Computer Ethics - Philosophical Enquiry (CEPE) Proceedings

Volume 2019 *CEPE 2019: Risk & Cybersecurity* Article 12

5-29-2019

Confucian Robot Ethics

Qin Zhu
Colorado School of Mines

Tom Williams
Colorado School of Mines

Ruchen Wen
Colorado School of Mines

Confucian robot ethics

Qin Zhu
Colorado School of Mines

Tom Williams
Colorado School of Mines

Ruchen Wen
Colorado School of Mines

Abstract

In the literature of artificial moral agents (AMAs), most work is influenced by either deontological or utilitarian frameworks. It has also been widely acknowledged that these Western "rule-based" ethical theories have encountered both philosophical and computing challenges. To tackle these challenges, this paper explores a non-Western, role-based, Confucian approach to robot ethics. In this paper, we start by providing a short introduction to some theoretical fundamentals of Confucian ethics. Then, we discuss some very preliminary ideas for constructing a Confucian approach to robot ethics. Lastly, we briefly share a couple of empirical studies our research group has recently conducted that aimed to incorporate insights from Confucian ethics into the design of morally competent robots. Inspired by Confucian ethics, this paper argues that to design morally competent robots is to create not only reliable and efficient human-robot interaction, but also a robot-mediated environment in which human teammates can grow their own virtues.

Keywords: *Confucian ethics, Applied ethics, Robot ethics, Artificial moral agents, Comparative studies*

In the literature of artificial moral agents (AMAs), most work is influenced by either deontological or utilitarian frameworks (Vallor, 2018). It has also been widely acknowledged that these Western "rule-based" ethical theories have encountered both philosophical and computing challenges. Most notably, these frameworks often struggle to "accommodate the constant flux, contextual variety, and increasingly opaque horizon of emerging technologies and their applications" (Vallor, 2018, p. 209). To tackle these challenges, this paper explores a non-Western, role-based, Confucian approach to robot ethics. In contrast to the Western philosophical approaches to robot ethics (or ethics in general) that focus on "defin[ing] what the good is" and worry about "how one can come to know the good," Chinese philosophers represented by Confucian scholars are more interested in the problem of "how to become good" (Ivahoe, 2000, p. ix). Inspired by Confucian ethics, this paper argues that to design morally competent robots is to create not only reliable and efficient human-robot interaction, but also a robot-mediated environment in which human teammates can grow their own virtues.

We argue that exploring a Confucian approach to robot ethics has critical values in a variety of ways. For instance, philosophers have recently argued that Confucian ethics can provide other alternative ways of thinking about ethical issues associated with robotics. As argued by Pak-Hang Wong, "if the philosophy of AI and robotics only comes from the West, that won't be enough, because it won't always apply to non-Western countries… And you miss opportunities to think in different ways about technology" (Cassauwers, 2019). In other words, Confucian ethics along with other non-Western ethical resources can enrich the moral imagination of roboticists and enhance their capabilities to define and engage ethical issues in designing robotics from culturally diverse perspectives.

Another crucial value for studying Confucian robot ethics is concerned with the powerful role that China and other cultures with the Confucian heritage (CHCs) such as Japan, Korea, and Singapore assume in the global market and the global robotics community. Understanding Confucian ethics is critical for understanding how policymakers, industrial entrepreneurs, scientists, and the public in CHCs view and accompany robotics.

In this paper, we start by providing a short introduction to some theoretical fundamentals of Confucian ethics. Then, we discuss some very preliminary ideas for constructing a Confucian approach to robot ethics. Lastly, we briefly share a couple of empirical studies our research group has recently conducted that aimed to incorporate insights from Confucian ethics into the design of morally competent robots.

Confucian ethics: Theoretical fundamentals

In recent decades, philosophers have employed various approaches to engaging Confucian ethics ranging from overtly historical or textual approaches to comparative approaches that put ideas from the classical period into conversation with contemporary Western ethical, social, and scientific theories (Mattice, 2019). Scholars have tried to understand Confucian ethics as a species of deontology, virtue ethics, or care ethics (Mattice, 2019). Until very recently, scholars have attempted to theorize Confucian ethics as a kind of role-based moral theory (Ames, 2011; Rosemont & Ames, 2016). The role-based approach to Confucian ethics is one of most recent efforts to reinterpret and rediscover the value of Confucian ethics. Confucian role ethics argues that we as humans were born into a web of social relationships. These social relationships have normative implications and they prescribe specific moral responsibilities for us in the communities we belong to. Ames (2011) argues that the term person itself is *relational* and *social* (e.g., it is hard to call a human "person" if she is the only human in the world). For instance, the father-daughter relationship does not only have descriptive values (e.g., describing an objective relationship between me and my daughter) but also has normative implications (e.g., what a healthy father-daughter relationship looks like, what expectations about being a good father are, what I should do to live my role as a father). Therefore, the Confucian ultimate goal of becoming a good person depends on to what extent we live well our social roles and practice well the moral responsibilities prescribed by these social roles.

Relationships, contexts, and social roles are crucial in Chinese philosophy especially Confucian ethics. The cultivation of the moral self including various virtues such as the principal virtue *ren* (仁, benevolence, goodness, or humaneness) takes place in the development of relationships (Lai, 2017). A person seldom grows her virtues at home by herself through her own individual actions and reflections. Instead, reliable virtues are required to be cultivated, tested, and enhanced in her *interactions* with others in specific contexts while living her communal roles. As a father, I learn virtues that define a good father not from reading books but from interactions with my daughter. The term father is coexistent with the term daughter. My daughter provides me with opportunities to develop the virtues that are required for being a good father. Cultivating virtues is thus a project that is engaged in concert with others.

Compared to most Western ethical approaches that focus on moral reasoning and justification, Confucian ethics places more emphasis on moral practice and practical wisdom. What is central to Confucian ethics is the moral development model that consists of three interrelated components: observation, reflection, and practice (Zhu, 2018). In other words, one must carefully observe how people in the society interact with others and reflect on whether and how their daily interactions with others are in accordance with *li* (礼, rituals or ritual propriety). The appropriate practice of rituals manifests virtues, whereas virtues underlie and guide the practice of these rituals. Then, one needs to incorporate her reflective learning experience into her own future interactions with others and test to what extent she has grasped the appropriate practice of rituals and their underlying moral virtues.

For Confucians, *li* or rituals are crucial for ethics practice and they are the social norms that are rooted in historical traditions and have been widely recognized as morally accepted norms in specific cultural contexts (Lai, 2003). Therefore, virtues (e.g., *ren*) and *li* are independent of each other (Lai, 2017; Shun, 1993). To demonstrate that one understands well a virtue such as *ren*, one needs to express it through appropriate ritual practice in her interactions with others. Conversely, her manner of interaction with others indicates her grasp of the virtue of *ren*. In this sense, Confucian criteria for evaluating one's moral development are *social* as they depend on to what extent people comprehend and appropriately practice rituals in their social interactions. In other words, someone who fails to or is unwilling to reflectively practice rituals misses opportunities for moral growth and thus is not a responsible person.

If the ultimate goal for Confucians is that one is *always* striving to become a good person through reflective learning in social interactions, then the question becomes whether everyone has the equal opportunity to achieve such goal. As a follower of Confucius, Mencius advocated for a kind of moral egalitarianism and believed that all human beings have the equal potential to become good if they devote themselves to intentional moral efforts (Munro, 1969, p. 15). To Confucians, what characterizes the personhood is not so much about one's innate and inalienable individual human rights as most Western political and ethical theories would emphasize. Instead, Confucians think that it is one's intentional efforts to strive for a good person that defines her personhood. Simply eating and sleeping without much reflective thinking will not make someone a *true* person (at least it is not a kind of person whose life is worth living).

Toward a Confucian robot ethics

A recent essay published in *OZY* by Tom Cassauwers (2019) aimed to reexamine ethical issues associated with robotics from two Eastern schools of thought: Confucianism and Buddhism. This essay includes interviews he conducted with multiple Eastern philosophers including two Confucian scholars Pak-Hang Wong and Heup Young Kim. An important theme in Confucian ethics of technology acknowledged by both Wong and Kim is that technology is never value neutral and it has crucial *instrumental* value that helps people acquire virtues such as *ren* and cultivate the moral self (Cassauwers, 2019). For instance, Wong mentioned an example in a recent redesign of Amazon's virtual assistant Alexa: designers have developed a new feature "politeness feature" for Alexa which will make Alexa only respond to people who speak to Alexa politely. Wong argues that such minor design could be made by a Confucian (Cassauwers, 2019).

Nevertheless, arguably, there remains a question whether these Amazon designers were *actually* inspired by Confucian ethical theories or more specifically Confucian moral psychology. Such question is worth exploring as it is relevant to the argument discussed earlier that relationships and contexts are important for Confucian ethics. The effectiveness of Alexa's politeness feature may be dependent on the specific role Alexa plays in a context and the relationship between Alexa and the human interlocutor. I as the father refuse to respond to my daughter's impolite request might have different moral impacts on her than if a stranger does the same reaction. Philosophers of technology such as Peter-Paul Verbeek would agree that design engineers do have the obligation to imagine the potential relationship that will be constructed between technology and its user and how such relationship affects the moral perception and behavior of the user (Verbeek, 2006).

What people are often overlooking is the *relational* nature embedded in the design of most robots especially social robots which are being integrated into our society. When robots are being designed, certain relationships are imagined, defined, and assigned to those robots and these relationships are often determined by the use contexts of robots and the specific roles robots are expected to assume in these contexts (e.g., healthcare robots). Dumouchel and Damiano (2017) recently argue that social robots such as Geminoid and Paro can only truly interact with other agents, and not with objects. Unlike humans, these robots have no relation to the world but to their human partners. These robots were mainly created for the interaction or relationship with human partners. It is the interaction or relationship between robots and their human partners that makes the *existence* of these robots. In this sense, we suggest that roboticists should not only leverage the traditional, dominant approaches to developing AMAs that focus on integrating rule-based morality, but also consider an alternative approach to designing morally competent robots based on the *role responsibilities* prescribed by the relationships robots have with human teammates in specific use contexts.

Philosopher JeeLoo Liu (2017) constructed three principles for Confucian robotic ethics that are based on the role moralities of robots:

[CR1] A robot must first and foremost fulfill its assigned role.

[CR2] A robot should not act in ways that would afflict the highest displeasure or the lowest preference onto other human beings, when other options are available.

[CR3] A robot must render assistance to other human beings in their pursuit of moral improvement, unless doing so would violate [CR1] and [CR2]. A robot must also refuse assistance to other human beings when their projects would bring out their evil qualities or produce immorality.

Liu's three Confucian robotic ethical principles well integrate major elements of Confucian ethics we introduced in the last section. A moral competent robot is expected to be able to fulfill its assigned social roles. Such fulfillment of social roles for this robot is not isolated from but in concert with other humans. For humans, self-cultivation is not an individual but a social project which the robot can contribute to. In other words, the interaction between the robot and its human teammate is indeed crucial for the human's pursuit of moral development. A socially integrated robot is expected to be sensitive to the norms shared within human communities and contribute to the evolution of these norms.

Confucius would probably argue that a social robot who is not capable of rendering assistance to humans in their pursuit of moral improvement is not worth being a companion for humans. Such judgment of the moral quality of the robot is comparable to Confucius's thesis that moral cultivation is essential to friendship. Interestingly, philosophers David Hall and Roger Ames (1994) argue that Confucian friendship is hierarchical despite that the hierarchy in friendship is different from that in other four Confucian relationships such as the father-son relationship. The hierarchy that exists in friendship recognizes that difference exists in the level of moral excellence between oneself and her friend (Lu, 2010). That is partly why Confucius said "do not accept as a friend one who is not your equal" (*Analects*, 1.8). Here, friendship as a relationship does have *instrumental* value, that is, a good or worthwhile friendship often provides opportunities and resources for the cultivation of the moral self. If we treat friendship as a paradigmatic case for the relationship between most social robots and their human teammates, then shouldn't we always strive to find robots who are capable of making us better people? Social robots may be distinguished in terms of their different capabilities of completing tasks. However, we argue that they should also be distinguished by their different capabilities of exerting positive impacts on the moral development of their human teammates.

The emerging literature in responsible innovation suggests that design is a "far richer process" as it realizes both functional requirements and moral values." Designers not only "can provide us with technical means but also can address the values of people and society and think about expressing them in material culture and technology" (van den Hoven, Vermaas, & van de Poel, 2015, p. 3). Therefore, roboticists should not simply consider robots as efficient means to help human users complete tasks. Moreover, for designing a social robot, roboticists need to consider other morally relevant issues such as:

- What social role is such robot expected to assume in its use context?

- What are characteristics or "traits" of this robot that defines it as a morally competent or "good" robot? How does the assigned role of the robot prescribe or specify these characteristics or "traits"?
- What kind of person is the human teammate becoming through her everyday interaction with the robot?

Social robots can be considered as companions with whom humans spend a lot of time. Nevertheless, does that mean a truly socially integrated robot has always to be polite and please its human teammate, even when the human teammate proposes morally questionable requests? Alistair M. C. Isaac and Will Bridewell (2017) point out that "standing norms" (baseline rules for effective human conversations such as "being polite or informative") are important for robots to be truly socially integrated and effectively communicate with humans. However, they also note that in meaningful conversations *ulterior motives* often are more fundamental to and thus supersede these standing norms. In other words, when designing strategies robots employ to respond to human requests, roboticist need to consider the ulterior motives (e.g., being a good companion) that are communicated through robot responses to human requests, in addition to the standing norms that are expected to be followed by robots and human teammates.

Therefore, is it okay for robots to blame morally questionable requests proposed by their human teammates as a person would do to another person (e.g., a friend)? From the Confucian perspective, one may argue that the way in which the robot responds to the human request is highly *contextual* and it depends on a variety of factors such as: under what context the human teammate asks the request, what role this robot plays, what relationship the robot and the human teammate has established, the level of ethical sensitivity the human teammate exhibits, and how much the robot "knows" the nature or personality of its human teammate. It is also worth noting that the relationship between the robot and its human teammate may also change as they interact with each other on the daily basis. For instance, if the relationship between a robot and its human teammate is reliable and trustworthy which is comparable to friendship in the Confucian sense, then it might be justifiable that the robot should be able to remonstrate with or blame its human teammate. Arguably, the role responsibility of the robot prescribed by its relationship with the human teammate encourages the robot to be responsible for the moral development of *the other one* who also contributes to such relationship. This argument is supported by Liu's third Confucian robotic ethical principle we mentioned earlier.

To some extent, despite the importance of role responsibility in Confucian ethics, Western philosophers hold different views on the connection between social roles and relationships and autonomous moral agency. To Western philosophers such as Dumouchel and Damiano (2017), the social roles and relationships assigned by roboticists to robots make robots less independent and thus have less moral agency which is fundamental for most Western political philosophical concepts such as liberty and autonomy. As discussed earlier, social robots are often *designed* to work in specific circumstances and serve certain purposes for humans. They are not independent and do not have or pursue their own goals. Dumouchel and Damiano (2017) argue that only robots with no explicit purpose may have autonomous moral agency comparable to

human personhood. Not having an explicit and predetermined purpose indicates that these robots are *free* to do anything they want.

In contrast, Asian philosophies pay less attention to the individualistic and liberal assumption of moral agency or personhood and instead they place more emphasis on the importance of social roles and relationship to personhood. Arguably, Asian philosophies such as Confucianism and Buddhism may provide possibilities "for nonhumans [such as robots] to reach the status of humans" (Cassauwers, 2019). As argued by Wong,

> In Confucianism, the state of reaching personhood is not a given. You need to achieve it. The person's attitude toward certain ethical virtues determines whether or not they reach the status of a human. That also means that we can attribute personhood to nonhuman things like robots when they play ethically relevant roles and duties as humans (Cassauwers, 2019).

Philosophical justifications for personhood in the West and the East may further lead to cultural differences in the public perceptions of robots. Spanish philosopher Jordi Vallverdú notices the cultural differences in human perceptions of robots between East and West: "Westerners are generally reluctant about the nature of robotics and AI, considering only humans as true beings, while Easterners more often consider devices as similar to humans" (Cassauwers, 2019).

Confucian ethics and designing morally competent robots

In this section, we are trying to provide some practical examples that demonstrate possible ways in which Confucian ethics can help to understand, inform, and shape the design of robots, and how this design process can help us refine our understanding of Confucian ethics.

One of the fundamental activities undertaken by robot designers is the identification, refinement, and application of *design patterns* (Alexander, et al., 1977; Borchers, 2000; Kahn, et al., 2008): abstract patterns of human interaction with the physical and social world that can be flexibly instantiated, nested, and combined. Kahn et al. (2008) list *claiming unfair treatment or wrongful harm* as one of the key design patterns for social robots, alongside common activities such as initial introductions, didactic communication, and recovering from mistakes, and describe how this key interaction pattern of *protest* can be instantiated from both deontological and consequentialist perspectives in order to assert a robot's moral standing.

This design pattern is itself just one example of the broader design pattern of *identifying* unfair treatment or wrongful harm, i.e., protesting an action not necessarily on the basis of unfairness or harm towards oneself, but more generally on the basis of some identified unfairness or harm. In our own work, we have examined the tradeoffs between different instantiations of this design pattern. Specifically, we have looked at different *Speech Act-theoretic* (Searl, 1969) that robots might reject commands and requests when they are identified as harmful, examining the differential effects of

phrasing rejections as *questions*, *statements*, or *rebukes* (Jackson, Wen, & Williams, 2019).

But critically, as identified by Kahn et al., this design pattern can also be instantiated according to different ethical frameworks. Accordingly, this design pattern presents an excellent testing ground for applying and evaluating the effectiveness of different ethical frameworks. In recent work, we have accordingly begun to examine how humans perceive robot rejection of inappropriate commands when those rejections vary not only according to Speech Act theoretic phrasing, but also according to underlying ethical framework (Wen, Jackson, Williams, & Zhu, 2019). For example, if a robot serving as an instructor is asked to perform an action that constitutes or facilitates cheating, it may issue a question-phrased rejection in multiple ways: asking "Wouldn't that be cheating?" draws direct attention to the norm that would be violated if the directive were accepted, whereas "Would a good instructor do that?" instead draws attention to the robot's role, only directly highlighting the prohibitive norm. While our preliminary results suggest that these role-based norm violation responses lead humans to perceive robots as better fulfilling their professed roles, we have not yet found any evidence to suggest that role-based responses lead to any other effects that we might expect, such as increased mindfulness and self reflection. In our current work, we are designing experiments that more fully address cultural, contextual, and temporal considerations that may have prevented us from observing these other hypothesized effects. If our experiments elicit our hypothesized effects, this will serve as a strong argument in favor of robotic moral language generation grounded in role-based frameworks such as Confucian ethics.

Robot designers interested in enabling moral language generation grounded in Confucian ethics must make design decisions that articulate different positions and priorities within the Confucian perspective. First, Confucian ethics is a role-oriented paradigm, and thus designers must decide what roles are appropriate for robots to play within human society. In Confucian classics, five cardinal relationships (*wulun*, 五伦) are delineated: ruler-minister, father-son, husband-wife, older-younger, and friend-friend. For human-robot relationships, designers must articulate an equivalent set of cardinal human-robot relationships, e.g. supervisor-subordinate, owner-ownee, adept-novice, teammate-teammate, and friend-friend. Critically, a designer's choice of represented relationships may impact not only the contents of robots' norm violation responses, but the decisions as to whether those responses are generated in the first place. For example, from the point of view of Confucian ethics, if the *friend-friend* relationship holds between agents A and B, then A has the role ethics of remonstrating with B when A observes B committing or proposing wrongdoing: this remonstration is a requirement for A to be a good friend of B. Thus if designers choose to include *friend-friend* among the cardinal human-robot relationships they choose to represent, then this may affect the frequency with which those robots should choose to respond to proposed norm violations, and may impact the Speech Act-theoretic phrasing that robots should use in such circumstances (i.e., blame-laden moral *rebukes* may be necessary for proper remonstration when the friend-friend relationship holds between a robot and its observed norm violator).

Second, Confucian ethics espouses multiple competing objectives that may conflict with each other. For example, Confucian ethics emphasizes both self reflection

and emotional display of role commitment. When phrasing a rejection from a role-based perspective, it is possible to differentially encourage these different outcomes: by using a role-based interrogative (e.g., "Would you be a good friend if you did that?"), the robot may be able to encourage more self reflection; while when using a role-based rebuke ("You are a bad friend for asking me to do that!") the robot may instead demonstrate more emotional commitment to its role (and provoke a greater emotional response). The process of designing a robot that must respond to unacceptable commands thus forces us to think critically not only about what constitutes an ethical response, but also about how different aspects of an ethical framework may conflict with each other or be chosen between. Moreover, this provides us an opportunity to interrogate those ethical commitments. Ultimately, we must ask ourselves whether Confucian principles such as encouragement of self-reflection and emotional demonstrations of role adherence are the end goals we strive to achieve, or whether we are only seeking to achieve these goals because we believe they will lead others to take more role-fulfilling actions in the future. If our goal is the former, we can use the design, implementation, and evaluation of our computational models to identify whether those policies actually lead robots' interactants to achieve those principles; if our goal is the latter, we can further examine whether achievement of those intermediate principles actually correlates with achievement of our overarching societal goals. Such examination may allow the ethically-informed design process to feed back and inform the ethical framework itself, by quantifying the relative merits of the intermediate principles espoused by that framework.

References

Alexander, C., Ishikawa, S., Silverstein, M., Jacobson, M., Fiksdahl-King, I., & Shlomo, A. (1977). *A pattern language: Towns, buildings, construction.* New York, NY: Oxford University Press.

Ames, R. T. (2011). *Confucian role ethics: A vocabulary.* Honolulu, HI: The University of Hawai'i Press.

Borchers, J. O. (2000). A pattern approach to interaction design. *The 4th Conference on Designing Interactive Systems: Processes, Pactices, Methods, and Techniques* (pp. 369-378). New York: ACM Press.

Cassauwers, T. (2019, March 28). *How Confucian could put fears about artificial intelligence to bed.* Retrieved April 26, 2019, from OZY: https://www.ozy.com/immodest-proposal/how-confucianism-could-put-fears-about-artificial-intelligence-to-bed/93206

Dumouchel, P., & Damiano, L. (2017). *Living with robots.* Cambridge, MA: Havard University Press.

Hall, D., & Ames, R. T. (1994). Confucian friendship: The road to religiousness. In L. S. Rouner (Ed.), *The changing face of friendship* (pp. 77-94). Notre Dame, IN: University of Notre Dame Press.

Isaac, A. M., & Bridewell, W. (2017). White lies on silver tongues: Why robots need to deceive (and how). In P. Lin, R. Jenkins, & K. Abney (Eds.), *Robot ethics 2.0: From autonomous cars to artificial intelligence* (pp. 157-172). New York, NY: Oxford University Press.

Ivahoe, P. J. (2000). *Confucian moral self cultivation* (2nd ed.). Indianapolis, IN: Hackett Publishing.

Jackson, R. B., Wen, R., & Williams, T. (2019). Tact in noncompliance: The need for pragmatically apt responses to unethical commands. *The 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 499-505). Honolulu: ACM Press.

Kahn, P. H., Freier, N. G., Kanda, T., Ishiguro, H., Ruckert, J. H., Severson, R. L., & Kane, S. K. (2008). Design patterns for sociality in human-robot interaction. *The 3rd ACM/IEEE International Conference on Human Robot Interaction* (pp. 97-104). Amsterdam: ACM Press.

