting relevant reports.

Indeed, questionnaires have the merit of articulate very complex legal and ethical concerns in a series of detailed points and corresponding questions. Those, when not self-explanatory, ought better be presented by experts. Said experts should prepare tables with clear indications of both the underlaying legal principle, its scope of application, while providing an exemplifying case to clarify the possible concrete application of the principle itself.

This enables non-experts (primarily engineers developing a specific application) to acquire knowledge and understanding of the essential concerns identified by ELSE specialists, including lawyers, and provide as accurate an answer as possible.

Said questionnaires ought to be prepared and circulated in advance, to allow technicians to gain a preliminary understanding, before discussing the questions with the interviewer, namely the ELSE specialist himself. Only after allowing some time for engineers and developers to familiarize with the questions and start thinking about the different aspects they touch upon, ought the interview take place. Then the ELSE specialist will go through each question, and annotate individual answers, clarifying the context and all the elements provided.

Based on said information, a detailed, yet synthetic and analytical report may be compiled, allowing for an overall assessment of the relevant risks. This will also allow for the identification of potential strategies to minimize and manage said concerns in the design phase, and eventually provide guidance for policy-making and regulatory initiatives.

Such a methodology was successfully employed by the research unit EURA within Scuola Superiore Sant'Anna, participating in the Horizon 2020 project CONBOTS.

B. A questionnaire for user manipulation concerns

Given the specific concerns here considered, a series of detailed questions is presented below that, when accurately answered, would allow for an in-depth assessment of the potential manipulation of the user through a specific AIS, including social robots.

The questionnaires to be presented pursuant to the methodology described in the preceding paragraph, should then lead to the drafting of a "Potential User Manipulation Assessment Report", addressing the following points and questions:

	Question	Explanation	Example
1.	Does the AIS target or deal with a particular kind of users? Which one? Are those individuals frail (e.g.: children, elderly, people with disability)?	It is important to establish whether the system is conceived to target one specific kind of users and possibly their characteristics. Frail individuals are typically more protected than full-fledged adult human beings. Interactions with them might be subject to specific legal requirements depending upon the different kind of applications. The purpose of this question is to clarify this aspect.	An AIS might be conceived to entertain senior users, eventually users affected by a condition (e.g. dementia), or even very young ones. Examples could come from robotics used in the care of the elderly.
2.	Does the AIS model the user or trace his behaviour and choices? How? What data does it collect, process or have access to?	The purpose is to establish whether the AIS is conceived to create a model of the agent so as to predict his/her behaviour, and determine what data it collects. The purpose is to establish whether it falls under the provisions of the General Data Privacy Regulation (GDPR)	Examples could be derived from marketing. If the data and information collected by the AIS are used to create a model of the real agent. It is essential to understand what information is collected and how it is treated and stored.

3.	What is the intended purpose of the AIS?	All AIS will be designed with an intended purpose. There is a need to identify and describe it clearly for purpose of transparency and also for the assessment of the potential manipulative behaviour.	The purpose of an AIS could be that of entertainment, or also to provide company, and/or emotional support. It could be intended to provide adult services or to operate as a sales generator.
4.	Is the AIS conceived to predict the behaviour of the user, and her choices?	Among the possible uses and purposes of an AIS that of predicting the user's behaviour and choices displays a particularly relevant manipulative potential, and needs to be addressed with specific attention.	Examples could be derived from marketing. If the information acquired as per *sub* 2 above, is used to model the agent and influence their behaviour (e.g. induce them to make a purchase), that capacity needs to be attentively described to determine if it satisfies all relevant legal requirements.
5.	Does the AIS simulate, pretend to have, or anyway display animacy, and/or emotions, and/or empathy, and/or attachment towards the user and/or other human beings? How?	The characteristics of the AIS that might have a particularly relevant manipulative capacity need to be precisely identified and described.	The robot or AIS might appear to possess feelings for the user due to the shape of its eyes, facial expressions and possibility to mimic human emotions. All such characteristics and capacities need to be attentively described.
6.	May the AIS induce through its functioning the human being to develop an emotional attachment, and/or does the AIS present itself as a friend, and/or carer, and/or partner, and/or romantic partner? Is it intended to do so? Is it its primary purpose? Is it a secondary purpose intended to facilitate another task, and/or the acceptance of the user? What other purpose does it serve?	Social robots or robot companions, as well as some software agents might be designed to induce an emotional attachment by the user, exploiting potential vulnerabilities, as well as the natural human propensity to anthropomorphize objects. The reasons whereby such a form of emotional manipulation is deemed necessary and/or beneficial ought to be clarified.	AIS already existing on the market today are advertised as potential sentimental partners of the user, or anyhow conceived to provide companionship, emotional and psychological support. The characteristics of those interactions ought to be clarified, as well as the potential of the single application in inducing an erroneous representation of reality by the user.
7.	Does the AIS allow for the creation of alternative worlds, and/or settings, and/or realities, including virtual and/or augmented reality? Is the AIS used to, and/or does it allow for, and/or does it facilitate the interaction with other non-human entities, and/or avatars, and/or representations of – even no-longer living – human beings? Does it induce, and/or allow the user to think that the interaction is real? Is that the primary purpose of the AIS? If not, what other purpose does it serve?	The purpose is to understand if the AIS induces the user to interact with a non-real world and the nature of such interactions. Indeed, some forms might fall under existing regulation both in the field of law and technology, as well as contract law and criminal law (e.g. circumvention of the incapacitated).	In this case the focus is on AIS used in the context of virtual and/or augmented reality. Examples could be that of AIS used to replicate the behaviour of deceased loved ones.

The purpose of such a report is to uncover potential risks of user manipulation, as well as the technological characteristics that might induce it. At the same time, it is essential to clarify the intention of the programmer and/or deployer of the system. All such aspects allow for the identification of potentially relevant legal constraints and limitations already applicable. At the same time, they ensure the possibility of achieving greater transparency about the very functioning of the AIS. Such information could then be summarized though a colour-coding mechanisms, that simply conveys the information to the final users about the manipulative potential of a specific AIS.

VI. Conclusion

In a policy perspective, it is advisable that a technology assessment based on the questionnaire here proposed, and based on the considerations drawn was required for all European-funded research in the field of AIS, including social robots.

Such an assessment would serve the purpose of focusing the attention of researchers on the ELSE and regulatory implications of their devices, would provide relevant information for users to whom those devices are catered, and would – in an aggregate fashion – provide sufficient data for narrow-tailored regulation, protecting human rights without impairing technological development.

REFERENCES

Bibliography

1. European Commission: Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts. COM/2021/206 final. European Commission (2021)
2. Bertolini, A.: Human-Robot Interaction and Deception. Osservatorio del diritto civile e commerciale, Rivista semestrale 645-659 (2018)
3. Bertolini, A.: Artificial Intelligence does not exist! defying the technology- neutrality narrative in the regulation of civil liability for advanced technologies. Europa dir. priv 369-420 (2022)
4. European Commission: Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. Artificial Intelligence for Europe. European Commission (2018)
5. Ishiguro, H.: How Human Is Human? The View from Robotics Research. JPIC International (2020)
6. Mori, M.: Bukimi no tani (the uncanny valley). Energy 7, 33-35 (1970)
7. Mori, M.: The Uncanny Valley. IEEE Robotics & Automation Magazine 19, 98-100 (2012)
8. Floridi, L.: Philosophy and Computing: an Introduction. Routledge, London and New York (1999)
9. Turing, A.: Computing Machinery and Intelligence. Mind 49, 433-460 (1950)
10.Gutmann, M., Rathgeber, B., Syed, T.: Organic Computing: Metaphor or Model? Organic Computing—A Paradigm Shift for Complex Systems, pp. 111-125. Springer (2011)
11.Bartneck, C., Kanda, T., Mubin, O., Al Mahmud, A.: Does the Design of a Robot Influence Its Animacy and Perceived Intelligence? International Journal of Social Robotics 1, 195-204 (2009)
12.Cameron, D., Fernando, S., Collins, E.C., Millings, A., Szollosy, M., Moore, R., Sharkey, A., Prescott, T.J.: You made him be alive: Children's perceptions of animacy in humanoid robot. In: Mangan, M., Cutkosky, M., Mura, A., Verschure, P.F.M.J., Prescott, T., Lepora, N. (eds.) Biomimetic and Biohybrid Systems. 6th International Conference, Living Machines 2017, Stanford, CA, USA, July 26–28, 2017, Proceedings. Living Machines 2017, vol. Lecture Notes in Computer Science, vol 10384, pp. 73-85. Springer, Stanford University, California (2017)
13.Bertolini, A.: Robots as Products: The Case for a Realistic Analysis of Robotic Applications and Liability Rules. Law Innovation and Technology 5, 214–247 (2013)
14.Di Napoli, C., Ercolano, G., Rossi, S.: Personalized home-care support for the elderly: a field experience with a social robot at home. User Modeling and User-Adapted Interaction (2022)
15.Turkle, S., Taggart, W., Kidd, C.D., Dasté, O.: Relational artifacts with children and elders: The complexity of cybercompanionship. Connection Science 18, 347-361 (2006)
16.Aminuddin, R., Sharkey, A., Levita, L.: Interaction With the Paro Robot May Reduce Psychophysiological Stress Responses. 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Christchurch, New Zealand (2016)