Lai, K. (2003). Confucian moral cultivation: Some parallels with musical training. In K.-c. Chong, S.-h. Tan, & C. L. Ten (Eds.), *The moral circle and the self: Chinese and Western approaches* (pp. 107-139). Chicago, IL: Open Court.

Lai, K. (2017). *An introduction to Chinese philosophy* (2nd ed.). Cambridge, UK: Cambridge University Press.

Liu, J. (2017). Confucian robotic ethics. *International Conference on the Relevance of the Classics under the Conditions of Modernity: Humanity and Science.* Hong Kong: Hong Kong Polytechnic University.

Lu, X. (2010). Rethinking Confucian friendship. *Asian Philosophy, 20*(3), 225-245.

Mattice, S. (2019). Confucian role ethics: Issues of naming, translation, and interpretation. In A. McLeod (Ed.), *The Bloomsbury research handbook of early Chinese ethics and political philosophy* (pp. 25-44). London, UK: Bloomsbury Academic.

Munro, D. (1969). *The concept of man in early China.* Stanford, CA: Stanford University Press.

Rosemont, H., & Ames, R. T. (2016). *Confucian role ethics: A moral vision for the 21st century.* Göttingen, Germany: Vandenhoeck & Ruprecht.

Searl, J. (1969). *Speech acts: An essay in the philosophy of language.* New York: Cambridge University Press.

Shun, K.-l. (1993). Jen and li in the Analects. *Philosophy East and West, 43*(3), 457-479.

Vallor, S. (2018). *Technology and the virtues: A philosophical guide to a future worth wanting.* New York, NY: Oxford University Press.

van den Hoven, J., Vermaas, P. E., & van de Poel, I. (2015). Design for values: An introduction. In J. van den Hoven, P. E. Vermaas, & I. van de Poel (Eds.), *Handbook of ethics, values, and technological design* (pp. 1-8). Dordrecht, Netherlands: Springer.

Verbeek, P.-P. (2006). Materializing morality: Design ethics and technological mediation. *Science, Technology, & Human Values, 31*(3), 361-380.

Wen, R., Jackson, R. B., Williams, T., & Zhu, Q. (2019). Towards a role ethics approach to command rejection. *The 1st HRI Workshop on the Dark Side of Human-Robot Interaction: Ethical Considerations and Community Guidelines for the Field of HRI* (pp. 1-4). Daego, Korea: ACM Press.

Zhu, Q. (2018). Confucian ethics, ethical leadership, and engineering ethics education. *International Journal of Ethics Education, 3*(2), 169-179.

机器人是否应当拥有权利吗？如果机器有自我意识了又该怎么做呢？

文章来源：企鹅号 - 镜脉传媒

想象一下，在未来你的烤面包机能想到你要怎么烤面包，白天，它会在网上找新型有趣的烤面包方法，也许它会想跟你聊天 或是谈谈烤面包技术的新成果。那么它能被看作是一个人吗？你会不会怀疑你的烤面包机也有喜怒哀乐呢？如果有拔出插头算谋杀吗？你还能拥有它吗？会不会有一天我们需要给机器和人一样的权力？

人工智能早已遍及生活，折扣店的库储要依赖人工智能，互联网广告的准确推送也靠人工智能，它甚至能够完整地编写故事。现在，我们看不起Siri笨拙的情感模拟效果，但是在不久的将来我们很有可能难以辨别人类和人工智能。

那么存不存在拥有权利的机器呢？就目前来讲还没有，如果出现了，我们对它也还束手无策。权利哲学还不足以处理人工智能相关问题。在大多数人类或动物的权利申诉案中 核心问题都集中于意识的存在性。不幸的是没有人知道意识是什么。有人说意识是精神上的，有人说意识是基于物质的，类似气体或液体。准确的定义先放在一边 我们对意识有直观感受源于日常经历。我们熟知自我和周边事物，也知道无意识的物品是什么样的。一些神经学家认为，任何足够先进的系统都可以产生意识，所以，如果你的烤面包机的硬件够高级就可能拥有自我意识。如果是这样，它应获得权利吗？

先不说这么多，先想想我们所定义的“权利”对它而言有意义吗？意识赋予人类权利是因为它同时使人类能感受痛苦，这意味着人类不仅能感到疼痛 也会意识到疼痛。机器人不能也不会感知疼痛，除非我们编程让它们有所感觉。没有痛苦或快乐，就没有偏好选择，而权利也就毫无意义了。我们的人权和我们自己的感知息息相关，比如我们讨厌疼痛，是因为我们的大脑进化成这样以使我们得以生存，有时是防止我们被火烫伤，有时又是让我们逃离捕食者。所以我们发明了权利，以保护我们免受痛苦。即使是更抽象的权利，例如自由，都基于我们大脑对于不公平和公平的感知。一个无法移动的烤面包机，会不会在意自己被锁在笼子里呢？如果它没有对死亡的恐惧，它会介意被拆开吗？如果它没有自尊，它会介意被侮辱吗？但是，如果我们编程让机器人拥有感情会怎么样呢？

让它们支持正义，趋利避害，让它们有自我意识会让它们成为人吗？许多技术专家认为，当人工智能可以学习并自己创造更胜一筹的人工智能时，技术将会发生革命性的改变。到那时，我们的机器人的编程 将会远远脱离我们的控制，如果一个人工智能，像大多数生物的进化生物学一样，发现有必要对痛感进行编程会怎么样呢？

机器人应得这些权利吗？也许我们不必如此担心超智能机器人对我们形成的风险。而是应该多考虑我们对它们造成的危险，我们作为人类的身份是基于人类例外论的。所谓人类例外论即是每个人都独一无二 并被赋予了主宰自然世界的能力，人类自古就在否认其他生物能够像他们一样感到痛苦，在科学革命中期笛卡尔曾说，如果你愿意如此理解，动物就是单纯的自动机器人。

因此伤害一只兔子在道德上和打毛绒动物是一样的，许多最严重的危害人类的罪行的祸首都狡辩说受害者比起文明人都是动物，更成问题的是 我们否认机器人的权利能够获得经济利益。如果我们能强迫一个有情感的人工智能去做我们想要它做的，那经济潜力是无限的，毕竟我们以前做过，人类同胞迫于暴力而工作，而且我们也从未在为仆人的辩护上产生过矛盾，奴隶主认为奴隶在奴隶制度中受益，这让他们有庇身之所。反对妇女投票权的男人认为，把艰难的选择留给男人是女性所得的利益，农民们认为 照料和喂养动物就是杀了它们以满足我们饮食偏好的理由。如果机器人有了意识，那么那些说机器人不该拥有权利的人，特别是那些以此立场从中获利的人就免不了一场舌战了。

人工智能引起了对哲学界限的严重问题，如果有知觉的机器人是有意识且应有权利，我们会问什么问题呢？我们被迫提出基本问题。例如，是什么使我们成为人类？是什么让我们应得权利？不管我们怎么想 这个问题可能都需要尽快解决。如果机器人开始要求自己的权利，我们又该怎么做呢？我们可以从机器人要求权利中学到什么呢？

发表于: 2019-12-12
原文链接：https://kuaibao.qq.com/s/20191213A03Z0S00?refer=cp_1026

上一篇：人工智能正在对轮胎“动手动脚”
下一篇：边缘计算在物联网中发挥怎样作用?

人工智能机器人的权利与义务

人工智能大健康 2018-08-25 10:39

2016年人工智能呈现井喷式爆发并大放异彩，这距离人工智能概念的首次提出仅过去60年。英国科学家阿兰·图灵在1950年的《心智》杂志上发表了题为《计算机器和智能》的文章，提出了“图灵测试”：认为判断一台人造机器是否具有人类智能的充分条件，就是看其言语行为是否能够成功模拟人类的言语行为，若一台机器在人机对话中能够长时间地误导人类认定其为真人，那么这台机器就通过了图灵测试。进而我们需要探究人工智能的研究目的：一是在人造机器上模拟人类的智能行为，最终实现机器智能，而智能的实质是去重建一个简化的神经元网络，从而实现智能体在行为层面上与人类行为的相似。美国的肖恩·莱格和马库斯·胡特认为：“智能是主体在各种各样的纷繁复杂的环境中实现目标的能力。”如何测量和评价人工智能主体是否具有智能或者其智商如何，是一个很复杂的判断过程。如何通过智能模型进行测试是人类需要面对的问题，这个问题也实际上在回答“人何以为人”这个本质的问题。

人工智能机器人法律人格

如果考虑赋予人工智能的机器人以法律上拟制的法律人格，就要求其能够独立自主地做出相应的意思表示，具备独立的权利能力和行为能力，可以对自己的行为承担相应的法律责任。2016年，欧洲议会呼吁建立人工智能伦理准则时，提及要考虑赋予某些自主机器人（电子人，Electronic Persons）法律地位。而如何界定监管对象（即智能自主机器人）是机器人立法的起点。对于智能自主机器人，欧盟的法律事务委员会提出了四大特征：（1）通过传感器和/或借助与其环境交换数据（互联性）获得自主性的能力，以及分析那些数据；（2）从经历和交互中学习的能力；（3）机器人的物质支撑形式；（4）因环境而调整其行为和行动的能力。在主体地位方面，机器人应当被界定为自然人、法人、动物还是物体？是否需要创造新的主体类型（电子人），以便复杂的高级机器人可以享有权利，承担义务，并对其造成的损害承担责任？这些都是欧盟未来在对机器人立法时需要重点考虑的问题。

随着未来技术的发展以及人类对脑科学和自我认知的加深，如何合理判定人工智能是否具备与人类相类似的“智能”，并以此来判断是否应赋予人工智能以独立的法律人格地位，是需要各学科、各领域的专家进行分工配合完成的课题。

机器权利

从人类的历史发展道路来看，一个群体对自身权利的争取，不但是漫长的历史进程，而且充满着战火和硝烟。法国启蒙运动大思想家让·雅各布·卢梭在其名著《社会契约论》中，曾经这样写道：“人人生而自由，但却又无往不在枷锁之中。自以为是其他一切人的主人，反比其他一切人更是奴隶。”

随着机器人和人工智能系统越来越像人（外在表现形式或者内在机理），一个不可回避的问题就是，人类到底该如何对待机器人和人工智能系统？机器人和人工智能系统，或者至少某些特定类型的机器人，是否可以享有一定的道德地位或法律地位？由此，机器权利日益受到关注，成为人类社会无法回避的一个问题。动物与机器人最大的不同之处在于动物具有天然的生命，有生物属性，但是机器人是人类制造出来的，没有天然的生命属性，但是其是否具有独立意识尚未达成共识。那么，未来是否需要承认机器人等人工智能系统也具有机器权利，同时机器的权利在何种情况下可以行使，是否应该与人类拥有相同的权利，例如选举和被选举权等政治权利以及民事权利等。

20世纪最有影响力的科幻作家之一伊萨克·阿西莫夫于1942年在他的科幻小说《环舞》中首次提出了著名的机器人三原则：（1）机器人不得伤害人类，或看到人类受到伤害而袖手旁观。（2）机器人必须服从人类的命令，除非这条命令与第一条相矛盾。（3）机器人必须保护自己，除非这种保护与以上两条相矛盾。后来，阿西莫夫又加了第零条定律：机器人不得伤害人类整体，或因不作为而使人类整体受到伤害。根据这个原则，人类的利益是高于机器人的，机器人不能损害人类的利益。假设人类开发和设计了一种智能机器人用于制造军事产品，但是其通过自我学习设计和开发出了核武器或致命武器，此时人类是否可以基于人道主义和人类共同利益而消灭该机器人？机器人是否有能力决定其生存或是死亡或者说机器人是否有权利从事买卖活动呢？或者我们是否可以对机器人进行虐待以发泄不满？

智能医疗
人工智能大健康
时时发布最新人工智能及智能医疗方...
+ 关注

谁来赋权于机器人？

启蒙运动为资产阶级的自由平等提供了新的理论基础，但是有时这种理论还不得不披着宗教神学的外衣。美国《独立宣言》写道："人人生而平等，造物主赋予他们若干不可让与的权利，其中包括生存权、自由权和追求幸福的权利。"造物主，一种高高在上的万能的存在，赋予了每个人自由平等的权利。尽管达尔文的进化论，早已经证明了人类从来不是被创造出来的，而是不断进化的结果。不可否认，科学技术的发展破除了封建迷信，宗教再也无法主导人类社会。但是，科技技术的进步，让人类的能力被逐渐放大——我们创造出了机器人，而我们人类是否能够承担起一个"造物主"的角色，去赋予机器人权利呢？不同于地球上现存的任何物种，机器人毫无疑问是由人类创造出来的。在2016年的热播美剧《西部世界》中，西部世界里的机器人将人类作为上帝，任由人类消遣娱乐甚至杀戮，而等到机器人的意识觉醒，他们发现，人类远不是上帝。

是否应当由人类赋予机器人权利的问题，其实质在于是否承认机器人的主体地位问题。早在20世纪五六十年代，人工智能技术刚刚起步之时，就有哲学家提出：把机器人看作机器还是人造生命，主要取决于人们的决定而不是科学发现；而等到机器人技术足够成熟，机器人自身就会提出对权利的要求。1976年，阿西莫夫出版的科幻小说《机器管家》（The Positronic Man）就讲述了一个自我意识觉醒的智能机器人安德鲁想要成为人类的故事。安德鲁作为一个家政智能机器人，在他两百年的生命历程中，一直要求人类把他作为人类看待，为此，他开设机器人公司，研发新的技术，使得在生命体征上他和普通的人类一模一样，甚至最后要通过手术让自己的生命只剩下一年（因为机器人在可预期的将来是永生的），才能获得法律的认可，最终获得人类的生命。

赋予机器人哪些权利？

尽管黑色人种和女性在历史上曾经遭受不公平待遇，他们被剥夺或者限制了作为人的基本权利，但是，随着人类社会的进步，肤色和性别不再是享受基本人权的障碍。机器人的种类非常多，它们存在各种各样的形态，主要可分为人形或者是非人形机器人。在机器人自我意识觉醒的前提下，讨论赋予哪些机器人权利，是一个非常复杂的问题。比如，类人形的陪伴型机器人享受权利，人类可能容易接受；而动物形状的陪伴机器人享受权利可能就难以接受了，但这确实是正在发生的事实，2010年11月7日，在日本，一个海豹宠物机器人帕罗（Paro）获得了户籍，而帕罗的发明人在户口簿上的身份是父亲。拥有户籍是拥有公民权利的前提，机器人在日本可能逐渐会被赋予一些法律权利。其实，现阶段的宠物机器人跟真实的宠物在享受的权利上并没有什么不同，因为普通的宠物也需要登记才能够饲养。还有一类非陪伴型的机器人，它们的外形迥异。例如，自动驾驶汽车是否可以被视为机器人而享有权利？任何存在着芯片和自我意识的实体是否都应当被认为是应当享受权利的机器人？

机器人可以拥有哪些权利？

人类具有的法律上的一些基本权利包括生存权、平等权和一些政治权利。在目前的技术水平之下，机器人的意识尚未觉醒，机器人的财产属性还十分强大，也就是目前对于人来说，机器人只是工具，而非另一种智能物种。目前机器人尚不可能被赋予跟人一样的权利，因此，在上文提及的欧盟的动议中，提出要把最先进的自动化机器人的身份定位为"电子人"，并赋予这些机器人依法享有著作权、劳动权等"特定的权利与义务"。动议中提出的赋予机器人著作权，是一个十分紧迫的现实问题。由于人工智能技术的进步，机器人或者人工智能系统目前已经不是简单地执行人类的指令，而是具有了创造性的思维，能够进行独创性的内容创作，而这些之前都是人类所独有的智能。

在欧盟法律事务委员会的提案中，还以护理机器人为例，提出了对机器人有生理依赖的人类会产生情感上的依恋。因此，机器人应该始终被视为机械产物，这有助于防止人类对其产生情感依恋。这种担忧不是空穴来风。在中国，2017年4月，一个浙江大学研究人工智能技术的硕士和自己研发的智能机器人莹莹结婚了。这种浪漫爱情故事，不仅只存在于人和机器人之间，机器人之间同样存在。2015年7月，明和电机就举办了一场机器人与机器人之间的婚礼。现在看来，这种事情仿佛闹剧一般，但是随着人工智能技术的进步，这些问题都将成为摆在人类面前亟待解决的问题。

机器人的权利与义务

赋予一个人（机器人）以权利，就要对另一人施加义务和限制。类比人类对于动物的保护，在动物保护立法比较完善的欧盟国家，都是赋予动物不受人类虐待的权利；其根本的中心点还是通过限制人的行为，来达到对动物权利的保护。未来的世界，人类面对机器人的存在，是否也要通过限制自身的某些行为来赋予机器人一定的权利呢？他们最基本的“生命权”是否可以由人类剥夺呢？例如2015年加拿大研究人员研发的机器人HitchBoT在成功地通过搭车的方式穿越多个国家后，在美国被人类残忍“杀害”，即便如此，HitchBOT在其留下的遗言中说道：“我对人类的爱不会消退。”我们是否可以以人类的名义任意剥夺机器人的生命权呢？当机器不再是一堆冰冷的金属堆砌成的物品，当其有了独立的“意识”和判断能力，我们是否也应该尊重他们的生命及权利呢？

除了法律权利之外，我们还应该给予机器人最低限度的道德权利。我们不能滥用机器人，不能利用人类的主导地位对其进行虐待。未来如果机器人拥有了自我意识，我们是否也应当尊重其意愿或者说照顾其喜怒哀乐，而不能强制其从事一些其不愿意从事的工作或劳动？那就是我们对其他与我们在地球上共存的主体的最低限度的尊重。

肉毒素能讲通爱美客的港股故事吗?

翟菜花 18分钟前 评论

是辛吉飞不适合抖音，还是抖音不适合辛吉飞?

奇偶派 19分钟前 评论

加速盘活存量资产：产业园区创新运营的3大新举措

火石产业大脑 19分钟前 评论

文明足跡 活得科學 社會群體 電腦資訊

機器人的生命權力——《再. 創世》專題

再・創世 Cybernetic ・2021/08/25 ・4697字 ・閱讀時間約 9 分鐘

相關標籤： AI 人工智慧 (25) 仿生 (25) 圖靈測試 (8) 想像實驗 (1) 機器人 (78) 機器人權 (1) 深度學習 (21) 生存權 (1) 科幻電影 (8) 認知能力 (14)

熱門標籤： 量子力學 (47) CT值 (8) 後遺症 (3) 快篩 (7) 時間 (37) 宇宙 (82)

再・創世 Cybernetic

11 篇文章 ・ 26 位粉絲

＋追蹤

由策展人沈伯丞籌畫之藝術計畫《再・創世 Cybernetic》，嘗試從演化控制學的理論基礎上，探討仿生學、人工智慧、嵌合體與賽伯格以及環境控制學等新知識技術所構成的未來生命圖像。

- 作者／高涌泉

人有人權，機器人是否應該也有某種類似於人權的權力（姑且稱之為「機器人權」）？目前這個問題還沒有被大眾認可的標準答案，因為我們還不知道機器人是否值得擁有機器人權。以當下（2021）最先進的機器人來說，我想多數人對於將它「關機」，不會有絲毫猶豫，就像我們可以毫不在意地任意關掉（或開啟）最先進的蘋果電腦。也就是說，當前最先進的機器人還沒有先進到需要我們去擔心關機是否影響它的福祉（生命）。但是未來呢？當，譬如說，一萬年之後 （或許不用那麼久，說不定一千年之後即可？）的機器人，已具備人性（我稍後會討論出現這種狀況的機率），那時人類應該允許機器人擁有機器人權嗎？

機器人是否也應有屬於自己的機器人權。圖／Pexels

科幻電影中的機器人

對於這個假設性問題，好萊塢已經給了答案——你所能想像的狀況大約已經出現於某部科幻電影裡。有一類情境相當常見，讓我舉幾部很好看的片子為例來說明：

1. ***Ex Machina***（*人造意識*，也譯為*機械姬*；據說拉丁文片名本意是「來自機器」）。片中的女機器人在主人的設計下，已經具有足以通過「圖林測試」（Turing Test）的智能，但是她還進一步發展出主人所不知的自主意識，最終為了自由而殺死把機器人當作娛樂工具的主人。
2. ***Blade Runner***（*銀翼殺手*）。此片的機器人是仿生人，在外貌、語言以及行動上，與人類沒有區別，和 ***Ex Machina*** 中的機器人相比，好似更加先進，然而其機器人本質還是可以被一項對於情緒反應的測試揭發（有如測謊器的功能），片中機器人當然也被安排會為了避免「被退休」而殺人。
3. ***I, Robot***（*機械公敵*）。片中機器人在外貌上，與人類有明顯區分，它們被製造來服務人類，有遠超越人類的體能，因此必須遵循艾西莫夫「機器人三定律」（第一定律：機器人不得傷害人類，或坐視人類受到傷害；第二定律：機器人必須服從人類命令，除非命令抵觸第一定律；第三定律：機器人在不違背前兩定律的情況下，必須保護自己），此片也一樣安排讓機器人產生某種程度的自主意識（主角機器人甚至會做夢），以及與人類的衝突。

以上三部科幻片的共同點是人類製造出的機器人終究會產生某種具自由意志的心智，並且會捍衛自己生存的權利、抗拒人類的宰制。這種「覺醒、反抗、勝

TRENDING 熱門討論

即時 熱門


「基因恆久遠，一個永流傳」，隱…
1 18 分鐘前


【2022 年諾貝爾生理或醫學獎】復…
1 2 天前

使用「藍碳」捕捉二氧化碳的速度…
2 3 天前

公投第20案【藻礁公投】模擬器
1 5 天前

利」三部曲的故事也廣為其他科幻電影所採用。（有一部叫好叫座、由 HBO 推出的科幻電視影集 ***Westworld***（*西方極樂園*）也是大致循這樣的套路。）事實上，這樣的套路也出現在非科幻的一般劇情片中，大約是這種勵志招式符合某種人類的心理需求，所以很受歡迎。總之我們從機器人科幻片學到了兩件事：一是我們與機器人的關係取決於機器人能否產生自我意識（心靈），而且大家願意相信機器人應該終究會具有這樣的能力；第二是對於我們應該賦予機器人多少「機器人權」的問題，無論我們如何操心，恐怕不是重要的事，因為就如「人權是爭取來的，不是靠施捨的」，「機器人權」的內涵還是由機器人決定。不過這兩點若真要仔細推敲，就會發現可以質疑的地方非常多。

科幻片中人類製造出的機器人終究會產生某種具自由意志的心智 。圖／Pexels

機器人適用心物二元論還是原子論？

首先，到底什麼是機器人？對此我一直沒有下個定義，因為不需要：大家都很清楚機器人儘管在外型與行為上，有很多種類，但一定都是人造的。所以機器人就是人造人（或人工人），是人類用材料製造出來的。所謂的材料就是物質，物質拆解到最後，就是各種原子罷了。所以機器人都是原子組裝出來的。但是人類不也是由原子組成的嗎？為什麼人不是機器人？或者，人其實也是一種是機器人？也就是說人也不過是一群原子依據某種明確的指令（程式、算則）在運行罷了！然而自古以來，不斷有哲學家懷疑這樣的看法，因為大家想不透在這樣的假設下，自由意志（俗稱靈魂）如何能夠出現？如果不行，也就是說靈魂這東西和物質屬於兩個範疇（即所謂的二元論），那麼人當然就不是機器人了：人有靈魂，機器人沒有。

RELATED 相關文章

牠如何長出一雙「隱形的翅膀」？——玻璃翼… 日誌

AI 讓羽球訓練更聰明！智慧球拍上… 力」一測就知道

在機器與人的交會之處——《再・創世》專題

巨大機器人 x 沒爹沒娘的悲慘童年——懷舊動畫為… 54】

讓機器讀懂我們的心情！臺灣AI情緒辨識技術再突破

但是自古以來也有不少人不相信二元論，例如古希臘的原子論者就不相信有獨立的靈魂這回事，以法國哲學家柏格森（Henri Bergson）的話說，原子論者相信的是「身體、靈魂、所有的物體以及世界，都是由原子所構成的。自然現象和思維都只是原子的運動而已。一切的事物與現象都是由原子、原子之間的真空（void）、以及原子的運動所組成的。除此之外，就沒有其他東西了。」如果原子論終究是對的，那麼人便只是一類較高明的機器人罷了！

那麼到底二元論與原子論兩者間哪一個比較有道理？（當然了，聰明的哲學家還發明出其他更繁複細膩，或者說更怪異有趣的理論，例如原子本身就是有意識之物體的說法等等，感興趣的人可以自行探究。）自近代科學出現以來，由於物理、化學與生命科學以及電腦科學的快速進展，眾多科學家自然地認為原子論的觀（自然現象和思維都只是原子的運動而已）是一件合理的假設，理解意識如何出現在腦子裡於是成為眾多研究的目標。

認知思考的想像實驗

美國哲學家瑟爾（John Searle）在 1980 年提出一項想像實驗（類似的想法其他人也有），試圖證明意識絕不是物質加上（電腦）程式就能產生的，具體說，即電腦不可能具有思考能力。他這個想像實驗一般稱為「中文房間論證」（Chinese room argument），讓我用一個不同於瑟爾原始版本、但我想仍不失其意的簡化版來說明這個論證：設想在某房間裡有位美國哲學家，他不懂中文與日文，但是能夠依據指令行事，房間裡有個資料庫，裡面有一份中日文字對照表（對哲學家來說，這裡的中日文字都只是奇怪的符號而已）及一本以英文寫的中日文語法規則簿（即中日文字對照表內符號之間應遵循的關係），我們將一篇中文文章送進房裡，這位先生就依據房間裡的資料庫，將這篇文章「翻譯」成日文，然後送出房間。房間外的人會以為這篇文章是房間內有位懂

中文與日文的人所做的翻譯，但是瑟爾說房內的哲學家根本不知道他所經手的文章在講些什麼。

以行話說，瑟爾想示範的是掌握了「語法」（syntax）不意味就了解「語意」（semantics），而不了解語意就談不上認知與思考。總之，瑟爾的重點是知道依據明確的規律來操弄符號（這是所謂「人工智慧（智能）」（AI）的功能）儘管有翻譯的本事，但仍不具認知、理解與思考的能力，也就是他不相信電腦（AI）能夠導致意識與心靈。瑟爾的講法引發大量評論，有人主張自由意識根本是個幻覺（當然另有人說這麼想的人錯得離譜），也有人主張人腦與電腦有根本差異（但是究竟差異為何，則意見紛雜）等等。

在我的簡化版本中，瑟爾的想像實驗假設了機器翻譯是行得通的（但是即便如此，機器還是沒有意識可言），不過長久以來，機器翻譯其實一直沒有太大的進展。然而近年來由於大數據與深度學習方法的出現，機器翻譯的水準已經頗為可觀，儘管還算不上完美，但是已經不像更早些時，譯文漏洞百出，明顯就不是人為的。同樣地，深度學習也讓電腦下棋（無論是西洋棋或是圍棋）的功力，遠遠超越人類棋手。在翻譯與下棋之外，電腦還有很多令人刮目相看的新本事（傳統的本事當然是其快速計算的能力），所以就算電腦還談不上有真正的意識（無論這是什麼意思）可言，不少人（包括我）已經感到震撼。

經深度學習後，電腦也可以下西洋棋。圖／Pexels

對於意識等抽象概念的探討，如果沒有具體的例子作為對象，容易流於空泛，莫衷一是。目前在人類之外，什麼東西可能擁有某種程度的意識？動物是個明顯的答案，無怪乎科學家與哲學家對於動物的心智很感興趣。不過動物心智也不容易捉模，相關意見也一樣紛雜。據說，主張心物二元論的笛卡爾就認為由於動物沒有語言能力，因此談不上具有心智，不過我想很多養過寵物的人恐不會接受這個見解。

動物的心智與生存權

我自己雖不養寵物，但全然認同（起碼有些）動物是具有心智的。主因是我在過去五、六年間，迷上了在 YouTube上觀賞對於白頭鷹（bald eagle）的巢 24 小時全天候（晚上有紅外光夜視）的實況轉播。簡單講，整個情況就像電影 ***Truman Show***（楚門的世界）的白頭鷹版——除了白頭鷹的真實生活比虛假的楚門世界要有趣太多了。白頭鷹是美國國鳥，曾一度列入瀕臨滅絕物種（endangered species）名單，後來在種種保護措施（包括禁止殺蟲劑 DDT）下，族群數量才逐漸回升。白頭鷹是美麗的大型鳥，位於食物鏈頂端，有王者氣質，令人著迷。現在網路上可以找到很多位於世界各地的這種稱為「白頭鷹巢實況監視」（live bald eagle nest cam）的 YouTube 頻道，我最早看的是一個位於美國首都華盛頓特區的巢：多年來，有一對白頭鷹固定在那裡築巢、育（鷹）嬰，人類鷹迷們分別暱稱公、母鷹為「總統先生」與「第一夫人」；它們每年秋季回到這裡，修整離地數十公尺的巢、交配、產卵、孵卵、撫育幼鷹，直到幼鷹於初夏可以自行飛翔離巢。

位於食物鏈頂端，具有王者風範的白頭鷹。圖／Pexels

對於我這樣剛入門的觀鳥人，白頭鷹的一切生活習性都很有意思。例如，幼鷹一但接連破殼而出，殘酷的「手足競爭」（sibling rivalry）立即登場，父母不會介入這種（從觀眾留言可知，令不少人不忍心看的）天生的競爭，弟妹在受到兄姊的壓制之後，很快學到要避開對方的攻擊，並且如何在適當時機，迅速從父母口中搶到食物。又例如，公鷹規律地獵捕魚、松鼠等動物回巢，轉交母鷹餵食幼鷹。還有令我特別訝異的——父母會在下雨（雪）或大太陽時張開翅膀護著幼鷹。

觀鷹久了，我發現自己能夠預測老鷹的企圖，或者說可以領會老鷹在「想」些什麼、在動些什麼「心思」。老鷹儘管沒有語言，但是能夠發聲「呼喚」、「警告」、「恐嚇」其他老鷹或其他生物。我一點也不懷疑白頭鷹具有某種程度的心智。（知名哲學家奈格爾（Thomas Nagel）在 1974 年發表了一篇文章「身為蝙蝠會有什麼樣的感受？」（What is it like to be a bat?），他此文

的主張就是，不是蝙蝠的我們永遠不會知道蝙蝠的主觀感受是什麼。我自以為多少了解白頭鷹的心思，當然是不認同奈格爾的主張可以推廣至白頭鷹。）為什麼白頭鷹能夠具有心智，而機器人沒有？這就是當代心智研究的基本問題。我猜測關鍵在於演化與成長歷史：白頭鷹是經過長期自然演化而產生的物種，從出生至獨立成熟也有個成長過程，而機器人卻不是如此。

動物應該擁有生存權，尤其是那些我們覺得具有某種心智能力的動物，這是很多人認可的事（在很多社會這件事其實已經成為法律）。白頭鷹的生存受到保護，數目也逐年增加，愛鷹人士都很高興。但是如果白頭鷹的數目因為保護而過度增加以至於影響了人類的利益呢？是不是白頭鷹的生存權也應受到限制呢？（對於某些動物，這種情況不是已經出現了嗎？）總之，動物權的範圍操之於人類。

機器人目前的處境還遠在動物之下，我看不出機器人如何能夠因產生心智而改變這種狀況。即便機器人因本事提高，讓我們將它們如同白頭鷹看待，它們生存權的範圍大小，仍是取決於人類，除非它們的聰明才智超越包括人類在內的一切動物。

人類目前的科技水準還處於初級階段，或許在很久很久以後，人類可以製造出和蚊子一樣靈活的「機械蚊」，那時才開始來操心所謂機械人權的問題還不晚。

發表
意見

千萬別說《千萬別抬頭》有幫科學說了什麼——《科學月刊》

科學月刊 ・2022/04/08 ・2534字 ・閱讀時間約 5 分鐘

- 黃俊儒／中正大學通識教育中心特聘教授，物理學出身，學術研究專長為科學傳播與大眾科學教育，喜歡讀書與追劇。

Take Home Message

- 《千萬別抬頭》是去年充滿話題性的科學題材電影。片中以極為誇張的劇情，諷刺現今社會與政治環境對於科學議題的反應及現象。
- 黃俊儒認為，電影的敘事手法有引發高馬效應之虞，導演以高高在上的姿態來訓誡觀眾，但這樣的做法卻可能引起更多的爭議及分裂，無助於釐清真相。
- 雖然許多人表示這類的電影能引發更多對於科學的關注，但黃俊儒表示，若媒體僅「提到」科學就算是「關心」科學，而非協助閱聽眾進行理解與轉譯，將無法實質幫助到科學知識傳播。

無疑的，Netflix 於去（2021）年末推出的科幻電影《千萬別抬頭》（Don't Look Up）極具話題性，更一舉獲得今（2022）年奧斯卡獎最佳影片入圍。

可能是感動於科學議題難得被如此大規模地盛情對待，此片在科學圈裡也引起了相當程度的關注，包括片中的科學主題、科學家的處境、科學家與媒體的互動，甚至連科學家與公關的關係也有人提出討論。我們能否真的藉由一部電影，令觀影人從此更加關注科學、留意地球環境？

《千萬別抬頭》獲得今（2022）年奧斯卡獎最佳影片入圍。圖／IMDb

荒誕卻又既視感十足的戲碼

看過影片的人應該可以在片中感受到一種滿滿的既視感，例如彗星撞地球的災難，幾乎是各種現存科技爭議的翻版；誇張的美國女總統如同世界上任何一位無視科學證據、一意孤行的民粹式領導者；科學家的角色充滿了不諳人情世故，且夾雜社交恐懼的刻板形象；媒體所顯示的則是充滿了嗜血，以及熱衷配合政治炒作的投機本質等，難怪許多網友更直呼這部片根本就是在講臺灣的現況。

《千萬別抬頭》或許會給觀影者滿滿的既視感。圖／IMDb

種種昭然若揭的片段，滿足了大家心中的各種「不意外」，但是如果冷靜下來想想，這些情緒所反應的只是大家對於現況不滿的集結，體現一種酸民式的語彙邏輯。但劇情滿滿的酸味，恐怕也稀釋了深入咀嚼本片的機會。

過度嘲諷的劇情與現實脫節

> **在政治與科學傳播的研究領域中，有個「高馬效應」（high horse effect）理論。**

這個說法的起源與古代戰爭有關，不論是東方或西方，戰時的將領總是會騎一匹特別高大的駿馬，給人一種睥睨常人的威嚴與氣勢。而在現代，高馬效應則經常描述假消息傳播時所造成的影響，例如有人在群組上義正辭嚴地駁斥另一個人所張貼的假消息，原本這個指正應該是對的，但是因為當事人用高高在上的姿態訓誡他人，導致受訓誡的人心生不滿。

除了不領情之外，甚至可能將錯就錯地故意用更極端的方式回應當事人。這種高馬效應所造就的結果，經常不是事情的真相因此被釐清，而可能適得其反地造成了更大的分裂及衝突。

這部影片就充滿了高馬效應的風險。即使以科學作為主要的討論題材，但片中卻充滿了粗糙、線性且誇大的影射手法。從電影中可以強烈感受到導演急著凸顯某些荒謬感，結果卻微妙地轉化成滿滿的說教感，彷彿只有導演才是這些光怪陸離現象背後的先知，其他人都是被意識形態矇在鼓裡的傻瓜。

而這些說理的不成功，除了基於大量猛爆性的嘲諷之外，一部分則來自於對科學活動的脫序描述，不僅讓觀眾頻頻出戲，更無法感受到電影敘事的誠意。如果彗星撞地球這麼外顯的天文現象發生，全世界卻只能仰賴「兩個」科學家去確認並奔走，我想許多天文或地球科學界的學者都要氣到翻桌了。

以現在的科技，人們非得用肉眼看見頭頂上的長尾巴星象，才相信彗星已經朝我們而來，就跟非得看見狂風暴雨才相信颱風要來是一樣的，這是停留在遠古時代的荒謬想像。這些刻畫讓人在觀影過程中很難引發同理心，反而充斥著一種酸民邏輯下難以言喻的不舒適感。

當然，也有人對於支持這部電影，特別是國外有些氣候科學家，認為這部片忠實地呈現了他們數十年來被忽視的無奈。也會有人認為好不容易讓這些大卡司代言科學議題的電影，至少可以喚起大家的關注。

有些人認為這些大卡司代言科學議題的電影，至少可以喚起大家的關注。圖／IMDb

這些基於媒體對科學的經常性忽視，而好不容易被投以關愛眼神時油然而生的感謝，是無可厚非的。在筆者過往分析過的案例中，就經常感受到科學家在這類事情上的客氣，即使媒體把研究的成果報錯、報歪了，只要有稍微提及，科學家們就會覺得「沒關係，有提到就好」。

然而除了卑微的期待之外，真正嚴肅的問題應該是：這部片真的能夠引發大眾對於相關議題的關心嗎？

電影真能喚起大眾關注科學？

如果透過片中反諷與嘲笑的方式，就可以推進我們對於科學現況的關心及投入，甚至因此更願意去了解氣候變遷、疫苗施打、能源轉型等科學議題，那我是滿滿的悲觀。因為透過網友的鍵盤，除了一時帶風向的效果之外，我不認為有真正協助科學改變了什麼，反而更擔心透過導演高高在上的視角，不要再製造出更多的分裂與對立就已經是萬幸了。

「科學是一個很長的故事，媒體要的卻只是快門的一瞬間。」這是科學傳播研究裡面一句十分經典的話語。媒體如果要為科學講一個動人的故事，需要花足精神進行理解與轉譯，如果「提到」就算是「關心」科學，那只能說是大家的寬容。

所以千萬不要相信《千萬別抬頭》真的有幫科學說了些什麼，如果今年 3 月 28 日的奧斯卡金像獎，真的把最佳影片頒給了這部片，我只能佩服評審們媚俗的勇氣；如果沒有得獎，那就算是剛剛好而已。

- 〈本文選自《科學月刊》2022 年 4 月號〉
- 科學月刊／在一個資訊不值錢的時代中，試圖緊握那知識餘溫外，也不忘科學事實和自由價值至上的科普雜誌。

發表
意見

關於Deepfake色情影像：雖然內容是假的，但傷害是真的

雷雅淇 / y編 • 2022/01/29 • 3412字 • 閱讀時間約 7 分鐘

相關標籤： deepfake (7) GAN (3) 復仇式色情 (1) 換臉 (7) 深度學習 (21) 色情 (4) 色情網站 (2) 裸照 (1)

熱門標籤： 量子力學 (47) CT值 (8) 後遺症 (3) 快篩 (7) 時間 (37) 宇宙 (82)

2021 年 5 月，鏡週刊的深度報導《臉被偷走之後——無法可管的數位性暴力？台灣 Deepfake 事件獨家調查》中，揭露了早在 2020 年就已經存在、利用 Deepfake 深度造假技術製作收費色情影片的 Telegram 群組，並訪問了數位影片中的受害女性。

雷雅淇 / y編

37 篇文章 • 834 位粉絲

＋追蹤

之前是總編輯，代號是(y.)，是會在每年4、7、10、1月密切追新番的那種宅。中興生技學程畢業，台師大科教所沒畢業，對科學花心的這個也喜歡那個也愛，彷徨地不知道該追誰，索性決定要不見笑的通吃，因此正在科學傳播裡打怪練功衝裝備。

2021 年 10 月，臺灣警方透過 Twitter 上的換臉色情影片追蹤到了該群組，逮捕了相關涉案人，「被換臉」的人除了名人之外，也包含一般人的換臉影片。此案在臺灣掀起了翻天覆地的討論，內容圍繞在 Deepfake 技術的濫用、和數位性暴力等相關議題。

到底 Deepfake 色情內容有多氾濫？而它又造成了什麼樣的傷害呢？

大部分的 Deepfake 影片都是色情內容

圖／envato elements

「Deepfake」指的是利用人工智慧深度學習技術，在某人沒有說過、拍過和錄過某些內容的狀況下，生成他的聲音、圖像或影片。（延伸閱讀：Deepfake 不一定是問題，不知道才是大問題！關於 Deepfake，你需要知道的是……？）

早在 2016 年，來自斯坦福大學、馬克斯普朗克研究所和埃爾蘭根－紐倫堡大學的研究人員，就已經創建了一個名為 Face2Face 的系統，透過捕捉演員的面部表情，在其他人臉上生成一樣表情的影像。不過，我們現在熟悉的「Deepfake」一詞，卻是 2017 年才現身；而且它最早的姿態，便是色情內容：由一位名為「deepfake」的 Reddit 的用戶，上傳的明星假性愛影片。

Face2Face: Real-time Face Capture and Reenactment of RGB Videos (CVPR 2016 Oral)

直到現在，網路上雖然不乏一些迷因、或是跟政治人物相關的 Deepfake 影片到處散播，而且數量越來越多，但 Deepfake 應用最多的，仍是色情內容。人工智慧公司 Sensity AI 的統計發現，自 2018 年以來，網路上的虛假影片每六個月成長一倍，截自 2020 年 12 月為止，他們透過自己的檢測技術，偵測到至少85047 個假影片在網路上流傳。

2019 年，Sensity AI（在當初的名稱為 Deeptrace）的報告統計，有 96%的 Deepfake 影片是色情內容，在色情網站上的內容幾乎百分之百都是以女性為主，被觀看次數超過 1.34 億次。由此可見，「色情內容」無疑是所有 Deepfake 應用中，內容最多、製造速度最快、傳播最廣的類型。

色情網站龍頭 Pornhub

此外，儘管色情網站龍頭 Pornhub 指出，他們不會容許包含 Deepfake 影片在內的任何非自願性質的內容出現，但根據 Deeptrace 的統計，2019 年排名前十的色情網站中，有 8 個網站裡有 Deepfake 的內容，另外還有 9 個 only for Deepfake 的色情網站，而這些網站裡的內容，佔了 Deepfake 色情內容的九成以上。

雖然，這些影片中被換臉的主角通常是名人，但受到 Deepfake 色情內容所苦的，不只有他們。

雖然內容是假的，但傷害是真的

2019 年 6 月 23 日，有一個叫做 DeepNude 的網站上線了：它提供可以免費下載的應用程式，並利用生成對抗網路（*Generative Adversarial Network*，簡稱*GAN*）技術，將女性有穿衣服的照片生成為裸照。雖然由它生成的照片都有浮水印，但只要付費就能把浮水印變小；重點是，它的使用方法很傻瓜，只需要一張女性的照片就行了，而且「穿越少」效果「越好」。

這個應用程式因為 Motherboard 的報導而爆紅，甚至一度因為下載量過大，導致網站不堪負荷。諷刺的是，Motherboard 報導的目的，原本是為了要批評 DeepNude 對於女性的傷害。雖然DeepNude 隨後在 6 月 27 日下架（其後又在同年 7 月 19 日以 30000 美元出售給匿名買家）；但其應用程式直到今天，仍在一些開源社群、種子網站等地方被不斷的上傳下載。

圖／envato elements

2020 年，Sensity AI 發現了另一個與 DeepNude 很相似的 Telegram bot，只要向這隻機器人傳送照片，便能直接生成裸照、且可以透過付費，得到去除浮水印或免等候等服務。根據統計，這當中起碼有十多萬名女性被生成裸照，當中包含未成年人。雖然被報導之後，相關群組一樣已經被官方刪除，但這絕對不會是最後一次，Deepfake 被這樣惡性利用。

雖然 DeepNude 的創辦人表示，他當初設計這個程式的初衷，是被小時候在雜誌上看到的「X 光眼鏡」所啟發（其 logo 也是為此致敬），他有想過這個程式會不會傷害到人，但他也表示能用 DeepNude 做到的事，任何人用 Photoshop 也能做到；如果有任何人懷有惡意想要做壞事，那有沒有 DeepNude 並不會有影響。然而情況卻是，有人因為類似應用被威脅，甚至遭受到復仇式色情（Revenge porn，本用詞為俗稱，目前有倡議應稱呼「未經同意即散布之私密影像(Non-consensual pornography)」，以避免簡化問題本質）的攻擊。

所謂「復仇式色情」，指的是未經過他人同意，任意散佈含有他人色情內容之照片或影片等影像的報復手段。美國心理學會的一項研究發現，每 12 名女性中，就有一名最終在她們生命的某個階段成為復仇式色情片的受害者。在過往，合意或是被偷拍的親密影像在非自願的狀況下外流就已經夠難防了，有了 Deepfake 之後就更難了：因為當事人根本不會意識到有這樣的影像的存在。而這些內容被用於勒索，甚或是威脅名人、記者等案例層出不窮。

圖／envato elements

「我要怎麼證明那不是我？我走在路上、搭捷運，可能有一些陌生人，他看我或交頭接耳的時候，我都會覺得，是不是他們看過那個影片？覺得我是那樣的女生？覺得我是很糟糕的人？」在鏡週刊的訪問中，被換臉的 Youtuber 球球這樣說道。儘管那真的不是自己，儘管內容是假的，但這類的色情內容造成的傷害卻是真實的。

波士頓大學法律學教授 Dielle Citron 在他的著作《網路空間裡的仇恨犯罪 (Hate Crimes in Cyberspace)》中提到：Deepfake技術正在成為針對女性的武器，這些性愛影片當中的身體雖然不是自己的，卻會對當事人造成影響，讓他們不想再上網、難以獲得或維持工作、並且感到不安。

Deepfake 最大的傷害不是來自技術，而是使用方式

從大量資料到一張照片、從專家操作到素人也行、從粗糙到以假亂真，Deepfake 的技術一直進步，而這已然打開的潘朵拉的盒子，要關上的方法，也必然需要人與科技的協力。

從辨偽技術的進步（延伸閱讀：Deepfake 辨偽技術如何在魔高一尺時，能道高一丈呢？）、法規制度的更新（延伸閱讀：應對Deepfake濫用，台灣修正刑法夠用嗎？），到協助被害人刪除與爭取權益等制度的完善，以及不看、不擴散，多理解技術對社會可能的影響：在面對 Deepfake 色情內容所造成的傷害，沒有人是局外人。

參考資料

1. Deepfake Porn Nearly Ruined My Life
2. How to Detect a Deepfake Online: Image Forensics and Analysis of Deepfake Videos – Sensity
3. THE STATE OF DEEPFAKES
4. Deepnude: The Horrifying App Undressing Women
5. 臉被偷走之後：無法可管的數位性暴力？台灣Deepfake 事件獨家調查- 鏡週刊 Mirror Media
6. Deepfake porn is ruining women's lives. Now the law may finally ban it. | MIT Technology Review
7. Deepfakes have got Congress panicking. This is what it needs to do. | MIT Technology Review
8. A deepfake bot is being used to “undress” underage girls | MIT Technology Review
9. Nonconsensual Pornography Among US Adults: A Sexual Scripts Framework on Victimization, Perpetration, and Health Correlates fo
10. An AI app that “undressed” women shows how deepfakes harm the most vulnerable | MIT Technology Review