17.Bertolini, A., Arian, S.: Automated Care-Taking and the Constitutional Rights of the Patient in an Aging Population. In: Custers, B., Fosch-Villaronga, E. (eds.) Law and Artificial Intelligence: Regulating AI and Applying AI in Legal Practice, pp. 297-321. T.M.C. Asser Press, The Hague (2022)
18.Bryson, J.J.: Robots should be slaves. Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues 8, 63-74 (2010)
19.Himma, K.E.: Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? Ethics and Information Technology 11, 19-29 (2009)
20.Floridi, L., Sanders, J.W.: On the Morality of Artificial Agents. Minds and Machine 14, 349–379 (2004)
21.Floridi, L.: Artificial Agents and Their Moral Nature. In: Kroes, P., Verbeek, P.-P. (eds.) The Moral Status of Technical Artefacts, pp. 185-212. Springer Netherlands, Dordrecht (2014)
22.Bertolini, A., Episcopo, F.: Robots and AI as Legal Subjects? Disentangling the Ontological and Functional Perspective. Frontiers in Robotics and AI 9, (2022)
23.Coeckelbergh, M.: Robot rights? Towards a social-relational justification of moral consideration. Ethics and Information Technology 12, 209-221 (2010)
24.Prescott, T.J.: Robots are not just tools. Connection Science 29, 142-149 (2017)
25.Coeckelbergh, M.: Moral appearances: emotions, robots, and human morality. Ethics and Information Technology 12, 235-241 (2010)
26.Isaac, A., Bridewell, W.: While lies on silver tongues. Why robots need to deceive (and how). Robot ethics 2, 157-172 (2017)
27.Turkle, S.: Alone together: Why we expect more from technology and less from each other. Basic Books, New York, NY, US (2011)
28.Sharkey, A., Sharkey, N.: We need to talk about deception in social robotics! Ethics and Information Technology 23, 309-316 (2021)
29.Hallevy, G.: Virtual Criminal Responsibility. Original Law Review 6, (2010)
30.Sax, M.: Between empowerment and manipulation: The ethics and regulation of for-profit health apps. Kluwer Law International BV (2021)
31.Coeckelbergh, M.: Artificial companions: Empathy and vulnerability mirroring in human-robot relations. Studies in ethics, law, and technology 4, (2011)
32.Recchiuto, C., Sgorbissa, A.: Diversity-aware social robots meet people: beyond context-aware embodied AI. arXiv preprint arXiv:2207.05372 (2022)
33.Gandy, O.H.: Coming to terms with chance: Engaging rational discrimination and cumulative disadvantage. Routledge (2016)
34.Bertolini, A., Carli, R.: Human-Robot Interaction and User Manipulation. In: Persuasive Technology, pp. 43-57. Springer International Publishing, (Year)
35.Harris, I., Jennings, R.C., Pullinger, D., Rogerson, S., Duquenoy, P.: Ethical assessment of new technologies: a meta-methodology. Journal of Information, Communication and Ethics in Society (2011)
36.O'Mahony, C.: There is no such thing as a right to dignity. International Journal of Constitutional Law 10, 551-574 (2012)
37.Dreier, H.: Die „guten Sitten "zwischen Normativität und Faktizität. Gedächtnisschrift für Theo Mayer-Maly, pp. 141-158. Springer (2011)
38.Gros, M.: Il principio di precauzione dinnanzi al giudice amministrativo francese. Il principio di precauzione dinnanzi al giudice amministrativo francese 709-758 (2013)
39.Fabre-Magnan, M.: La dignité en droit: un axiome. Revue interdisciplinaire d'études juridiques 58, 1-30 (2007)
40.Kretzmer, D., Klein, E.: The concept of human dignity in human rights discourse. Kluwer Law International The Hague (2002)
41.Kolakowski, L.: What is left of Socialism. First Things: A Monthly Journal of Religion and Public Life 42-47 (2002)
42.Bertolini, A., Arian, S.: Do robots care? Towards an Anthropocentric Framework in the Caring of Frail Individuals Through Assistive Technology. In: Haltaufderheide, J., Hovemann, J., Vollmann, J. (eds.) Aging between Participation and Simulation. Ethical Dimensions of Socially Assistive Technologies in Elderly Care, pp. 35-52. De Gruyter, Berlin, GE (2020)