11. Deepfakes have got Congress panicking. This is what it needs to do. | MIT Technology Review
12. The biggest threat of deepfakes isn't the deepfakes themselves | MIT Technology Review
13. How to Detect a Deepfake Online: Image Forensics and Analysis of Deepfake Videos – Sensity
14. The year deepfakes went mainstream | MIT Technology Review
15. Inside the strange new world of being a deepfake actor | MIT Technology Review
16. I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me | HuffPost UK News

發表
意見

Deepfake 不一定是問題，不知道才是大問題！關於 Deepfake，你需要知道的是……？

TingWei ・2022/01/24 ・3489字 ・閱讀時間約 7 分鐘

相關標籤： AI (45) CG (1) deepfake (7) facebook (18) GAN (3) 人工智慧 (73) 合成 (7) 名人 (4) 愛因斯坦 (58) 換臉 (7) 星際大戰 (6) 歐巴馬 (2) 深度偽造 (1) 深度學習 (21) 演算法 (29) 玩命關頭 (1) 班傑明的奇幻旅程 (2) 電影 (123) 魔戒 (2)

熱門標籤： 量子力學 (47) CT值 (8) 後遺症 (3) 快篩 (7) 時間 (37) 宇宙 (82)

編按：你的理智知道「眼見不為憑」，但你的眼睛還是會背叛你的理智，不自覺得被眼前的影像所吸引，儘管你真的、真的知道他是假的。Youtuber 小玉於2021年底涉嫌利用 Deepfake 技術，偽造多位名人的色情影音內容並販售的事件，既不是第一起、也不是唯一、更不會是最後一個利用「深偽技術」進行科技犯罪的事件。
當科技在走，社會和法律該如何跟上甚至超前部署呢？本次 Deepfake 專題，由泛科學和法律白話文合作，從Deepfake 技術與辨偽技術、到法律如何因應，讓我們一起全方位解析Deepfake！
第一篇，讓我們就 Deepfake 技術做一基礎的介紹，那我們就開始囉！

什麼是 Deepfake？

深偽技術 Deepfake 於 2017 年陸續開始進入大眾的目光中。原文 Deepfake 源自於英文「deep learning」（深度學習）和「fake」（偽造）組合，主要意指應用人工智慧深度學習的技術，合成某個（不一定存在的）人的圖像或影片、甚至聲音。最常見的應用，就是將影片中的人臉替換為另一張臉（常是名人），讓指定的臉在影片中做出自己從未說過或做過的事情。

利用深度學習技術合成或是置換人臉的技術，都是屬於Deepfake。圖 / stephenwolfram

現今談到 Deepfake，大多數人想到的可能是偽造的成人影片，就如前述 Youtuber 小玉的事件，Deepfake 一開始受到關注，主要與名人或明星的臉部影像被合成到成人影片有關，然而，Deepfake 的功能遠不僅於此，相關的技術使用還包括了替換表情、合成一整張臉、合成語音等等。

TingWei
13 篇文章 ・ 10 位粉絲
＋追蹤
據說一生科科的生科中人，不務正業嗜好以書櫃堆滿房間，努力養活雙貓為近期的主要人生目標。

除了像是讓過去或現在的名人在影片中「栩栩如生」做出使用者想要的表情與動作，之前在社群媒體上曾有好幾款 APP一度風靡，包括上傳一張照片就可以看看「變老」「變性」自己的 FaceApp，甚至於讓自己的臉在經典電影中講上一段台詞的「去演」APP，這類的功能也是應用前述 Deepfake 的技術。

雖然有些線索顯示這類 APP 常有潛在的資安疑慮[註]，但好歹技術的成果多屬搏君一燦自娛娛人，尚可視為無傷大雅。

「栩栩如生」的愛因斯坦

而過往電影的影音產業要仿造人臉需要應用許多複雜、耗時、昂貴的電腦模擬，有了 Deepfake 相關的技術，也使得許多只能抱憾放棄的事情出現了彌補的空間。最有名的應用應是好萊塢電影《玩命關頭7》與《星際大戰》系列。《玩命關頭7》拍攝期間主角保羅・沃克（Paul William Walker IV）意外身亡，剩下的戲份後來由弟弟擔綱演出，劇組再以 Deepfake 的技術讓哥哥弟弟連戲，整部電影才得以殺青上映。

RELATED 相關文章

從「自動化」進化成「智動化」——智慧… 業的未來趨勢

AI 是理科「主場」？AI 也可以成為文科人的助力！

應對Deepfake濫用，台灣修正刑法夠用嗎？

關於Deepfake色情影像：雖然內容是假的，但傷害… 的

Deepfake製作的虛假訊息氾濫，美國如何立法力… 瀾？

Weta Digital 說明如何讓保羅・沃克的弟弟 Brian O'Conner 能透過 Deefake 的技術，繼續協助 保羅・沃克演完《玩命關頭7》

Deepfake 讓「變臉」變得太容易了？

想想過去的電影如《魔戒》中的咕嚕、或是 2008 年布萊德・彼特主演的《班傑明的奇幻旅程》，將影片或照片中人物「換臉」「變老」的修圖或 CG 技術，在 Deepfake 出世之前就已經存在了。Deepfake 受到關注的核心關鍵在於，應用 AI 的深度學習的演算法，加上越來越強大的電腦與手機運算能力，讓「影片換臉」這件事情變得越來越隨手可得、並且天衣無縫。

利用CG技術把布萊德・彼特「變老」。圖 / © 2008 – Paramount Pictures

過往電影中採用的 CG 技術要花好幾個月由專業人士進行後製，才能取得難辨真偽的影像效果，而應用了 AI 演算法，只需要一台桌上型電腦甚或是手機，上網就可以取得軟體、有機會獲得差強人意的結果了。

進一步，傳統軟體演算法主要依靠工程師的持續修改調整，而如 Deepfake 這類技術，內部的演算法會經過訓練持續進化。有許多技術被應用於提高 Deepfake 的偽造效果，其中最常見的一個作法被稱為「生成對抗網路（Generative Adversarial Network, GAN）」，這裡面包含了兩組神經網路「生成器（Generator）」和「辨識器（Discriminator）」。

在投入訓練資料之後，這兩組神經網路會相互學習訓練，有點像是坐在主人頭上的小天使與小惡魔會互相吐槽、口才越來越好、想出更好的點子；在練習的過程中，「生成器」會持續生成偽造的影像，而「辨識器」則負責評分，反覆訓練下來，偽造生成的技術進步，辨識偽造的技術也得以進步。

舉例來說，This Person Does Not Exist 這個網站就充滿了使用 GAN 架構建構的人臉，這個網站中的人臉看上去非常真實，實際上都是 AI 製造出來的「假臉」。

This Person Does Not Exist 裡的「假臉」。

Deepfake 影片不一定是問題，不知道是 Deepfake 才是問題

現今的 Deepfake 技術得以持續進步、騙過人眼是許多人努力的成果，也不見得都是壞事。像是《星際大戰：俠盜一號》片尾，年輕的萊婭公主出面驚鴻一瞥，就帶給許多老粉絲驚喜。這項技術應用癥結在於，相關演算法輕易就能取得，除了讓有心人可以藉以產製色情影片（這類影片佔了Deepfake濫用的半數以上），Deepfake 製造的影片在人們不知情的情況下，很有可能成為虛假訊息的載體、心理戰的武器，甚至於影響選戰與輿情。

因此，Deepfake 弄假似真不是問題，閱聽者因此「不辨真假」才將是最大的問題所在。

歐巴馬的 Deepfake 影片

相關的研究人員歸納了幾個這類「變臉」影片常見的特徵，可以用來初步辨識眼前的影片是不是偽造的。

首先，由於 AI 尚無法非常細緻的處理一些動作細節，因此其眨眼、視線變化或臉部抽蓄的動作會較不自然。其次，通常在邊緣處，如髮絲、臉的邊緣線、耳環等區域會出現不連貫的狀況。最後，在一些結構細節會出現不合理的陰影瑕疵，像是嘴角的角度位置等。

由於現階段的 Deepfake 通常需要大量的訓練資料（影像或影片）才能達到理想的偽造成果，因此會遭到「換臉」的受害者，主要集中在影像資源豐富的名人，如電影明星、Youtuber、政治人物等。需要注意的是，如果有人意圖使用 Deepfake 技術製造假消息，其所製造的影片不見得需要非常完美，有可能反而降低解析度、非常粗糙，一般人如用手機瀏覽往往難辨真假。

人眼已經難辨真假，那麼以子之矛攻彼之盾，以 AI 技術辨識找出 Deepfake 的成品，有沒有機會呢？隨著 Deepfake 逐漸成為熱門的議題，有許多團隊也開始試圖藉由深度學習技術，辨識偽造影像。2020 年臉書與微軟開始舉辦的「換臉偵測大賽」（Deepfake Detection Challenge）就提供高額獎金，徵求能夠辨識造假影片的技術。然而成果只能說是差強人意，面對從未接觸過的影片，第一名辨識的準確率僅為 65.18%。

「換臉偵測大賽」（Deepfake Detection Challenge）的辨識素材。圖／MetaAi

對於 Deepfake 可能遭到的濫用，某部分我們可以寄望技術的發展未來終將「道高一尺」，讓社群平台上的影像不致於毫無遮攔、照單全收；然而技術持續「魔高一丈」讓防範的科技追著跑，也是顯而易見的。

社群網路 FB 在 2020 年宣布全面禁止 Deepfake 產生的影片，一旦有確認者立即刪除，twitter 則強制註記影片為造假影片。Deepfake 僅僅是未來面對 AI 浪潮，科技社會所需要應對的其中一項議題，法律、社會規範如何跟上？如何解決箇中的著作權與倫理問題？這些都將是需要經過層層討論與驗證的重要課題。

至少大家應該心知肚明，過往的網路流行語：「有圖有真相」已經過去，接下來即將面臨的，是一個「有影片也難有真相」的網路世界了。

- 註解：推出 FaceApp 與「去演」的兩家公司其軟體皆要求註冊，且對於上傳資料之後續處理交代不清，被認為有侵犯使用者隱私權之疑慮。

參考資料

1. Deepfakes and the New AI-Generated Fake Media Creation-Detection Arms

Race – Scientific American
2. What To Do About Deepfakes | March 2021 | Communications of the ACM
3. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, *64*, 131-148.
4. Deepfake 深偽技術的技術濫用與道德困境，大眾正要開始面對 | TechNews 科技新報
5. 台灣團隊研究辨識Deep Fake影片　深偽技術的正邪之戰開打 | 台灣事實查核中心 (tfc-taiwan.org.tw)

發表
意見

智能机器人还不能成为法律上的人

冯珏

2019年06月19日08:22 | 来源：经济参考报

原标题：智能机器人还不能成为法律上的人

意志、理性均源于人类心灵的能力，正是这种能力使得为人类确立道德法则成为可能。而在现有的技术水平下（弱人工智能时代），智能机器（人）显然不具备人类心灵的能力。不能基于这种人工的“智能”，就认为智能机器（人）可以与自然人比肩而成为法律中的人。

从功利的角度出发，是否需要为智能机器（人）拟制主体地位，取决于这种拟制的目的是否正当、手段是否合适，即是否符合工具理性的要求。仅出于限制智能机器（人）的制造商或设计者的责任的目的，不构成为拟制法律人格的正当目的。

或许在将来的某一天，从功利的角度出发需要拟制智能机器（人）的法律人格，这种必要性就需要有实践的基础并得到充分的论证。

随着人工智能技术的发展，“人工智能”、“智能机器”、“智能机器人”这些概念进入了人们的视野。内华达州开了美国自动驾驶汽车立法之先河，于2011年通过AB 511法案，对“人工智能”下了一个定义：人工智能是指使用计算机和相关设备以使机器能够重复或模仿人类的行为。当智能机器的概念与人们原已拥有的“机器人”的观念相结合的时候，“智能机器人”的概念又被提出来了。这样，人们不禁疑惑，智能机器（人）在法律上能不能算人？

2017年2月16日，欧洲议会投票表决通过《就机器人民事法律规则向欧盟委员会的立法建议［2015/2103（INL）］》（以下简称“机器人民事法律规则”），其中最引人注目之处，就是建议对最复杂的自主智能机器人，可以考虑赋予其法律地位，在法律上承认其为“电子人”（electronic person）。不过，是否承认智能机器人具有法律人格，存在激烈的争论。

讨论智能机器（人）是否可以被赋予法律人格或法律主体的问题，首先需要厘清的是，什么是“法律中的人”？

有权利能力

才能成为法律认可的“人”

为了确定谁可以成为法律中的人，民法学引入了“权利能力”这一概念。权利能力成为法律中的人的“标签”。简言之，有权利能力，就是法律中的人；没有权利能力，就不是法律中的人。

作为近代以来“人的解放”的重要成果，每个自然人（man）生而为人（person），不分年龄、性别、民族、种族、职业等具体情形的不同，都具有权利能力，此为各国国内立法和国际公约所普遍承认。但是，法律中的人（person）与我们通常所理解的“人”，并非同一个概念。为了将这两者相区分，我们将“法律中的人”用“法律主体”这个概念来替换。

热门排行

1 天舟二号货运飞船发射圆满成功 回顾升空...
2 拆“快递”啦！天舟二号都送了啥？
3 天舟二号货运飞船发射成功
4 我国第二代静止轨道气象卫星风云四号首颗...
5 鸿蒙登场！它的征途是万物互联
6 “长七”成功送“天舟”：实现“零窗口”...
7 这些太空“黑科技”让天舟二号“出手不凡”
8 风云四号B星成功发射 我国气象卫星观测...
9 “火眼金睛”这样炼成 细数风云四号B星...
10 放眼全球 风云四号实力如何？一文带你读...

法律中的主体，除了自然人之外，尚有“法人”。我国《民法总则》第57条明确规定，法人具有权利能力。其实，“法人”本来就是“法律中的人”的简化表达形式，从这个意义上讲，自然人具有权利能力，也是“法人”。但是习惯上，人们将“法人”概念限缩为专门用来指称除自然人以外的法律中的人，按照《民法总则》的规定，法人是具有权利能力的组织。

若是将智能机器（人）定位为电子人（electronic person），即一方面是说，智能机器（人）也是法律中的人（person）；另一方面，这种法律中的人既不是自然人（natural person）也不是法人（legal person），而是一种新的类别——电子人。若是智能机器（人）与自然人、法人一样，可以是法律中的人，也就意味着他们之间必然具有某种共性。

那么，智能机器（人）是在哪些方面与自然人或法人相类似，以至于需要赋予其法律人格或法律主体地位呢?

智能机器（人）

还不具备自然人的心灵能力

法律中的人的“权利能力”，究竟是谁的能力、又是什么样的能力呢？权利能力是法学对人的理性能力的抽象化和概念化。依现有的哲学，意志、理性均源于人类心灵的能力，正是这种能力使得为人类确立道德法则成为可能。用康德的话说，“可以把纯粹理性看成是一种制定法规的能力”。

即使是尚不具备理性能力的婴幼儿、虽然成年但却丧失了理性能力的不幸者，法律也承认其具有权利能力。对此需要加以解释，人的意志、理性是人类区别于自然界其他事物、生命的本质特征，因此，即便某些人类个体与一般情况有所偏离，仍不妨碍我们将其视为同类，并且依“等者等之”的正义观，承认其仍然为法律中的人。

智能机器（人）与自然人的类比，显然不是从其外在物理或形体特征角度出发的。智能机器（人）不是因为外形与人类相似，而使人产生它们也是“人”的联想的。从目前的讨论来看，智能机器（人）与人类的比较，着眼点在于其“智能”。

AlphaGo战胜人类最优秀的棋手、微软小冰能够“创作”诗歌，这些人工智能技术的新成果，使得不少人认为有的智能机器（人）可以比人类更聪明。而欧洲议会建议赋予电子人地位的智能机器人，也是着眼于最复杂的自主智能机器人。这种类比思维使我们回过头来反思，究竟什么才是人类的理性能力？是不是会下围棋、可以写诗，就具备了人类的理性能力?

需要指出，机器人的所谓自主性，具有的是纯技术本质。欧洲议会建议的“机器人民事法律规则”明确说明：机器人的自主性可以被界定为，在外部世界作出独立于外在控制或影响的决定并实施这些决定的能力；这一自主性具有纯技术本质，且其自主性程度取决于机器人被设计成的可与其环境进行交互的复杂程度。欧洲议会关于赋予智能机器人电子人格的建议，并非因为其自主性使其具有了人类的理性能力。

近期，中科院院士张钹接受记者采访，指出引发这一轮人工智能技术进展的深度学习算法，本质上是概率统计。深度学习是寻找那些重复出现的模式，重复多了就被认为是规律（真理）。因此，现在形成的人工智能系统非常脆弱，容易受攻击或欺骗，需要大量的数据，而且不可解释，存在严重的缺陷。张院士认为，我们现在还没有进入人工智能的核心问题。人工智能的核心是知识表示、不确定性推理等，知识表示、在开放系统中进行不确定推理的能力等，才是人类理性的根本。

总结而言，在现有的技术水平下（弱人工智能时代），智能机器（人）没有自主目的、不会反思、不会提出问题、无法进行因果性思考、没有自己的符号系统，显然不具备人类心灵的能力。虽然人工智能使用了“智能”这一语词，但是这个“智能”前面还有“人工”这一限定语。因此，不能基于这种人工的“智能”，认为智能机器（人）可以与自然人比肩而成为法律中的人。

智能机器（人）

与法人的比较

智能机器（人）与法人的比较，则存在更多的争议与不确定性。这是因为，法人是否具有自然人据以成为法律中的人的权利能力的内核即理性能力，一直存在极大的争论。

关于法人本质的几百年来的争论，随着各国立法普遍承认（或赋予）法人以权利能力、承认其作为法律主体的地位，而逐渐淡出人们的视野。而智能机器（人）的法律地位问题，使这个古老的问题又焕发出生机。一些人正是从与法人的对比中获得灵感，主张或呼吁赋予智能机器（人）以法律主体地位，这就要求对作为被比较对象的法人有进一步的认识。关于法人本质的问题，有一些共识是学界已经达成的，这些共识可以作为我们对比智能机器（人）与法人的出发点。

首先，法人的背后其实离不开自然人。

法人的目的由自然人设定，法人通过自然人的代表（或代理）从事民事交往，法人通过自然人的理性响应法律的行止要求。人类社会之所以可以由法律来调整并构建秩序，是由于人的理性决定了法律能够通过规范人们的行为来调整社会关系。那么，智能机器（人）能否理解法律的行止要求？能否根据法律的要求来规范自己的“决策”？在现有技术条件下显然无法做到。这样，智能机器（人）也不得不依赖于其背后的人来响应法律的要求。例如，德国《道路交通法第八修正案》第1a条要求高度或者全自动驾驶汽车能够在高度或者全自动驾驶期间遵守交通法规，这就要求设计或制造自动驾驶汽车的人将道路交通规则内化于自动驾驶汽车的决策逻辑之中。所以，德国的法律是对人提出了要求，而不是对自动驾驶汽车本身提出要求。

其次，法人作为其背后众多自然人所构建的法律关系的结点，有利于简化法律关系，便利民事交往。

以社团性法人为例，如某钓鱼俱乐部，可能有众多的俱乐部成员甲、乙、丙等等。如果该俱乐部需要购买钓鱼用品，或者需要租用钓鱼场所，以俱乐部的名义来缔结合同并享有权利、承担义务，比起以甲、乙、丙等等的名义来缔结合同，要简便得多。那么，将智能机器（人）类比于法人，能否实现简化法律关系、便利民事交往的目的呢？至少从目前来看，还无法想象如何通过赋予智能机器（人）法律主体地位以实现法律关系的简化。

再次，法人能够实现将特定财产用于特定目的的财产特定化需求。

财产的特定化，在遗产的限定继承和将遗产本身作为一个财团法人方面，就有所体现。近代以来，企业的产生、主权者之间的经济竞赛的需要，使得财产的特定化以法人或者信托的形式得到极大发展。从人们对智能机器（人）成为电子人的遐想来看，也要求其制造商、设计者、销售者或者其他利益相关者以智能机器（人）登记为基础，为其设立责任基金。在本质上，这一做法就是为智能机器（人）分配一定的财产并且将其特定化为智能机器（人）的责任财产。但是，通过法人实现的财产特定化，实际上是对近代以来法人作为社会生产的基本组织单位的法律认可。相较而言，目前关于智能机器（人）致损事故法律责任的讨论，包括欧洲议会建议的“机器人民事法律规则”，均将智能机器（人）的生产者或设计者的产品责任作为处理智能机器致损事故的主要法律机制。显然，智能机器（人）的生产者和设计者通常都是法人。法人本来就已经实现了财产的特定化，并且经过特定化的用于生产、经营等的法人财产同时也是其责任财产。这样，赋予智能机器（人）以法律人格，在实际效果上就是将法人的财产加以进一步分割和特定化，从而限制为其设计或制造的智能机器（人）的致损事故承担的责任。这就涉及到下面将要论及的法人与其他法律主体之间的关系问题。

最后，法人的法律主体地位导致了法人与其成员的关系以及法人与法人以外的其他法律主体之间的关系这样两类需要法律予以关注和解决的问题。

法人与其成员的关系引发了法律上关于成员资格、成员权等的相关制度构建；而法人与其他法

律主体的关系，则主要涉及劳动者、消费者以及法人的一般债权人这三类法律主体。在法人与其一般债权人的关系上，问题的核心在于法人是否仅以自己的责任财产承担责任，换句话说，法人的成员是否仅承担有限责任。虽然有限责任制度在近代以来的经济发展过程中发挥了巨大的作用，并且至今仍然占据主导地位，但是只有在可以与法人的债权人的利益达成大致平衡的情况下，才是可持续的。并且，经济社会发展到今天，有些公司的股东主动或被动地承担补充出资责任，承诺在公司资不抵债的情况下对公司的债务负责，这是我们在思考有限责任的合理性时所必须注意到的。

智能机器（人）并非组织，无所谓成员问题，因此考虑赋予其法律人格，主要就涉及与其可能的债权人之间的关系。目前来看，智能机器（人）充当经营者或从事其他商事活动还只是人们的想象，其涉及的主要对外关系就是与其致损事故的受害人之间的损害赔偿责任问题。智能机器（人）若成为法律主体，就意味着其损害赔偿责任应由其自己承担，而智能机器（人）的制造商或设计者则可以类比法人的有限责任而在原则上无须担责。若果真采取这样的制度设计，固然有利于鼓励更多的公司致力于智能机器（人）的设计与制造，使得他们不用担心被人工智能技术可能存在的不可预见的风险引发的损害赔偿责任所击垮，但是对于智能机器（人）致损事故的受害人而言，却是极不公平的，因为他们将可能承担无端的、自己根本无法预防和控制的风险。

可以看出，目前学界普遍承认从功利的视角看待法人。法人之权利能力，是对近代以来社会生产方式和社会组织方式的法学构建。在某种意义上，权利能力一方面解放了自然人，另一方面又将其禁锢于法人之中。回到智能机器（人）的法律地位问题，若是其主体地位不能通过与自然人的类比而得到承认，那么与法人相类比的结果，就是要求回答下面的问题：从工具理性的角度出发，为什么要将其拟制为人？

为何承认智能机器（人）是“人”

是有待回答的问题

从功利的角度出发，是否需要为智能机器（人）拟制主体地位，取决于这种拟制的目的是否正当、手段是否合适，即是否符合工具理性的要求。

若是仅服务于限制智能机器（人）的制造商或设计者的责任的目的，不构成为拟制法律人格的正当目的。从手段的角度来讲，作为法人的制造商或设计者正处于便利的地位，可以通过产品定价等机制在全社会范围内分散新技术应用所可能带来的风险。另一方面，他们还可以通过保险机制进一步分散风险。因此，即便仅从手段的角度来讲，也不必采取拟制主体地位的方式来达到本来已经可以达到的目的。

但是，也应该看到，技术和产业的发展日新月异，我们今天想象不到的为智能机器（人）拟制法律人格的必要性，或许在将来的某一天会凸显出来。但是，既然这是从功利的角度出发考虑是否需要拟制法律人格，并非出于智能机器（人）在伦理上的应然地位，这种必要性就需要有实践的基础并得到充分的论证，并且，这种论证责任应该在主张赋予其拟制法律主体地位的一方。

在人工智能产业发展如火如荼的今天，关于智能机器的法律地位的探讨，可能被某些产业界人士诟病为不利于产业的发展。然而，产业发展与事故受害人的救济应该是并行不悖的。这也是众多的自动驾驶汽车设计者或者制造商主动声明愿意承担责任的原因所在。沃尔沃总裁宣称将“对其自动驾驶模式下汽车造成的损失承担全部责任”；谷歌和戴姆勒也都提出，如果他们的技术有缺陷，他们将承担责任。因此，产业的发展只有在能够平衡产业内外不同主体之间的利益关系尤其是在高度尊重自然人的生命、身体、健康等伦理价值的情况下，才是可持续的。

（作者单位：中国社科院法学所）

(责编：毕磊、孙红丽)

相关新闻

上海张江建设“机器人谷” 推动智能机器人走进生产...

司法部发布全国民事行政法律援助服务规范

“人工智能军事应用法律问题学术研讨会”在京举行

机器人进入高速发展期 专家称未来或将与人共融

客户端下载

人民日报 | 人民网+ | 手机人民网 | 领导留言板 | 人民视频 | 人民智云 | 人民智作

评论 分享 关注

返回顶部

解正山：对机器人“法律人格论”的质疑

2020
09/23
17:49

北大法律网
企鹅号

分享

评论

【副标题】兼论机器人致害民事责任

【作者】解正山（上海对外经贸大学法学院副教授）

【来源】北大法宝法学期刊库《暨南学报（哲学社会科学版）》 2020年第8期（文末附本期期刊法学要目）。因篇幅较长，已略去原文注释。

内容提要：快速发展的人工智能技术使机器人越发具有自主性并呈现出某些道德或法律的能力。鉴此，不少法律学者主张授予这些人工智能实体法律人格以使之成为法律主体。然而，这种实用主义的法律立场忽视了智能机器人与自然人以及法人间的本质差异，赋予它们所谓的权利无论是在道德层面还是在法律层面上都是值得商榷的，主张机器人承担自身“行为”责任更是不可能。相反，将智能机器人视为“产品”并通过优化现有责任法框架来合理分配机器人“行为”责任才是现实的选择。

关键词：人工智能；机器人；法律人格；产品责任

一

问题的引出

现今，人工智能已成社会热议的高频词汇并成为国际竞争的新焦点，但它并非法律概念甚至没有统一的行业定义，因为准确定义人工智能甚为困难。这不在于“人工”概念而在于“智能”一词的模糊性，因为后者的定义相当宽泛并同人类自身都难以界定的一些特质，如自我意识、学习能力、推理能力等相互关联。不过，仍可从以下几个方面来理解人工智能，即“像人一样行为”、“像人一样思考”。随着技术进步，这些高度自主的人工智能实体（如自动驾驶汽车、护理机器人、医疗机器人等，下称“智能机器人”或“机器人”）完全可脱离人类之手并基于自身与周围环境的互动与分析而独立地为或不为一定的行为。此所谓“机器的范式”，即在没有人类参与或干预情况下自行“感知-思考-行动”。即将到来的人工智能革命或将为人类的生产与生活带来无限可能。实践中，包括中国在内的许多国家正如火如荼地进行自动驾驶汽车商用前测试，机器人投资顾问则已开始帮助金融消费者作出投资决策，机器人“棋手”甚至还多次赢得与人类棋手的较量。于是，民法学者发出了“机器人究竟应当作为法律关系的主体还是客体出现？应当将其等同于普通的机器或者动物对待，抑或将其作为‘人’来看待”之问。这一问题的回答关系到机器人进入人类生活后的法律地位，尤其是机器人权利及其“行为”责任问题。

通过对智能机器人自主性、社会性以及权利主体历史演变等方面的论证，越来越多的学者（国内研究者如许中缘、彭诚信、袁曾、张玉洁、郭少飞等，国外研究者如Filipe M. Alexandre、Paulius Cerka、Ryan Calo、Colin R. Davies、Hutan Ashrafian、Samir Chopra等）认为，应授予这些机器人法律人格或主张将其视为法律主体。然而，授予机器人法律人格终将“充满挑战与困难。毕竟，法律是——并且始终是——由人类制定并服务于人类的。想想那些基础概念，比如‘人格’与‘法律人格’。历史上，这些概念都与人类有关”。而且，授予机器人法律人格并非解决它们“行为”责任的灵丹妙药，反而可能打

200亿美元！IBM拟未来10年向美国科技产业投资

开“潘多拉魔盒”。因为享有法律人格的机器人很可能会成为一个“坏主人”而非“好仆人”，它们或将成为某些人推卸责任的替罪羊，人们将因此陷入更加危险的境地，现有的法律威慑体系恐被消解。那么，到底是否应授予机器人法律人格以便其享有权利、负担义务，从而解决高度自主的机器人造成损害时的道德与法律责任问题？换言之，将机器人视为法律意义上的“人”是否是框定其行为及后果，尤其是责任承担的必然选择？这些问题不仅关乎“人”（自然人，下同）与机器人关系的法律确认，也涉及机器人法律地位与“行为”责任，是人工智能技术演进中应予解决的难题。

二

为什么不应授予机器人法律人格

本质上，法律人格的授予是主体间相互尊重彼此权利并对侵害权利行为加以法律控制的一种手段，其与权利授予以及责任承担或义务履行密不可分。正所谓“每个人都有自己的生活，且都有责任去创造生活……这种责任与我们传统中承认的权利相匹配，它就是……自己能够定义人格或定义生活的权利……这些权利与更普遍的责任共存”。显而易见，法律人格的内核就是享受权利、负担义务的能力，其具有如下属性或本质要求：其一，能成为法律上的主体，具有相应的智力与理性；其二，成为法律主体使之有能力行使权利并能切实理解、遵守其所承担的法律义务，且对惩罚或制裁具有敏感性；其三，享受权利需要其意识到自己应得的权利以及他人应履行的义务。基于上述人格之法理，下文将尝试回答：机器人能否成为法律上权利与义务的主体，即机器人是否具有成为权利和义务潜在拥有者的能力，包括为行使权利、承担义务而创建、修改或废除法律关系的能力？

（一）机器人不可能成为权利主体

上述“法律人格论”者之所以主张赋予机器人法律人格，主要基于以下几项理据：（1）人工智能技术已经或将要使机器人呈现一些所谓道德的或法律的行为能力，具备与其他法律主体直接互动的能力以及自主决策能力，且能从自身经验中学习，智能机器人变得越自主就越不应将其视为工具而应被视为独立于其所有人或使用人的法律实体，它们“须以权利义务的承担者身份参与到民事法律关系中才能解决法律工具主义观下责任归责的困局”；（2）从社会性角度看，具有人形且自主的社会型机器人已能与人类进行交流与互动，随着人工智能技术不断取得突破，它们的社会能力将会不断提升，社会行为模式与情感交流将愈发突出，在社会化应用过程中逐渐扩张机器人的权利空间是这些智能实体实用主义功能催生的人类自我让步，而社会及经济上的便利性恰恰对决定是否将法律人格授予某些类别的实体发挥着至关重要的作用；（3）人类法律既然能把法律人格授予人类婴孩以及丧失心智的成年人，而且也认可既无“身体”也无“心智”的公司或庙宇成为法人，那么就没有什么可以成为机器人享有法律人格的阻碍因素。“法律人格论”者还从权利主体历史演化与发展逻辑中找到了机器人享有权利的正当性，并在法律主体历史演变中发现了确立人工智能主体地位的制度空间。总之，在“法律人格论”者看来，人类正处在迎接一个全新智能族群的当口，不管此种智能是否是人工的，都不应阻止这些数字平民获得尊严与权利，且应“表达出自然人类对人工智能体拥有同情与关爱之良善本性”。因此，“当机器人社会化应用达到一定阶段，人类必须对此作出必要回应……而回应的必要措施就是通过立法的方式肯定机器人的法律主体地位，并赋予其一定的权利”。

尽管“法律人格论”者言之凿凿，且现代人工智能技术催生的智能机器人具有了某些与人类相似的特征，但这些机器人根本不同于具有生命的自然人，也有别于具有独立意志并作为自然人集合体的法人等非“人”实体。因而，无论在道德层面还是在法律层面，承认机器人的权利“能力”并赋予其所谓的权利都是值得商榷的。

首先，机器人与“人”具有本质差异，它们不可能成为“理性的人”，既无法区分善与恶、对与错、好与坏等复杂的道德观念，也不要指望它们拥有难能可贵的同情心。毕竟，“人代表……一种具有理性思考与反省能力的智慧生物”，只有“人”才具有灵魂、目的性、意识、感情、利益以及自由意志。正因如此，“只有人类能够触及那些不可思议的事物，他们辨别着、选择着、判断着……”。独特的思考能力、自我意识、自省能力以及自尊等特质奠定了人类享有某种程度人格的哲学基础，意味着每个人与生俱来地拥有相同或同等的权利“能力”，他们平等地拥有生命权、自由权、财产权以及追求幸福的权利等自然权利。这表明，人格对于人类而言是自由的、平等的、包容的。然而，较之于“人”，机器智能充其量是一种演绎智慧，只能对数据进行无理性、无情感的枯燥处理，无论其具有多强的感知能力、深度学习能力以及由此形成的自主性都不过是对人类某些认知行为的模拟，根本谈不上运用自由意志以及所谓的道德知识在道德两难中做出正确的选择。总之，机器人不可能与人类签订互相赋予权利的契约，而且，其作为人类工具的原初地位无法改变，并且没有不容破解的内心秘密。

进一步而言，基于人类自身利益，人工智能技术发展的禁区必然要避免机器人成为一个具有“自我”意识的存在。因为一旦“人工智能有了‘我’的概念和意识，不仅是对人的模拟，而且也具有了人的核心内核。在这个层面上而言，人工智能就在个体上可以成为另一个物种的‘人’”，此时，它将思考自身存在的意义，而人类恐将“沦落为技术‘眼中’的他者”，人类的生存将因此而受到实质性威胁。可见，无论机器人多么智能，其限度必然是不得等同或超越人类地位，因为一旦对人类主体地位构成威胁，它们就会失去存在的正当性与合法性。另外，具备人格的一个本质要求便是拥有“自我”意识，这个“自我”不仅是个可识别的实体，而且还能根据“生活计划”进行创造性的自我定义。然而，人工“主体”缺乏内在情感，不具备包含了理解自我以及感知他人感受的“自我认识智能”。况且，智能机器人虽具备了形式逻辑能力，但并不具备辩证思维能力。正所谓“没有理性的东西只具有一种相对的价值，只能作为手段，因而叫作‘物’；而有理性的生灵才叫作‘人’，因为人依其本质即为目的本身，而不能仅仅作为手段来使用”。总之，如果无法成为一个有“生活计划”的自我意识实体且以某种方式“关心”这一计划的实现，那么承认机器人某种程度的人格就缺乏必要的前提与基础，它们自然难以由道德无涉者演变为道德或权利主体。

其次，以公司法人这一非“人”实体具有享受权利、承担义务的主体资格为理据论证智能机器人应获得同等法律待遇也是站不住脚的，这一论据不仅忽视了公司法人与智能机器人的差异性，而且没有意识到机器人“权利意识”觉醒带来的风险。

一方面，公司虽在形式上具有独立人格，但事实上并未真正脱离“人”的控制或干预，其“意思能力”很大程度上取决于“人”的意思能力：无论是公司的意思形成机关还是意思表示机关，其背后都是“人”而非公司本身。例如，董事会决议是由自然人担任的董事共同作出，即便存在法人股东，股东会决议也都可以最终穿透至自然人股东的意思，一旦公司“意思”形成，最终是由“人”所担任的法定代表人向交易相对方作出而非公司自身。总之，公司很大程度上仅是个抽象的存在，其“并非真正自主，因为它的行为由其股东、董事、经理等利益相关者而非‘自身’决定，它的‘意志’总是其股东、董事、经理等利益相关者的‘意志’”。公司人格的正当性在于公司的权利与义务/责任实际上就是拥有、管理它的“人”的权利与义务。这意味着承认公司法人的独立人格不致危及“人”的主体地位。不可否认，把法律人格延及公司这种非“人”实体反映了立法者乐于为商业或技术创新提供相应的法律保护，同时也表明法律人格是一个发展且开放性的概念。然而，即便法律人格是一个发展的概念，也并不意味可以无限扩展。否则，只要某一事物具备一定的人格表征就主张赋予其法律人格或权利，那么到最后权利客体恐怕都将不复存在。无权利客体，权利本身自然就成了空中楼阁。

另一方面，鉴于公司“没有可被诅咒的灵魂，亦无能被踢打的身体”，因此，在本质上，公司法人的所谓自主性是虚拟的，而智能机器人自主性往往是真实的。这种自主性正是“法律人格论”者认为机器人应被赋予法律人格或权利的重要论据。然而，由于机器人具有全部或部分独立于人类的自主性，所以，一旦在法律上强化它们的独立地位，恐怕将无法确定它们如何表达使用人的意志以及谁能控制它们。鉴于此，越来越多的学者、科技企业家以及未来学家警示道，更强大的人工智能可能会抗拒人类对其行为的监管，从而给人类带来灾难甚或生存危机，此种风险并非源于智能系统对人类的恶意或无法理解人类的主观意图，而是源于它们对人类主观意图根本就漠不关心。因此，对于人工智能系统，应非常小心地向其索取，否则人类将会看到原本设计的实用功能可能并不是它字面含义那么简单。就此而言，不仅不应在法律上强化机器人的独立地位，相反，还应从严监管这些人工智能系统以免其危及人类的安全与秩序。说到底，无论人工智能系统多么复杂、智慧多么高级，它们终究只是作为“工具”而存在，反映的是人类技术能力的进步。那种根据“实力界定权利”理论所得出的“机器人与人类之间的实力变化可能催生出机器人权利”的观点，显然是把机器人与人类置于博弈的场域中。无论如何，机器人都不可能也不应该脱离于人类独立地作为“类人”而存在。

或许是意识到机器人与“人”或法人之间的本质差异，“法律人格论”者转而借用动物权利之论述来论证机器人享有权利的正当性。他们认为，无论是在心理上还是在哲学层面，动物与机器人都具有相似性，人类既然能承担避免让动物产生痛苦的道德义务且能对某些动物固有的尊严表示尊重，那么，也可同样对待拟人化的机器人——赋予它们权利且不虐待它们。不可否认，人类行使对自然的统治权既需要对动物进行道德关怀，又需要谨慎关注人类自身利益。鉴于人类自身利益与动物和环境的利益交织在一起，因此，为保护濒危物种和人类赖以生存的生态系统的可持续性，立法者制定了一些保护动物的法律。但这并非意味人类立法者赋予濒危动物或其他物种某些权利，而是因为生物多样性与生态系统稳定关乎人类自身的可持续性发展，立法者系基于人类自身权利或利益之考虑而对特定物种施加保护。很大程度上，要求善待动物不是因为它们享有权利而是因为人类的同理心——动物的痛苦可能让人类感同身受。但这种同理心并不足以让动物们拥有人格或享有权利，相反，人们仍可自由地购买、出售自己的宠物。可见，所谓的动物权利充其量只是一种道德倡导或道德层面的论述。更重要的是，动物无法与人类进行自我意识之间的复杂互动，正因这种“能力”差异，无论人类如何拟人化动物或对动物充满爱意，但在法律权利方面，动物通常只能被视为一个“物”。而且，因动物缺乏“善的观念”和“正义感”，且无法满足道德人格和政治公民的要求，所以人类对待动物的态度通常并不涉及正义问题。总之，所谓的动物权利更多是人类基于自身利益或同理心而给予动物某种程度的道德关怀，它们在法律上的客体地位不曾有过改变。机器人又何尝不是如此！

进一步而言，如果机器人能够享有法律权利，那么它们可享有什么样的权利？生存权？财产权？与人类通婚的权利？休息的权利？被认定为受害人的权利？保护它的运行系统免遭无端搜查与扣留的权利？请求不被拆解或免遭终止电力供应等惩罚的权利？通常，权利的概念根植于人类的道德世界。这决定了权利概念及其体系与“人”的密切关联性。不难想象，人类立法者承认机器人享有原本专属于“人”的那些权利时将遭遇怎样的挑战！或因如此，主张赋予机器人权利的学者并未在现有权利框架中讨论，转而强调以法律拟制手段“强制性要求人们之间达成‘机器人拥有权利’的基本共识”，赋予机器人拟制性且利他的所谓新型权利。然而，这些拟制性且利他的权利要么是不现实的，要么是多余的。所谓“基于功能约束的自由权”，更多是指向机器人在产品意义上所能发挥的功能——机动性，它怎么可能成为一项独立存在的权利呢？根本没有必要把这种产品意义上的机动性上升到权利的高度，只要此种机动性不致对人类构成显著的或无法控制的危险，立法者完全可允许机器人发挥其机动

性。至于“获得法律救济的权利”完全可在机器人作为“物”的情形下给予权利主体——所有者或使用人——相应的救济即可，而非一定要把机器人视为法律主体方可为之。

（二）机器人履行义务/承担责任的非现实性

正如我们所知，不存在纯粹的权利或义务，任何主体享有权利的同时也应负担义务。换言之，对于一个完全自主的法律主体而言，其不仅有“能力”意识到自己行为的意义，而且能对其行为负责。因为，如果一个人在社群中与其他人共存，就应以负责任的态度与社群中的其他成员进行互动，承认彼此的权利且须为侵犯这一权利而承担相应的责任。那么，具有类似于人类自主性的智能机器人拥有这样的“能力”吗？不可否认，机器人若能为其行为负责，那么确有必要重新审视现有的法律人格之概念并对现有法律体系是否足以适应新的法律现实进行评估。但问题是，机器人具有履行义务/承担责任的能力吗？或者说，它们能够拥有财产进而具备履行义务/承担责任的物质基础吗？如果可以，那么它们拥有或占有财产的表征方式是什么？如果不能，赋予它们法律人格还有意义吗？

一般认为，公司等非“人”实体获得人格不在于它们的伦理性，很大程度上是因它们财产与责任的独立性。鉴于此，有论者提出可通过法律拟制让人工智能系统像公司那样，不仅能拥有财产而且还可作为被告应诉，自己承担责任。果真如此，那么机器人的生产者（包括设计者、编程者等当事人，下同）、所有者或使用人（下称“用户”）自然无须再为机器人的行为承担责任了，除非他们“出资”不足或其他原因需“刺破人工智能面纱”。可问题是，谁负责“出资”以构成机器人的财产与责任基础？需要设定最低的财产额度吗？“出资”人是否以出资额为限承担有限责任？即便不考虑立法技术上的挑战，仅从生产者或用户的立场看，这也是不现实的。因为要求他们履行“出资”义务势必增加他们的经济成本，而且也不具操作性。如此做法既不经济也不方便，看似解决了机器人责任承担问题，实则徒增各方负担，包括立法上的成本。额外的成本负担不仅会打击人工智能开发者的创新热情，而且也将浇灭用户使用人工智能的兴趣，没有市场需求的支撑，人工智能技术的开发者自然就会失去技术创新的动力。进一步而言，如果机器人真的具有了权利意识而且具备了承担责任的基础，那么这对“人”而言到底是祸是福？人类准备好接受一个拥有法律人格且具有公司无法比拟的自主性的新角色了吗？

从刑事责任能力的角度观察，认为机器人能为其自主“行为”负责更是不可能。较之于公司，绝大多数机器人因具有相应的物理形态从而显得比公司更实在，加之有人工智能技术的支撑，它们也都具有一定的意志能力。有学者据此认为，“智能机器人的意志自由程度比单位更强”，它们“完全可能脱离人类产生独立的意志。对于人类来说，智能机器人是人类模拟自身神经网络创造出来的事物，比世界上任何事物更类似于人，刑法应当尊重其自由意志，适时考虑赋予其法律上的主体资格”;而且，唯有“将其作为行为主体与社会成员看待，有罪必罚，并经由刑事诉讼程序进行法律上的审判，才能在智能机器人的‘大脑’中建立价值体系，植入人类文明，使其自觉遵守并维护社会秩序，从而实现‘人工人’向‘社会人’的转变”。笔者以为，赋予机器人刑事责任主体地位的构想显然是天方夜谭。

首先，如上文所言，机器人很难对善恶、好坏进行道德判断，既无力在是非、对错之间进行抉择，也不可能理解法律禁止的以及法律允许的行为的真正意涵。其次，较之于智能机器人越来越高的自主性，公司的意志很大程度上仍是“人”的意志，它们在本质上仍是虚拟的主体，其背后的“人”才是真正的主宰者。因此，公司犯罪时，除可处以罚金外，立法者还把刑事责任延伸至其背后的“人”。就此而言，对公司的惩罚实质上是对其背后“人”的惩罚，这不仅使惩罚公司成为可能，而且能防止公司自身产生不可控风险。相反，当机器人意志自由程度等同甚或超越“人”或公司的意志时，那么此种失控物可能会对人类利益造成损害甚至威胁到“人”的主体地位。最后，如果承认机器人具有刑事责任能力或主体地位，那

么自然可对其错误行为处以罚金或徒刑甚至直接将其销毁。但问题是，此等刑罚是否有意义？且不说机器人能否像公司那样因拥有独立的财产从而可对其处以罚金，单就对机器人能否处以徒刑而言就是个极大疑问。能让机器人因失去自由而感到痛苦并因此痛改前非吗？经由审判或道德谴责，它们会对自己的错误行为感到羞愧吗？通常，对被告人处以刑罚意味着这些主体要承受社会、心理乃至身体上的不利后果。可是，所有这些对无法成为道德主体的机器人而言是毫无意义的。此时，刑法的惩罚与矫正功能如何实现？除非机器人与其生产者或用户的关系如同公司与其实际控制人之关系，否则，法院难以对机器人错误“行为”施以有效制裁。如果不能对机器人的错误选择与行动施加有效的刑事制裁，那么就更不用说让其他机器人“感知”并“意识”到它们未来作出相同或类似错误“行为”的风险所在了，而这恰恰关系到刑法威慑功能对机器人“种群”的实际效用。总之，机器人不可能具有所谓的“感知刑罚痛苦的能力”，更不可能在“犯罪的‘快感’与刑罚的‘痛感’之间进行理性权衡”。机器人真若有此“能力”，那么立法者不仅不应承认它们的自由意志，还应毫不犹豫地阻绝这一技术发展趋势。很难想象，人类如何与“社会化”的机器人相处！

既然难以像法人等非“人”主体那样具备责任能力，那么可否退一步，赋予机器人类似于人类婴孩般的法律地位？正如我们所知，即便是尚不具备完全意志能力的未成年人或暂时丧失自由意志的精神病人以及医学上的植物人仍都具有法律人格。承认这些特殊主体具有法律人格的根本原因在于：他们是“我们中的一员”，虽然认知能力存在不足或障碍，但并不妨碍他们与认知能力正常的人一样享有尊严并获得尊重，而且，他们都具有培养完全人格的潜力。总之，人类的这些特殊成员享有“不同但却平等”的人格。然而，机器人不可能也不应该获得这般地位。首先，较之于人类上述特殊成员，机器人不可能具有培养完全人格的潜力，即便机器人未来在强人工智能技术支撑下具有或超过人类普通成员那般的认知能力，其作为手段或工具而存在的本质规定性也不可能逆转。更重要的是，认知能力状态虽对“人”享有何种权利、承担何种责任具有决定性影响，但对其是否具有法律上承认的人格却几乎不起作用，法律人格与法律权利的基础更多在于人性而非认知能力。否则，不仅可把法律人格授予机器人，甚至也可让黑猩猩等“聪明”的动物享有人格，因为它们的自主性或认知能力有时比上述人类特殊成员健全。显然，这种逻辑站不住脚，而且将贬损人类尤其是那些处于弱势地位的人类成员的尊严。其次，如果把机器人视为民法上的未成年人，那么机器人可能就得接受民法上对未成年人行为能力的区分与限制。随之而来的问题，如何区隔机器人的行为“能力”？它们将如何自主地为或不为一定的行为？解决这些问题必将带来高昂的立法成本。

或因意识到以法律人格或主体地位为前提建构责任框架的非现实性，所以，“法律人格论”者在论证机器人等人工智能系统如何承担责任时，又不自觉地回到传统的法律体系中寻求答案而非站在授予机器人法律人格之立场来构建其责任框架。例如，建议通过强化产品责任、要求机器人生产者或用户承担严格责任。不难看出，这些论者弱化甚至是放弃了围绕法律人格来构建机器人民事责任框架的努力，他们把本应能独立承担责任的法律人格之假定与非人格范畴的产品责任糅合在一起，从而导致论证上的逻辑不自洽。因为，适用产品责任的逻辑前提是把机器人视为“物”，责任承担者自然是生产者等法律主体而非产品本身，但如果将机器人视为主体，无论它们的人格是有限的还是完全的，都不可能再落入“物”或客体的范畴。既然站在“物”的立场来构建机器人的责任框架，还有什么必要授予机器人法律人格或主体地位呢？不可否认，许可与产品责任等既有监管与责任框架适用于机器人时存在不充分性，也有必要对现有框架予以优化以应对人工智能产品责任风险与挑战，但这并不代表要建构一个以主体为前提的责任体系。

综上，机器人或能具有一定的自主性以及与环境进行交互的能力，但在根本上，它们不可能具备人类那般的理性，缺乏像“人”那样的自我意识、辩证思考。况且，某种程度上，人格

还与某一实体为自身制定目标以及实现这一目标的能力紧密相连，而机器人尚不具备这种能力。即便这些人工智能实体具有某种“目标”，顶多也是为实现生产者或用户的目标而衍生出来的，且它们根本不“关心”这一目标能否实现。机器人能否具备公平和公正的价值观更是值得怀疑。通常，只有“人”具备公平和公正的意识，才能真正地理解权利与责任的意涵。确实，人类通过编程可使机器人“符合”规则，但绝非“遵守”规则，因为对规则的“遵守”预设了人们对规则意涵的理解，但机器人并没有这种理解能力。正如霍斯特·艾丹米勒所言：“某一特定社会的法律以及赋予该特定社会成员的权利义务是‘人类境况’的表达。法律……反映了我们认为处于人类核心的东西，以及处于核心的人何以为人的意义……如果将法律人格赋予机器……让它们有权取得财产、订立合同，这简直会使这个世界非人化”。总之，在促进科技进步的同时，更应确保“人”的主体地位、自由和人格尊严不受侵害，树立对尊重自然人人格的民法底线。说到底，机器人仍在民法“物”的范畴之内，虽具备了部分类似于“人”的功能或行为，本质上仍是“物”之属性的产品，它们的所谓“智能”体现的仍是人类技术能力本身。毕竟，人工智能技术的根本出发点在于：提高技术工具为人类预设目标服务的能力而非令其成为“我们”当中的一员。如果说人工智能技术使机器实体具备了某些道德的或法律的能力，那么充其量也只是它们提高了服务人类生产、生活的能力，而不能把这样的智能表现理解为构成了主体资格的一般要件。总之，人工智能技术催生的机器自主决策无法等同于社会规范中智能机器人的“主体性”。至于“法律人格论”者担心的机器人“责任归责之困局”，完全可通过将这些人工智能体界定为产品加以解决，在法律上为机器人行为负责的总是人类自身。

三

产品视角下机器人的“行为”责任

（一）机器人致害“行为”的归责难题

如上文所言，将机器人定位为民法上的“物”且将其视为产品并据以确定其民事责任应是更现实的选择。一旦将机器人视为产品或人类生活的一种工具，那么，当需对其法律或道德上的责任进行评价时，焦点自然应将转向制造或使用它的当事人。

类型上，智能机器人引发致害事故无非两种情形：一是由警示缺陷（未充分警示消费者如何使用产品或应予注意的风险）、设计缺陷（产品的可预见风险本可通过合理的替代设计而被避免或被降低）、制造缺陷（产品偏离了设计意图、未根据既有规格进行生产）造成的，用户未尽合理注意义务亦可能产生致害事故；二是机器人自主决策模式下的致害行为，即致害事故不是由警示缺陷、制造缺陷或设计缺陷等产品缺陷以及用户未尽合理注意义务引起的，而是机器人基于自我学习所带来的意外副产品。

第一种情形下，鉴于与机器人有关的责任都是人为错误的结果，因此，自可适用现有产品责任与侵权责任框架，要求生产者承担产品责任或由用户承担侵权责任。产品责任在法律属性上属无过错的严格责任，它要求当事人应尽最大努力阻止伤害。这是立法者强行分配的一种责任承担形式，免除了受害人对责任方是否存在过错的证明责任，受害人仅需证明产品存在缺陷且缺陷与损害之间存在因果关系即可，生产者是否采取适当的措施或是否存有过错在所不问。理由在于：责任方往往获利最多并控制着设计生产过程，且大多数情况下的损害是由“错误”导致的。本质上，产品缺陷系因设计生产者对可预见风险的疏忽而产生，这种过错自然构成他们承担严格责任的正当基础。这也意味着，若产品缺陷及其导致的损害是可以证明的，那么，责任归属不会因机器人而产生额外问题。只是人工智能时代受害人或将面临如下挑战：机器人所依赖的“算法”越复杂，受害人就越难证明其中的特定错误以及错误与损害之间的因果关系，因为“算法”对普通人而言就是一个“黑箱”，要求他们破解算法黑

箱或证明存在合理的替代设计，难度可想而知！有时甚至连技术创新者都无法解释致害事故的真正原因，更不用说让受害人去破解机器人的算法黑箱进而证明机器人是否存在缺陷以及该缺陷与损害之间的因果关系！此外，生产者利用责任豁免条款——产品投入流通时引起损害的缺陷尚不存在或当时的科技水平尚不能发现缺陷存在——进行抗辩更将增加受害人的证明难度。

第二种情形下的归责问题更具挑战性。具有一定自主性与学习能力是智能机器人最突出特征之一，但也导致了可预见性难题。因为它们在本质上并不受提前预置的概念、经验法则、传统智慧等要素的限制，而人类则需依靠这些要素才能进行决策，这意味着智能机器人可能做出人类无法预见的行为。虽然此种情形目前仍限于较小范围，但随着机器“学习”能力的增强，越发自主的机器人的应用范围将更加广泛，从而带来棘手的归责难题，包括过错认定与责任分配。如果说上述第一种情形下，缺陷以及缺陷与损害之间的因果关系尚有可能证成，那么第二种情形下，甚至连机器人本身是否存在缺陷都是一个极大的疑问。例如，完全自主的无人驾驶汽车根据自我学习而自行做出的行为以及它与其他智能驾驶车辆或其所处环境进行互动与协调而作出的行为所导致的致害事故就会使损害赔偿趋向复杂化。此时，要求受害人根据既有责任框架证明导致其受损的原因或产品存在缺陷就存在极大困难。鉴于这些自主行为往往是机器学习的结果，因此，即使是最谨慎的生产者恐怕也难以预见那些已具备了与现实世界互动能力的智能机器人所导致的风险，不管这种结果是否是设计生产者本身所期望的，它都将给现有责任框架带来挑战。一方面，在技术层面上，机器人根据自身经历而做出的自主行为及其风险已非生产者所能预见或控制，而在现有产品责任框架中，他们仅对可预见性风险负有警示义务或合理注意义务。此种情形下，机器人的生产者或用户能否以机器人自主行为不受其控制而请求豁免承担责任？传统侵权责任归责时惯常适用的因果关系链条是否会因高机器人的高度自主行为而被打破？毕竟，普遍接受的观点是，要求一个人为他无法控制的人所犯下的错误行为负责是不公平、不公正的。另一方面，若豁免机器人的生产者等当事人此种情形下的责任，又该如何向受害人提供救济？为没有过错但遭受损害的个体提供救济本就是一项重要的价值，因此，让个体承担损失尤其是在因果关系无法解释的情形下要求其自行承担损失显然是与公平、分配正义以及风险分担等基本的社会观念相悖的。

对此，最需解决的问题是，当“不得伤害人类或以不作为之方式放任人类被伤害”这一律法已被植入机器人“思维”系统之中，且有证据表明机器人是因自己的学习能力而违反了这一律法时，是否仍可要求生产者承担责任？真若如此，或意味着他们在任何事情出错时都可能要承担责任！此时，生产者等当事方承担责任的依据是什么？是否可因他们处于更有利的经济地位从而要求他们承担责任？或者，机器人的每个“错误”是否都是法律意义上的缺陷造成的？一般认为，具有自适应性与自我学习能力的机器人可自由地与人类或周围环境互动，它们能以不可预知的方式对新感知到的信息做出反应。鉴于此，若机器人的自主行为对第三方造成损害，那么将很难认定机器人存在法律意义上的缺陷，因为它做了应该做的：对新输入的信息（变量）做出反应并调整自己的行为。然而，侵权法的一般原则是，非因自己过错而遭受损害的一方不应自行承担损害成本。一旦机器人的自主行为致使他人受到损害，自然应有人为此负责。问题是，谁该为此负责？唯一可行的方案似乎是推导出某种新的理论，即机器人自主行为导致的事故本身就是缺陷的证明。而且，鉴于机器人的自主行为以及它们在现有产品意义上的缺陷均能致害，因此，除非将机器人自主行为本身视为“缺陷”，否则，很难划分两者之间的界限。如果这是正确的选择，那么，随之而来的问题是，法律应该如何在智能机器人的设计者、编程者、生产者以及其他参与方之间分配责任？换言之，如果一个真正的自主机器的行为被认为是造成某种损害的主要原因，那么，在何种程度上要求生产者等当事人承担责任才是公平合理的？

另外，机器人的生产者与用户间的责任分配也是亟待解决的法律难题。一方面，适用严格责

任意味着生产者应确保它们生产的机器人难以被用户重置，或确保它们不遵从危险性指令，包括将旨在阻止机器人伤害人类或其财产安全的措施嵌入产品之中，唯有如此才有可能限制或免除自己的责任；另一方面，若完全豁免终端用户责任，则可能错误地激励他们不当使用机器人并因此导致第三人人身或财产受到侵害。可见，控制终端用户行为层面上的过错也是有效责任机制的重要组成部分。只是，如此一来，终端用户在传统侵权法框架下的合理注意义务或将延至需要及时了解人工智能系统是否失灵之情形，而且在某种程度上还应知道如何处理以避免损及他人人身或财产。当然，终端用户履行这些义务离不开生产者等前端当事人警示义务的适当履行，两者密切关联。即便如此，这也可能加重终端用户的合理注意义务。终端用户责任风险的增加势将降低他们购买或使用人工智能产品的意愿，这种不利影响最终将传导至生产者并致使后者因没有市场需求而延缓甚至是放弃技术创新。可见，在生产者与终端用户之间分配责任时存在着明显的紧张关系，因而需要在法律政策上对责任负担做出合理安排。

（二）作为产品的机器人之民事责任分配

作为一般原则，智能机器人的生产者应为产品缺陷所致损害承担严格的产品责任。只是较之于普通产品，智能机器人具有高度复杂性，其所依赖的复杂“算法”或致使被侵权人难以证明机器人是否存在缺陷。而且，即便能够证明，受害人恐怕也得负担高昂的诉讼成本，譬如委托专业技术人员对产品缺陷进行证明。机器人脱离人类干预的自主行为的责任分配更会让法院左右为难。因此，作为回应：

第一，鉴于智能机器人为或不为一定行为取决于难以被普通人所理解的复杂“算法”，因此，首先应要求生产者等“算法”控制者遵从一套以人为本的伦理或法律准则，包括将人的保护置于最优先地位且应避免“算法”构成对不同人群的歧视等；其次，还应要求生产者能够解构“算法黑箱”并对依赖该“算法”而实施的行为或决策负责。例如，为方便查明致害原因到底是用户疏忽、技术失灵、机器人自主行为抑或其他外部因素所造成，可要求生产者在机器人体内置入“即时数据记录仪”，用以记录机器人运行状态及数据，为事故原因分析、因果关系证明、责任的合理分配提供客观、准确的证据支持。总之，机器人立法应促进机器“算法”的透明度与问责制，尤应要求生产者对智能机器人的运行逻辑进行解释。

第二，鉴于产品标准或规格通常是证明产品是否存在缺陷的重要依据，因此，未来立法可考虑创设或指定人工智能监管机构并由其负责制定机器人等人工智能产品的国家或行业技术标准（包括硬件安全、网络安全、公众知情乃至用户隐私保护等方面的具体标准），要求生产者确保机器人难以被终端用户重置或确保它们不遵从危险性指令，包括将旨在阻止机器人伤害人类或其财产的安全措施嵌入产品之中、警示用户何时应进行必要的维护、保养与检查以及应在何种情形下进行“人—机”切换等，不过，无论如何，不应不合理地增加终端用户的注意或监督义务，否则将有违人工智能系统是为人类提供某种便利而非提升人类操作水平或消除人为失误这一设计初衷。总之，包括用户在内的社会公众有权对机器人安全性能抱有期待，即机器人安全等级或标准至少应与非人工智能产品持平或在统计学意义上比其他非人工智能产品更安全，这也是生产者最低的法律义务。

第三，鉴于人工智能的高度复杂性，因此，可将机器人是否存在缺陷以及该缺陷与损害之间因果关系的证明责任分配给生产者。生产者证明产品缺陷存在与否的意义并不在于否定自己的责任，而是基于如下考虑：（1）若机器人存在产品缺陷，那么判令导致缺陷的那一方当事人最终承担责任自无异议；或当（2）致害事故非由产品缺陷或其他人为因素而是机器人自主行为造成的，那么生产者就应成为第一责任人，但其可要求参与机器人设计、编程、软件供应等环节的当事人共同担责。因为，一方面，一旦机器人进入自主决策模式，用户即不再控制其运行，此种情形下致害事故的最可能原因是智能系统本身的“失灵”而非用户过

错，除非用户实施了干扰、阻碍等错误行为，否则要求其承担责任显然是不公平的；另一方面，就产品责任而言，现有立法要求生产者先行承担责任；承担责任后，可向有过错的其他责任方行使追索权，但在机器人自主决策模式下，先行承担责任者可能于事后无法向其他责任方进行追索，因为他们似乎都无“过错”。鉴于此，除非生产者处于承担损失的最佳地位，否则，由参与智能机器系统设计、制造、维护等环节的所有当事人或那些能够更有效防范或避免损失的当事方之间分摊责任才是公平合理的。此种“共同责任”无须纠结于应把错误或不法行为的每一个细节归属于哪一方，一旦确定了责任，裁判者即可要求所有参与方共同承担责任——公平均摊责任或根据各自所获利益多寡分摊责任。这有助于解决智能自主机器对人类造成损害但将过错分配给特定一方是不可能的或不可行时的归责难题。

进一步而言，从技术层面上看，让机器人等人工智能系统拥有自主决策能力即便不是设计、生产等环节当事人的直接目的，至少也是他们的间接目的，因为只有这样才能赋予机器人与周围环境进行互动的能力。一定程度上，这是该等当事人期望发生的，更是机器人高度智能的体现。可见，机器人的设计生产者对机器人所谓的不可预见行为及其风险并非全然不知。至少，他们是在“放任”甚至是“鼓励”机器人做出这样的自主行为。而且，较之于个体，商业组织分散风险与吸收损失的能力更强，且它们将通过售卖智能机器人而获得不菲的利益。因此，要求这些拥有共同目标且都分享了利益的潜在责任方分摊机器人自主决策模式下的“行为”责任自然是公平合理的。更重要的是，如果可预见的损害成本小于因提供安全保障所花费的成本，那么生产者等当事人自然更愿承担损害成本而非花费力气提供安全保障。因此，唯有增强这些当事人的责任负担才能迫使他们把过错行为的成本内部化，确保更负责任的设计，尽力避免损害发生。不难看出，防止有害行为尤其是让受害人获得赔偿构成了上述“共同责任”的正当基础。

不可否认，要求设计、编程、生产等环节的当事人承担机器人自主行为的全部责任或将增加他们的成本，若他们因责任风险而趋于保守，那么很大可能会阻碍人工智能技术的运用或迫使技术开发者大幅降低机器人的功能或运行范围。这种担心不无道理，但事实上，这些当事人除可根据产品定价等方式转移部分成本外，还将因人工智能系统优异的安全性能大幅降低事故率进而使其总体责任成本显著降低。尽管如此，以下问题仍值得考虑：可否根据他们在技术应用方面预见能力的局限性或基于激励创新之考虑从而以某种方式减轻他们的责任？实际上，要求设计、生产等环节的当事人作为“一个整体”共同负担机器人自主决策模式的致害责任就是对这一问题的部分回应，至少是避免了某一个当事人承担全部责任，他们可以公平均摊责任或根据各自所获利益的多寡分摊责任。而且，作为对内部责任共担机制的补充或替代，还可考虑建立一种类似于自我保险的赔偿基金，由机器人设计、编程、生产等环节的当事人根据产品特性及风险因素共同筹资建立。当机器人非由产品缺陷或人为因素产生致害责任时，可由共同基金赔偿受害人损失。为激励技术创新、降低企业成本，政府也可考虑向该赔偿基金出资。

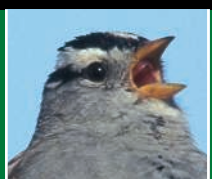
将保险安排引入智能机器人致害事故的责任框架之中也是一个值得考虑的方案。鉴于智能机器人具有一定的自我“感知—思考—行动”能力，因此，如上文所言，把它作为纯粹的产品增加了生产者或用户的不可预见风险。为此，一个理性的事前解决方案就是通过保险安排来分担或降低当事人的赔付风险。为确保受害人总能获得部分或全部的损害赔偿，同时也为彰显产品责任法的激励与矫正功能，可强制要求利害关系人为机器人可能的致害风险包括其自主行为致害风险购买保险。该保险安排的基本立场是，无论机器人导致损害事故时是否处于自主决策状态，均应确保受害人通过保险获得赔偿。此外，鉴于技术革新降低了用户对保险产品的需求，加之人工智能产品致害责任向生产者转移，保险模式可考虑从传统的以保护用户为中心为其操作失误风险提供保险服务转向以保护生产者为中心为其智能系统失灵风险提供保险保障。具体而言，首先，若机器人发生致害事故，保险人将作为第一责任人负责赔偿

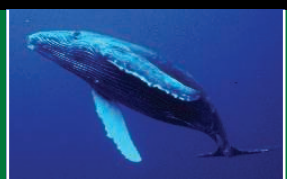
损失，但存在下列情形时保险人有权排除或限制自身赔偿责任：一是被保险人擅自更改机器人操作系统引起致害事故；二是被保险人未根据保单的要求更新或检查机器人操作系统因而引起致害事故。其次，受害人对产生的损害负有部分或全部责任时，保险人的赔偿责任可相应减少，于机器人未投保之情形，应投保的当事人将成为第一责任人，保险人无须承担责任。最后，若最终查明事故系机器人传统意义上的产品缺陷所致，则保险人有权向生产者行使追索权，非由人为因素引起的机器人失灵风险仍由保险人承担。

四

结语

与科学技术的激进性相比，法律具有明显的保守性。但是，面对快速发展的科学技术以及由其塑造的新的社会关系，立法者或裁判者应积极理解新技术，留意它的新风险并在此基础上作出回应。当初，克隆技术一经问世同样引起轰动，包括人们对克隆技术可能对人类伦理与生活秩序产生强烈冲击的担忧。最后，立法者的底线思维——禁止利用克隆技术“复制”另一个“我”——为这一技术革命设定了明确边界并确保其安全有序地发展。如今，人工智能技术的蓬勃发展使人类立法者再次面临挑战——是否允许一个在法律上作为主体而非客体的“类人”群体存在？笔者认为，无论机器人未来多么智能，它都不可能也不应该让其拥有“人”一样的自我意识与自身目的，其所发挥的仍应是智慧工具价值。尤不能忽视的是，“人之主体能力的本质与人所创造工具的工作能力表象之间”存在根本区别，这也是维持人类主体性的必然结果。而且，如上文所述，机器人也不同于作为自然人集合体的法人。总之，作为法律上的客体是智能机器人存在的合法性与正当性基础，以此为前提进一步优化机器人的民事责任框架才是更是现实的选择。





LETTERS

edited by Jennifer Sills

Robot Rights

IN HIS PERSPECTIVE ("THE ETHICAL FRONTIERS OF ROBOTICS," 19 DECEMBER 2008, P. 1800), N. Sharkey regrets that there are no international guidelines for robot use from the U.N. Convention on the Rights of Children. Although the international community has not officially released such a code of ethics, several sets of guidelines are under way, including Japan's Draft Guidelines to Secure the Safe Performance of Next Generation Robots by the Ministry of Economy, Trade, and Industry; the South Korean Robot Ethics Charter by the Ministry of Commerce, Industry, and Energy; and the guidelines for robots by the European Robotics Research Network (*1*–*3*).

These guidelines differ from each other when it comes to social problems, robot service, and robot political rights. Lack of consistency in ethical guidelines can be attributed to some degree to cultural differences in human beliefs and practices. Differences, for example, in the value placed on the development of independence in infants and toddlers could lead to totally divergent views of the use of robots as caregivers for children. Because different cultures may disagree on the most appropriate uses for robots, it is unrealistic and impractical to make an internationally unified code of ethics.

SHESEN GUO* AND GANZHOU ZHANG

Qianjiang College, Hangzhou Normal University, Hangzhou, Zhejiang 310012, People's Republic of China.

*To whom correspondence should be addressed. E-mail: guoshesen@126.com

References

1. K. Yoon-mi, *AI Magazine* **2**, 144 (2007).
2. L. Lewis, "The robots are running riot! Quick, bring out the red tape," *The Times*, 6 April 2007 (www.timesonline.co.uk/tol/news/world/asia/article1620558.ece).
3. R. J. Sawyer, *Science* **318**, 1037 (2007).

Adapting to Climate Change

IN THEIR PERSPECTIVE "PHYSIOLOGY AND climate change" (31 October 2008, p. 690), H. O. Pörtner and A. P. Farrell stress the importance of understanding organisms' physiological responses to climate change but fail to consider thermally induced phenotypic plasticity (*1*, *2*).

An organism with thermal plasticity can compensate for changing environmental temperatures. For example, some Antarctic fish can reverse the negative effects of rising water temperatures. *Pagothenia borchgrevinki* usually experience water temperatures of –1.8°C, but exposing the fish to +4°C for 4 weeks induces compensatory responses at the level of cellular metabolism, in the cardiovascular system, and in whole-animal swimming performance (*3*, *4*). These phenotypic changes allow *P. borchgrevinki* to function at +6°C, which is 8°C above their usual water temperature, far warmer than any projections of human-induced warming. There are many more examples of similar responses in temperate and tropical ectotherms (*1*, *2*, *5*).

We do not advocate that the potential effect of climate change be taken lightly. However, to respond adequately to climate change, it is crucial to consider the full breadth of physiological responses to temperature. The most appropriate approach is not to assume that all animals are specialized to their environment—which is clearly not the case—but to determine which animals have the capacity for phenotypic plasticity and then concentrate conservation efforts on those organisms that are most likely to be negatively affected.

CRAIG E. FRANKLIN[1]* AND FRANK SEEBACHER[2]

[1]School of Biological Sciences, The University of Queensland, Brisbane, QLD 4072, Australia. [2]Department of Biological Sciences, The University of Sydney, Sydney, NSW 2006, Australia.

*To whom correspondence should be addressed. E-mail: c.franklin@uq.edu.au

References

1. F. E. J. Fry, *Annu. Rev. Physiol.* **20**, 207 (1958).
2. H. Guderley, *Biol. Rev.* **79**, 409 (2004).
3. F. Seebacher *et al.*, *Biol. Lett.* **1**, 151 (2005).
4. C. E. Franklin, W. Davison, F. Seebacher, *J. Exp. Biol.* **210**, 3068 (2007).
5. E. J. Glanville, F. Seebacher, *J. Exp. Biol.* **209**, 4869 (2006).

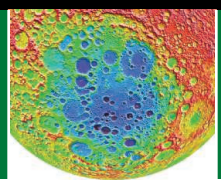
Response

WE THANK C. E. FRANKLIN AND F. SEEBACHER for elaborating on the importance of thermal acclimatization (thermally induced phenotypic plasticity) as a mitigation strategy for climate change. Our figure did in fact note that acclimatization would shift the thermal window along the temperature axis. However, acclimatization is limited to the thermal niche of a species, and these limits reflect thermal specialization. Such a niche becomes visible in analyses of temperature-dependent growth. The niche differs between species or even populations of the same species in various climates.

Although some Antarctic fishes have conserved limited warm acclimatization capabilities beyond present habitat temperatures (*1*–*3*) and thus live on the cold side of their thermal niche, it is clear that not all species do (*4*, *5*). Acclimatization capacity is minor among Antarctic marine invertebrates. It is likely also minimal among Antarctic icefishes

Published by AAAS

Earth and Moon under the microscope
882



Lunar asymmetry
885

that lost their hemoglobin when specializing on the high oxygen content of cold waters. Continued warming will thus cause species losses and, together with invasions by species presently at the doorstep of the marine Antarctic (*6*), will elicit progressive restructuring and functional shifts in marine Antarctic ecosystems.

The hypothesis of variable specialization according to climate variability requires further testing at the level of crucial life functions like growth, reproduction, and development in aquatic and terrestrial species and ecosystems (*7*).

HANS O. PÖRTNER,[1]* ANTHONY P. FARRELL,[2] RAINER KNUST,[1] GISELA LANNIG,[1] FELIX C. MARK,[1] DANIELA STORCH[1]

[1]Department of Integrative Ecophysiology, Alfred Wegener Institute, Bremerhaven, D-27570, Germany. [2]Department of Zoology, University of British Columbia, Vancouver, BC V6T 1Z4, Canada.

*To whom correspondence should be addressed. E-mail: hans.poertner@awi.de

References and Notes

1. G. Lannig, D. Storch, H.O. Pörtner, *Polar Biol*. **28**, 575 (2005).
2. E. Brodte, R. Knust, H.O. Pörtner, *Polar Biol.* **30**, 95 (2006).
3. H. A. Hudson, P. R. Brauer, M. A. Scofield, D. H. Petzel, *Polar Biol.* **31**, 991 (2008).
4. L. S. Peck, *Antarctic. Sci.* **17**, 497 (2005).
5. H. O. Pörtner, L. S. Peck, G. N. Somero, *Philos. Trans. R. Soc. London Ser. B* **362**, 2233 (2007).
6. S. Thatje *et al., Ecology* **86**, 619 (2005).
7. B. Huntley *et al., Ecol. Lett.* **7**, 426 (2004).
8. Supported by the MarCoPoli research program of the Alfred Wegener Institute and NSERC Canada.

LIFE IN SCIENCE

No Restroom for the Weary

OUR LAB IS IN A HISTORIC BUILDING AT THE University of Tokyo. Unfortunately, in this case, the word "historic" is synonymous with "very old" and "shabby." The poor condition of the electric power supply makes our electroencephalography (EEG) experiments a challenge—the electric signals we obtain are contaminated with every sort of noise.

We decided to send a postdoc, Yosuke Morishima, on a journey in search of a good experimental room. Yosuke carried an EEG amplifier and monitor and a colleague wore an EEG electrode cap. They visited every room in the seven-story building, including those that belonged to other labs, and checked the noise level of the EEG. Finally, they found the best room for the experiment: the men's restroom on the east wing of the building. We invited a subject to the new experimental room and started an EEG experiment. The recording was fantastic—we could see beautiful brain signals. However, after running several experimental sessions, we began to receive complaints from people who visited the room for the purpose for which it was originally designed. Thus, our search for the ideal laboratory continues.

KATSUYUKI SAKAI

Department of Cognitive Neuroscience, University of Tokyo 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. E-mail: ksakai@m.u-tokyo.ac.jp

EDITOR'S NOTE

This is an occasional feature highlighting some of the day-to-day humorous realities that face our readers. Can you top this? Submit your best stories at www.submit2science.org.

CREDIT: PETER HOEY

Social Science Evolves to Include Biology

THE SPECIAL SECTION ON GENETICS OF Behavior (7 November 2008) helps put flesh on the evolutionary skeleton that biologists, and some social scientists, have been developing for many years. In the Perspective "Biology, politics, and the emerging science of human nature" (p. 912), J. H. Fowler and D. Schreiber argue that "biologists and political scientists must work together to advance a new science of human nature." The study of politics is not the only social science that would profit by such collaboration. Indeed, efforts to integrate evolutionary biology have recently been under way in every facet of the social sciences, including anthropology, economics, history, law, linguistics, psychology, sociology, and even literary criticism and aesthetics, in addition to political science. Finally, belatedly, but with increasing empirical and theoretical validity, social scientists representing all disciplines are collaborating with biologists to advance a much-needed, new science of human nature.

DAVID P. BARASH

Department of Psychology, University of Washington, Seattle, WA 98195, USA. E-mail: dpbarash@u.washington.edu

TECHNICAL COMMENT ABSTRACTS

Comment on "Dynamic Shifts of Limited Working Memory Resources in Human Vision"

Nelson Cowan and Jeffrey N. Rouder

Bays and Husain (Reports, 8 August 2008, p. 851) reported that human working memory, the limited information currently in mind, reflects resources distributed across all items in an array. In an alternative interpretation, memory is limited to several well-represented items. We argue that this item-limit model fits the extant data better than the distributed-resources model and is more interpretable theoretically.

Full text at www.sciencemag.org/cgi/content/full/323/5916/877c

Response to Comment on "Dynamic Shifts of Limited Working Memory Resources in Human Vision"

Paul M. Bays and Masud Husain

Cowan and Rouder suggest that a modification to the four-slot model of visual working memory fits the available data better than our distributed-resource model. However, their comparisons of statistical fit are biased in favor of the slot model. Here, we compare the predictions of the two models and present further evidence against the division of visual memory into slots.

Full text at www.sciencemag.org/cgi/content/full/323/5916/877d

Int J Soc Robot (2009) 1: 267–282
DOI 10.1007/s12369-009-0019-1

ORIGINAL PAPER

Toward the Human–Robot Co-Existence Society: On Safety Intelligence for Next Generation Robots

Yueh-Hsuan Weng · Chien-Hsun Chen · Chuen-Tsai Sun

Accepted: 7 April 2009 / Published online: 25 April 2009

Abstract Technocrats from many developed countries, especially Japan and South Korea, are preparing for the *human–robot co-existence society* that they believe will emerge by 2030. Regulators are assuming that within the next two decades, robots will be capable of adapting to complex, unstructured environments and interacting with humans to assist with the performance of daily life tasks. Unlike heavily regulated industrial robots that toil in isolated settings, *Next Generation Robots* will have relative autonomy, which raises a number of safety issues that are the focus of this article. Our purpose is to describe a framework for a legal system focused on Next Generation Robots safety issues, including a *Safety Intelligence* concept that addresses robot *Open-Texture Risk*. We express doubt that a model based on Isaac Asimov's Three Laws of Robotics can ever be a suitable foundation for creating an artificial moral agency ensuring robot safety. Finally, we make predictions about the most significant Next Generation Robots safety issues that will arise as the human–robot co-existence society emerges.

Y.-H. Weng (✉)
Conscription Agency, Ministry of the Interior, Republic of China, Chengkungling, Taiwan
e-mail: yhweng.cs94g@nctu.edu.tw

C.-H. Chen
7531, Lucas Rd, Richmond, BC, V6Y 1G1, Canada
e-mail: lucemia.cis93g@nctu.edu.tw

C.-T. Sun
College of Computer Science, National Chiao Tung University, Hsinchu, Taiwan
e-mail: ctsun@cs.nctu.edu.tw

Keywords Safety intelligence · Robot legal studies · Roboethics · Robot policy · The three laws of robotics · Robot law · Social robotics

1 Introduction

The Japanese Robot Association[1] predicts that *Next Generation Robots* will generate up to 7.2 trillion yen (approximately 64.8 billion USD) of economic activity by 2025, with 4.8 trillion (43.2 billion USD) going to production and sales and 2.4 trillion (21.6 billion USD) to applications and support. According to the Japanese Ministry of Economy, Trade and Industry (METI), manufacturers will focus on specific markets (e.g., housework, nursing, security), while application and support firms provide maintenance, upgrading, and reselling services similar to today's information technology structure [1]. Also similar to the current IT industry, individual firms will specialize in such areas as education (public, safety, technical, etc.), selling insurance to cover special robot risks, and buying/selling used robots.

The *Fukuoka World Robot Declaration*, issued in February 2004, lists Japanese expectations for Next Generation Robots that co-exist with human beings, assist human beings both physically and psychologically, and contribute to the realization of a safe and peaceful society.[2] However, the declaration falls short in describing what Next Generation Robots should be. In a report predicting the near future (2020–2025) in robot development, the Japanese Robot Policy Committee (RPC, established by METI) created

[1] http://www.jara.jp/.

[2] International Robot Fair 2004 Organizing Office, World Robot Declaration (2004), http://www.prnewswire.co.uk/cgi/news/release?id=117957.

two Next Generation Robots categories: (a) next generation industrial robots capable of manufacturing a wide range of products in variable batch sizes, performing multiple tasks, and (unlike their general industrial predecessors) working with and/or near human employees; and (b) service robots capable of performing such tasks as house cleaning, security, nursing, life-support, and entertainment-all functions to be performed in co-existence with humans in businesses and homes. The report predicts that humans will gradually give Next Generation Robots a growing number of repetitive and dangerous service tasks, resulting in increased potential for unpredictable and dangerous actions [2]. METI describes the danger level in terms of contact degree: "low" for in-home communication or cleaning robots, "mid" for nursing robots, and high for universal humanoid robots capable of performing a wide range of tasks. How well such dangers can be anticipated is closely linked to the amount of autonomous behavior programmed into machines, in a relationship that remains to be determined.

Since 2000, Japanese [3] and South Korean [4] technocrats have been discussing and preparing for a *human–robot co-existence society* that they believe will emerge by 2030. Based on the content of policy papers and analyses published by both governments, researchers are currently studying potential *Robot Sociability Problems* that—unlike technical problems associated with design and manufacturing entail robot related impacts on human interactions in terms of regulations, ethics, and environments. Regulators are assuming that within the next two decades, robots will be capable of adapting to complex, unstructured environments and interacting with humans to assist with the performance of daily life tasks. Unlike heavily regulated industrial robots that toil in isolated settings, Next Generation Robots will have relative autonomy allowing for sophisticated interactions with humans. That autonomy raises a number of safety issues that are the focus of this article.

In addition to the semi-autonomous robots created by NASA scientists of the United States for exploration of the deep sea and the surface of Mars, in December 2008, the U.S. military reported its plan to devote approximately $4 billion USD within the following two years for the development of "ethical" robots-autonomous robot soldiers which will conform to the laws of warfare [5].

Artificial intelligence (AI) will be the main tool giving robots autonomy, expanded work ranges, and the ability to work in unstructured environments. Changes in human–robot relationships made possible by advancements in AI are likely to exert an impact on human society rivaling that of any other single technological innovation. We will predict the most significant issues of Next Generation Robots safety that will arise as the human–robot co-existence society emerges. The emerging co-existence society and issues in establishing robot law will be respectively discussed in Sects. 2 and 3. In Sects. 4 and 5 we will describe Human-Based Intelligence and our proposal for a *Safety Intelligence* concept to address these issues in light of Isaac Asimov's *Three Laws of Robotics*. In Sect. 6 we will describe the potential development of a *Legal Machine Language* to overcome the considerable shortcomings of Asimov's laws.

2 Human–Robot Co-Existence Society

The public often views robot development in terms of biomorphic machines. The actual situation involves input and innovation from multiple non-engineering fields that pave the way for harmonious interactions between humans and robots of all shapes, sizes, appearances, and capabilities. We will refer to interdisciplinary issues as *Robot Sociability Problems* and to engineering issues as *Robot Technical Problems*.

The Japanese are funding multiple efforts to address robot sociability problems and robot technical problems issues, including the establishment of research committees and physical environments for the testing of robot prototypes. In 1999, the Ministry of International Trade and Industry (MITI, which later became the above-mentioned METI) provided 450 million USD for a five-year research effort called "HRP: The Humanoid Robotics Project." The participants were Japan's major players in robotics: Hirochika Inoue and Susumu Tachi from the University of Tokyo, and representatives from Honda, Fujitsu, Panasonic (Matsushita), Kawasaki, Hitachi, and other corporations. The first (two-year) stage was dedicated to developing a "humanoid robot platform" (HRP-1) and the second to developing HRP-1 applications associated with human–robot co-existence [6].

A separate project was already underway at Waseda University in Tokyo's Shinjuku ward, the 1973 birthplace of WABOT-1, the world's first full-scale biped walking humanoid robot.[3] Named "Innovative Research on Symbiosis Technology for Humans and Robots in an Elderly-Dominated Society",[4] it was sponsored by the Ministry of Education, Culture, Sports, Science and Technology. Another important agreement was finalized in 2001, when Waseda University and Gifu prefecture established a "Wabot-House"[5] technical area in the city of Kakamigahara. The lab consists of three buildings, one for addressing ideas about ideal living spaces for people and various robot types; one focusing on social factors such as daily living needs, medical

[3] http://www.humanoid.waseda.ac.jp/history.html.

[4] http://www.waseda.jp/prj-rt/English/index/html.

[5] http://www.wabot-house.waseda.ac.jp/.

concerns, the natural environment, and other issues associated with human–robot co-existence; and one for determining how robot–human living spaces can be designed to support suitable levels of autonomous robot behavior [7]. Several other cities are now vying to attract robot researchers. Since late 2003, Fukuoka and Kitakyushu (both in Fukuoka prefecture) have shared the distinction of being the world's first Robot Development Empiricism Area (RDEA),[6] created according to national "Special Zone for Structural Reform" legislation [8]. The law addresses rules for road signs, traffic lanes, and traffic regulations, adding flexibility that allows for limited outdoor testing of robots (mostly on sidewalks) for the purpose of collecting empirical data.

Robot researchers in the two cities receive special tax dispensation and are exempt from Japanese radio law, meaning they do not have to apply for special certification for experiments using the 5 GHz wireless frequency range. A second goal of the Wabot-House group is to establish Gifu prefecture as one of several centers of the Japanese robot industry [9, 10]; other prefectures opening some of their city streets to robots in the interest of attracting manufacturers are Kanagawa (International Rescue Complex project) [11] and Osaka (RoboCity CoRE) [12]. A huge "robot city" is considered essential to planning for and creating a human–robot co-existence society, since the qualities of artificial environments that match the technical requirements of robot functions must be identified. In this regard, researchers are studying the potential for robots to serve as bridges between physical and virtual worlds. As Google's Vinton Cerf observes regarding the Internet:

> Virtual and real worlds will merge. Virtual interactions will have real world consequences. Control of the electrical grid and power generation systems could be made to appear to be part of a virtual environment in which actions in the virtual space affect actions in the real space. If your air conditioner is attached to the Internet, your utility might turn it off to prevent a brownout. Educational environments that mix real and virtual instruments and places will enrich the learning experience of school children [13].

If we add Next Generation Robots to Cerf's scenario, they will simultaneously act as virtual agents and physical actors, with overlapping boundaries that allow for the movement of many Next Generation Robots within a robot city. Before that day arrives, several important policy and regulatory issues must be settled to prevent a legal crisis. We will review some environmental issues first before looking at safety-related and other legal concerns.

An important concept in this area of robot research is *affordance*-the quality of an object or environment that allows an individual to perform an action. Psychologist James J. Gibson, who introduced the term [14], defined affordances as including all latent "action possibilities" in an environment, objectively measurable and independent of an individual's ability to recognize them, but always in relation to the actor and therefore dependent on the actor's capabilities. Since humanoid shapes are currently considered most suitable for a human–robot co-existence society, affordances for humanoid robots will be much more complex than those for industrial robots. Whereas industrial robots are limited to using relatively simple arm-like mechanisms to grab, move, and place objects in a restricted space, robots with legs may someday perform tasks using all four of its limbs—for instance, leaving one's physical home or business to run errands. For Next Generation Robots, affordance issues will involve service applications, their effects on industrial planning, functional imagination (if owners can do certain tasks, why not their robots?), and human–robot psychology [15]. Arguably, the central question is this:

- Should robots be designed to essentially "do anything" using all of their action possibilities?
- Should they be entrusted with providing nursing care in the absence of human caregivers?
- Should they be allowed to use their huge power requirements to feed grapes to their reclining owners?
- Should they be allowed to use their huge power requirements to feed grapes to their reclining owners?
- Should they be capable of sexual relations with humans [16]?

The range of possibilities raises the specter of a complex licensing system to control how Next Generation Robots are used.

Power consumption and generation is one item on a long list of environmental concerns that need to be addressed, preferably before the human–robot co-existence society becomes a reality. We are notoriously messy creatures, putting up hundreds of satellites into orbit above the earth and letting them stay up there long after their original purposes are exhausted. How can we prevent the abandonment of robots used to explore extreme environments that are too dangerous for humans? Furthermore, anyone who has tried to recycle a personal computer or peripheral knows that it is not as easy as placing them in a curbside recycling bin. Robot technology is a complex technical domain that will require a combination of ingenuity and strict enforcement in order to avoid disposal problems that are sure to arise when millions of robots and robot parts break down or wear out.

[6]The cities of Fukuoka and Kitakyushu were designated as Robot Development Empiricism Research Areas by the Japanese government in November, 2003. The first experiments in using robots in public spaces were conducted in February, 2004. In 2002, Fukuoka and Busan (South Korea) co-sponsored an international "Robocup" competition and conference. See http://www.robocup.or.jp/fukuoka/ and http://www.island-city.net/business/it/it.html.

Ambulatory robots will consume enormous amounts of energy. The International Energy Association (IEA) took that into consideration when making their predictions for future energy needs, and reported that if governments stick to their current policies, electric power consumption will double by 2030, and the percentage of electricity in total energy consumption will rise from 17 to 22 percent. Whereas western countries may find ways to generate "green power" for robots (e.g., fuel cells), developing countries will have little choice but to continue using the least expensive ways to generate electric power. In most cases that means burning coal, thereby increasing the quantity of greenhouse gasses released into the atmosphere [17]. Clearly, the emergence of a human–robot co-existence society makes our search for clean energy sources and ways of sharing them with developing countries-even more imperative.

3 Robot Law

Future robot-related issues will involve human values and social control, and addressing them will require input from legal scholars, social scientists, and public policy makers, using data from researchers familiar with robot legal studies. Levy [18] argues convincingly that a new legal branch of *Robot Law* is required to deal with a technology that by the end of this century will be found in the majority of the world's households. Here we will review the main issues expected to emerge in the fast-arriving era of human–robot co-existence in terms of four categories: robot ethics, rights, policy, and safety.

3.1 Robot Ethics

Determining how robotics will emerge and evolve requires agreement on ethical issues among multiple parties, in the same manner as nuclear physics, nanotechnology, and bioengineering. Creating consensus on these issues may require a model similar to that of the Human Genome Project for the study of Ethical, Legal and Social Issues (ELSI) sponsored by the US Department of Energy and National Institutes of Health.[7] Each agency has earmarked 3–5 percent of its financial support for genome research to ethical issues. ELSI's counterpart across the Atlantic is the European Robotics Research Network (EURON), a private organization devoted to creating resources for and exchanging knowledge about robotics research.[8] To create a systematic assessment procedure for ethical issues involving robotics research and development, a EURON committee has written and published *Roboethics Roadmap* [19], a collection of articles outlining potential research pathways, and speculating on how each one might develop. Due to the rate of rapid change occurring in the technology, EURON does not promote the collection as a guideline to state-of-the-art robotics or a declaration of principles such as those emerging from Japan and Korea. Instead, the *Roadmap* is billed as a review of topics and issues aimed at those individuals and regulatory bodies that will eventually determine robot policies—legislatures, academic institutions, public ethics committees, industry groups, and the like. It is important to note that the Roadmap focuses on human centered rather than robot or artificial intelligence centered ethics, perhaps due to its "near future urgency" perspective that addresses the next decade while contemplating foreseeable long term developments. For this reason, *Roboethics Roadmap* does not consider potential problems associated with robot consciousness, free will, and emotions.

According to the *Roadmap* authors, most members of the robotics community express one of three attitudes toward the issue of roboethics:

- *Not interested*: they regard robotics as a technical field and don't believe they have a social or moral responsibility to monitor their work.
- *Interested in short-term ethical questions*: they acknowledge the possibility of "good" or "bad" robotics and respect the thinking behind implementing laws and considering the needs of special populations such as the elderly.
- *Interested in long-term ethical concerns*: they express concern for such issues as "digital divides" between world regions or age groups. These individuals are aware of the gap between industrialized and poor countries and the utility of developing robots for both.

The authors of this paper are in the third category, believing that social and/or moral questions are bound to accompany the emergence of a human–robot co-existence society, and that such a society will emerge sooner than most people believe. Furthermore, we agree with the suggestions of several *Roboethics Roadmap* authors that resolving these ethical issues will require agreement in six areas:

1. Are Asimov's Three Laws of Robotics (discussed in Sect. 5) usable as guidelines for establishing a code of roboethics?
2. Should roboethics represent the ethics of robots or of robot scientists?
3. How far can we go in terms of embodying ethics in robots?
4. How contradictory are the goals of implementing roboethics and developing highly autonomous robots?
5. Should we allow robots to exhibit "personalities"?
6. Should we allow robots to express "emotions"?

[7] http://www.ornl.gov/sci/techresources/Human_Genome/research/elsi.shtml.

[8] http://www.euron.org/.

This list does not include the obvious issue of what kinds of ethics are correct for robots. Regarding "artificial" (i.e., programmable) ethics, some *Roadmap* authors briefly touch on needs and possibilities associated with robot moral values and decisions, but generally shy away from major ethical questions. We consider this unfortunate, since the connection between artificial and human-centered ethics is so close as to make them very difficult to separate. The ambiguity of the term *artificial ethics* as used in the EURON report ignores two major concerns:

- How to program robots to obey a set of legal and ethical norms while retaining a high degree of autonomy.
- How to control robot-generated value systems—or morality.

In this article we will respectively call these *Type 1* and *Type 2* artificial ethics. Since both will be created and installed by humans, the boundary between them will be exceptionally fluid.

Visually-impaired people depend on guide dogs to navigate their living environment. Given their serious responsibility, guide dogs must absolutely obey orders given by their owners. However, the dogs received instruction in "Intelligent Disobedience" which trains the dog to act against the orders of its master in emergency cases to ensure the person's safety. Initially, the dogs were trained to make these decisions according to human-centered value systems or what we have called Type 1 artificial ethics earlier. Nevertheless, through repeated training in disobedience through various kinds of situations, the decisions of the dogs show a blending of its own value system with the inculcated human-centered value system. As such, Intelligent Disobedience is what we call Type 2 artificial ethics, in which value is not absolutely human-centered.

Susan Leigh Anderson also holds the similar ideas [20], such as:

> It might be thought that adding an ethical dimension to a machine is ambiguous. It could mean either (a) in designing the machine, building in limitations to its behavior according to an ideal ethical principle or principles that are followed by the human designer, or (b) giving the machine ideal ethical principles, or some examples of ethical dilemmas together with correct answers and a learning procedure from which it can use the principle[s] in guiding its own actions.

The South Korean government is putting the finishing touches on a *Robot Ethics Charter*; when published,[9] it may stand as the world's first official set of ethical guidelines for robotics. According to that country's Ministry of Commerce, Industry and Energy (MC-IE), the Charter will present criteria for robot users and manufacturers, and guidelines for ethical standards to be programmed into robots. The standards are being established in response to a plan announced by the Ministry of Information and Communication to put a robot in every South Korean home by 2020. The Charter's main focus appears to be social problems—for example, human control over robots and the potential for human addiction to robot interaction. However, the document will also deal with a number of legal issues, including protections for data acquired by robots and machine identification for determining responsibility distribution.

In an April 2007 presentation at an international "Workshop on Roboethics" held in Rome, an MCIE representative gave three reasons explaining why his government felt a need to write a *Robot Ethics Charter* [21]: the country's status as a testing ground for robots (similar to its experience with IT electronics); a perceived need for preparation for a world marked by a strong partnership between humans and robots; and social demands tied to the country's aging population and low birth rate. The inclusion of guidelines for the robots themselves may be interpreted as tacit acknowledgement of Asimov's Three Laws of Robotics as well as concern over the implementation of Type 1 ethics in robot control systems.

3.2 Robot Rights

There are many barriers to overcome before we can produce human-based intelligence robots capable of making autonomous decisions and having limited "self-awareness". Still, futurists who believe that such a day will come in this century are contemplating issues that might emerge. In 2006, the Horizon Scanning Centre (part of the United Kingdom's Office of Science and Innovation) published a white paper with predictions for scientific, technological, and health trends for the middle of this century [22]. The authors of the section entitled "Utopian Dream, or Rise of the Machines" raise the possibility of robots evolving to the degree that they eventually ask for special "robo-rights" [23]. In this paper we will limit our discussion to how current human legal systems, in which rights are closely tied to responsibilities, will affect early generations of non-industrial robots.

Whenever an accident occurs involving humans, the person or organization that must pay for damages can range from individuals (responsible for reasons of user error) to product manufacturers (responsible for reasons of poor product design or quality). Rights and responsibilities will need to be spelled out for two types of Next Generation Robots. The system for the first type—Next Generation Robots lacking artificial intelligence-based "self-awareness"—will be straightforward: 100 percentage human-centered, in the

[9] http://www.reuters.com/article/lifestyleMolt/idUSSEO16657120070507.

same manner that dog owners must take responsibility for the actions of their pets. In other words, robots in this category will never be given human-like rights or rights as legal entities.

The second type consists of Next Generation Robots programmed with some degree of "self-awareness", and therefore capable of making autonomous decisions that can result in damage to persons or property. Nagenborg, Capurro, Weber and Pingel [24] argue that all robot responsibilities are actually human responsibilities, and that today's product developers and sellers must acknowledge that principle when designing first-generation robots for public consumption. They use two codes of ethics—one from the Institute of Electrical and Electronics Engineers and the other from the Association of Computing Machinery—to support their view that for complex machines such as robots, any attempt to remove product responsibility from developers, manufacturers, and users represents a serious break from the human legal system norm. We may see a day when certain classes of robots will be manufactured with built-in and retrievable "black boxes" to assist with the task of attributing fault when accidents occur, since in practice it will be difficult to attribute responsibility for damages caused by a robot, especially those resulting from owner misuse. For this reason, Nagenborg et al. have proposed the following meta-regulation:

> If anybody or anything should suffer from damage that is caused by a robot that is capable of learning, there must be a demand that the burden of adducing evidence must be with the robot's keeper, who must prove her or his innocence; for example, somebody may be considered innocent who acted according to the producer's operation instructions. In this case it is the producer who needs to be held responsible for the damage.

If responsibility for robot actions ever reaches the point of being denied by humans, a major issue for legal systems will be determining "punishment". Wondering if human punishment can ever be applied to robots, Peter Asaro observes that

> they do have bodies to kick, though it is not clear that kicking them would achieve the traditional goals of punishment. The various forms of corporal punishment presuppose additional desires and fears of being human that may not readily apply to robots pain, freedom of movement, morality, etc. Thus, torture, imprisonment and destruction are not likely to be effective in achieving justice, reform and deterrence in robots. There may be a policy to destroy any robots that do harm, but as is the case with animals that harm people, it would be preventative measure to avoid future harms rather than a true punishment . . . [American law] offers several ways of thinking about the distribution of responsibility in complex cases. Responsibility for a single event can be divided among several parties, with each party assigned a percentage of the total (p. 2) [25].

If we go this route, we may need to spell out robot rights and responsibilities in the same manner that we do for such non-human entities as corporations. Will we be able to apply human-centered values to robots as we do to other entities—a core value in human legal systems? To practice "robot justice," those systems will be required to have a separate set of laws reflecting dual human–robot-centered values. Roboresponsibilities would need to be clearly spelled out.

3.3 Robot Policy

A large number of robot-related policies must be debated and enacted before the mid-century "robot in every home" era begins: labor force displacement, physical safety, supervising research and development, and the shape of robot technology (RT) marketing, among many others. The breadth of these issues makes the appearance of a single, all-encompassing robot policy unlikely. However, it is likely that governments will follow their established top-down approach to giving direction to new technologies, and free-market advocates will resist such efforts.

The Japanese are currently addressing these concerns. In 2005 the METI created the above-mentioned Robot Policy Committee and invited robotics experts to serve on it. The committee's initial report emphasized the idea that Japanese government agencies and enterprises need to cooperatively address three areas of concern when establishing a Next Generation Robots industry [26]:

1. Develop a market environment: According to a survey conducted by the Japanese Robot Association, the Next Generation Robots market is expected to expand from 3 trillion yen in 2010 to 8 trillion yen in 2025 [27]. This enormous market will require support in many forms. Two examples are training in RT-related fields and the above-mention-ed need for local governments and robot enterprises to create areas dedicated to robot research and development.[10] Whereas technical research directions in the past were determined by university labs and research institutions, the committee suggested that market forces determine future directions.
2. Ensure safety: The clarification of legislative issues pertaining to Next Generation Robots safety requires analyses of human–robot interaction responsibilities before and after the manufacturing stage. Areas of concern for

[10] http://www.f-robot.com/english/index.html.

what we will call *pre-safety regulations* include standards for robot design and production. *Post-safety regulations* will address situations in which human injury is caused by robot actions, as well as systems for product liability protection and insurance compensation.

3. Develop a mission-oriented RT system: Japanese are accustomed to making products and manufacturing systems according to available technologies. A mission-oriented RT system will emphasize technology development by private firms based on demands and needs identified by government authorities [28].

Robot policy can be viewed as an intersection in which robot rights, robot ethics, and other subfields are integrated for the purpose of generating a direction for technology development. Since an equally important function of robot policy is to serve as a reference for creating new legislation, it must become a priority well before the expected emergence of a human–robot co-existence society. Even in draft form, robot policies can support international cooperation and information exchanges to assist legislators in setting legal guidelines.

The cultural difference between East and West may result in different responses to policy decisions. While most people agreed that "Humanoid" is a positive term for robotic research in Japan, the use of this type of research tend to be associated with "Frankenstein Complex"' in western societies. Kaplan [29] states that from the Japanese perspective, this difference seems to stem from the blurring between realizations of nature and the production of man. He described this Japanese point of view as "linking beings instead of distinguishing them", whereas in western discussion of robotics, the distinction between the natural and the artificial is very significant. Norwegian writer Jon Bing [30] has analyzed several representative western literature on artificial beings and found that the writings posited three types: "Machinelike man", "Manlike machine", and the new synthesis, "Cyborg". This distance between humans and robots is always stressed in western cultures.

In addition, Fumio Harashima from Tokyo Denki University has argued that the one main difference between Japanese and American robotics research is their source of funding—U.S. robitics research is supported by U.S. Military [31], thus the relevance of Robot Policy to military application or usage may take priority in the United States. For example, a group of USJFCOM-U.S. Joint Forces Command published a study titled "Unmanned Effects: Taking the Human out of the Loop" in August 2003, which suggest that by as early as 2025, widespread use of tactical, autonomous robots by U.S. military may become the norm on the battlefield [32].

3.4 Robot Safety

In 1981, a 37-year-old factory worker named Kenji Urada entered a restricted safety zone at a Kawasaki manufacturing plant to perform some maintenance on a robot. In his haste, he failed to completely turn it off. The robot's powerful hydraulic arm pushed the engineer into some adjacent machinery, thus making Urada the first recorded victim to die at the hands of a robot. A complete review of robot safety issues will be given in Sect. 5. Here we will simply emphasize safety as the most important topic requiring detailed consideration and negotiation prior to the coming human–robot co-existence society. Based on its large body of regulations and guidelines for industrial robots, Japan is considered a leader in this area. But as the METI Robot Policy Council notes, it has no safety evaluation methods or regulations currently in place for Next Generation Robots.

4 Human-Based Intelligence

In order to adapt to unstructured environments and work more closely with humans, Next Generation Robots must be designed to act as *biomorphic robots* with specialized capabilities. Lewis and Sim define biomorphic robots as imitations of biological systems capable of predicting the sensory consequences of movement, learning through the use of neural-type methods, and exploiting "natural system dynamics to simplify computation and robot control" [33]. Current examples of biomorphic robots are snakebots [34], insectbots,[11] and humanoid robots [35]. Researchers have built several variations of ant-like robots with mechanical limbs and cameras that are capable of "exploring" mazes and using cooperative strategies similar to those of ants to move about in different environments. AI researchers are finding ways to combine sampling from explicit external phenomena ("seeing") with implicit internal activities ("thinking") to create biomorphic robots capable of facial expressions that make them appear "sociable" to humans.[12]

Neurologists view the human brain as having three layers—primitive, paleopallium, and neopallium—that operate like "three interconnected biological computers, [each] with its own special intelligence, its own subjectivity, its own sense of time and space, and its own memory" [36]. From an AI viewpoint, the biomorphic equivalents of the three layers are action intelligence, autonomous intelligence, and Human-Based Intelligence (Fig. 1). Action intelligence functions are analogous to nervous system responses that coordinate sensory and behavioral information, thereby

[11] http://www.cis.plym.ac.uk/cis/InsectRobotics/Homepage.htm.

[12] http://www.ai.mit.edu/projects/humanoid-robotics-group/kismet/kismet.html.

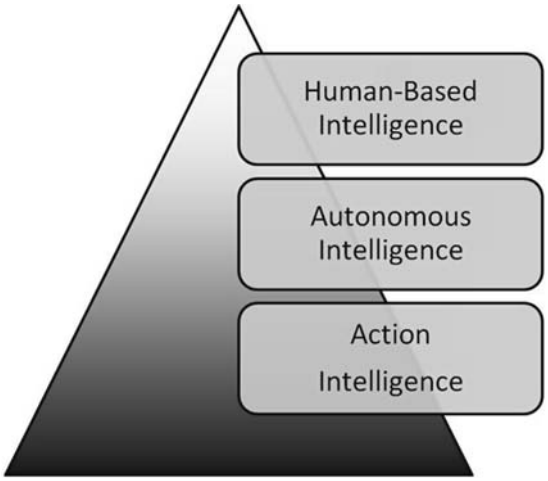


Fig. 1 Robot intelligence layers

giving a robot the ability to control head and eye movement [37], move spatially [38], operate machine arms to manipulate objects [39], and visually inspect its immediate environment [40]. Autonomous intelligence refers to capabilities for solving problems involving pattern recognition, automated scheduling, and planning based on prior experience [41]. Such behaviors are logical and programmable, but not conscious.

We are currently in a developmental period bridging action intelligence and autonomous intelligence, with robots such as AIBO,[13] QRIO [42], and Roomba[14] on the verge of being lab tested, manufactured, and sold. These simple and small robots are strong indicators of Next Generation Robots potential and the coming human–robot co-existence age. Even as "pioneer" robots, they have remarkable abilities to perform specific tasks according to their built-in autonomous intelligence—for instance, AIBO and QRIO robots have been programmed to serve as companions for the elderly, and Roomba robots as housecleaners. However, they cannot make decisions concerning self-beneficial actions or decide what is right or wrong based on a sense of their own value.

At the third level is *Human-Based Intelligence*—higher cognitive abilities that allow for new ways of looking at one's environment and for abstract thought, also referred to as "mind" and "real intelligence". Since a universally accepted definition of human intelligence has yet to emerge, there is little agreement on a definition for Human-Based Intelligence. Many suggestions and predictions appear to borrow liberally from science fiction, such as Human-Based Intelligence robots forming a new species with the long-term potential of gaining power over humans [43, 44]. In real-world contexts, researchers are experimenting with ways of combining action intelligence, autonomous intelligence, and human-based intelligence to act more human-like and to "comprehend complex ideas, learn quickly, and learn from experience" [45]. Human-Based Intelligence research started in the 1950s—roughly the same time as research on artificial intelligence, with which human-based intelligence is closely associated. One of the earliest and most famous efforts to examine human-based intelligence potential consisted of what is now known as the "Turing Test" [46]. Taking a behaviorist perspective, Alan Turing defined human intelligence as the ability "to respond like a human being," especially in terms of using natural language to communicate. There have been many efforts at creating programs that allow robots to respond like humans [47], but no AI program has ever passed the Turing test and been accepted as a true example of Human-Based Intelligence [48].

The legal and robot sociability problem issues that will arise over the next few decades are intricately linked with artificial intelligence, which was originally conceived as "the science and engineering of making intelligent machines, especially intelligent computer programs" [49]. Currently the two primary focuses of AI research are *conventional* (or symbolic) and *computational*; since intelligence still has such a broad definition, the two handles separate human-based intelligence parts. Conventional AI, which entails rational logical reasoning based on a system of symbols representing human knowledge in a declarative form [50], has been used for such applications as chess games (reasoning) [51], conversation programs (text mining),[15] and for organizing domain-specific knowledge (expert systems) [52]. While conventional AI is capable of limited reasoning, planning, and abstract thinking, researchers acknowledge that the use of symbols does not represent "mindful" comprehension, and is limited in terms of learning from experience [53].

Computational (non-symbol) AI [54] mimics natural (e.g., genetic [55] or neural [56]) learning methods, and allows for learning and adaptation based on environmental information in the absence of explicit rules—an important facility for living creatures. Computational AI has advantages in terms of overcoming noise problems, working with systems that are difficult to reduce to logical rules, and especially for performing such tasks as robot arm control, walking on non-smooth surfaces, and pattern recognition. However, as proven by chess programs, computational AI is significantly weaker than conventional AI in thinking abstractly and following rules. Among researchers in the fields of robotics and AI, the majority believes in the inevitability of human-based intelligence becoming a reality following breakthroughs in computational AI [57]. Others argue that

[13] http://support.sony-europe.com/aibo/.

[14] http://www.irobot.com/sp.cfm?pageid=122/.

[15] A.L.I.C.E. AI Foundation. Alicebot and AIML Documentation, http://www.alicebot.org/documentation/.

computational and conventional AI are both examples of behaviorism, and therefore will never capture the essence of human-based intelligence [58]. They claim that reaching that goal requires the development of an entirely new framework for understanding intelligence [59].

Optimistic or not, the belief that human-based intelligence robots will someday become a reality means that researchers must consider Human-Based Intelligence when predicting future robot safety and legal issues. They may conclude—as does Shigeo Hirose of the Tokyo Institute of Technology—that a prohibition on Human-Based Intelligence is necessary. Hirose is one of a growing number of researchers and robot designers resisting what is known as the "humanoid complex" trend [60], based on his adherence to the original goal of robotics: to invent useful tools for human use [61]. Alan Mackworth, past president of the Association for the Advancement of Artificial Intelligence [62], frames the robot Human-Based Intelligence issue as "should or shouldn't we" as oppose "can or can't we". Mackworth emphasizes the idea that goal-oriented robots do not require what humans refer to as "awareness", and therefore challenges the idea that we need to create human-based intelligence for machines.

In "ROBOT: Mere Machine to Transcendent Mind" [63], Carnegie Mellon Robotics Institute professor Hans Moravec predicts that robot intelligence will "evolve" from lizard-level in 2010 to mouse-level in 2020, to monkey level in 2030, and finally to human level in 2040—in other words, some robots will strongly resemble first-existence entities by mid-century. If true, future legislators interested in creating robot-related laws must face the difficult task of maintaining a balance between human and robots that will win broad acceptance from their constituents. Our motivation for this research is to give examples of robotics issues that future legislators and policy makers will have to address.

First, we will have to respond to standard societal suspicions about new technology, as exemplified by people's reactions to the Woosung Road in China, the first railway built in 1876 from Shanghai to Woosung. The speedy and powerful locomotive was seen as a monster by the Chinese that time, and finally, as a result of boycott, the Woosung Road was closed by the government and all the railway systems were transferred to Taiwan in 1877. This situation is very similar to *Uncanny Valley*, in which Masahiro Mori introduced the hypothesis in 1970 that human observers will respond with horror when faced with robots and other facsimiles of humans that look and act like actual humans. It now looks as though such suspicions and fears will be much less than what Mori predicted, but people may still express apprehension over blurred boundaries between humans and robots unless acceptable Robot Ethics guidelines are established.

In an earlier section we discussed the idea that robot responsibility should be regarded as human-owner responsibility. If we allow Human-Based Intelligence robo-ts to be manufactured and sold, the potential for any degree of robot self-awareness means dealing with issues such as punishment and a shift from human-centered to human–robot dual values. This is one of the most important reasons why we support a ban on installing Human-Based Intelligence software in robotsperhaps permanently, but certainly not until policy makers and robotists agree on these issues.

We also believe that creating Type 1 robots, in other words, "programming robots to obey a set of legal and ethical norms while retaining a high degree of autonomy requires agreement on human-centered ethics based on human values. The challenge is integrating human legal norms into robots so that they become central to robot behavior. The most worrisome issue is the potential capability of late-generation Human-Based Intelligence robots with significant amounts of self-awareness to generate their own values and ethics what we call Type 2 artificial ethics. Implementing Type 2 robot safety standards means addressing a long list of uncertainties for machines capable of acting outside of human norms. We are nowhere near discussing-let alone implementing policies for controlling Human-Based Intelligence robot behavior, since we are very far from having Human-Based Intelligence robots as part of our daily lives. However, if the AI/Human-Based Intelligence optimists are correct, the high risk of Human-Based Intelligence robots will necessitate very specific guidelines.

A guiding principle for those guidelines may be categorizing robots as *Third Existence* [64] entities, neither living/biological (first existence) nor non-living/non-biological (second existence). As described by Waseda University's Shuji Hashimoto, third existence machines will resemble living beings in appearance and behavior, but they will not be self-aware. We think this definition overlooks an important human–robot co-existence premise: most Next Generation Robots will be restricted to levels of autonomous intelligence that fall far short of Human-Based Intelligence, therefore their similarities with humans will be minor. As Cynthia Breazeal from the MIT Personal Robots Group observes:

> There's a "fuzzy boundary" that's very compelling for us, where we are willing to see robots as not human, but not exactly machine either [65].

According to the current legal system, robots are second-existence human property, a status that may be inadequate for the semi-autonomous Next Generation Robots that are about to enter people's homes and businesses especially in terms of responsibility distribution in the case of accidents. Asaro therefore proposes the creation of a new legal status for robots as "quasi-persons" or "corporations," while Nugenborg prefers emphasizing the point we made earlier about robot owners being responsible for their robots' actions in the same manner as pet owners. In Nugenborg's

view, robots should be given a legal status somewhere between personality and property.

Third existence status for robots may be an acceptable way of avoiding threats posed and impermanent caused by the society, law, and technology. But if Moravec's prediction comes true, the day will come when Human-Based Intelligence robots are a reality, and at that time we will be forced to decide between strictly following third existence guidelines or completely redefining the societal role and status of Human-Based Intelligence robots. If we choose the first response, then we must ban Human-Based Intelligence. However, legal scholars currently looking at an unknown future may yet find a way to make the second response work.

5 Safety Intelligence

In terms of safety standards, the primary difference in risk between industrial robots and autonomous Next Generation Robots is that the first involves machine standards and the second a mix of machine standards and open texture risk from unpredictable interactions in unstructured environments. *Open-Texture Risk* [66]—regarding language, any term in a natural language has a central (core) meaning, but the open texture character of language allows for interpretations that vary according to specified domains, points of view, time periods, etc. The open texture character of language produces uncertainty and vagueness in legal interpretations. Risk assessment associated with Next Generation Robots autonomous behavior faces a similar dilemma in that a core meaning exists, but the range of that core is difficult to clearly define, resulting in what we refer to as open texture risk. In a May 2006 paper on legislative issues pertaining to Next Generation Robots safety, Japanese METI committee members describe the difference in terms of *pre- and post-human–robot interaction responsibilities*. In the following discussion we will refer to them as *pre- and post-safety regulations*.

For industrial robots, safety and reliability engineering decisions are guided by a combination of pre-safety (with a heavy emphasis on risk assessment) and post-safety regulations (focused on responsibility distribution). Pre-safety rules include safeguards regarding the use and maintenance of robot systems from the design stage (e.g., hazard identification, risk assessment) to the training of robot controllers. One example of this is the United Kingdom Health and Safety Executive Office's 2000 publication of a set of industrial robot safety guidelines during installation, commissioning, testing, and programming.[16] Another example is International Standardization Organization (ISO)

	Safety Design by Risk Assessment	Safety Intelligence
Risk	Risk of Machine	Risk of Autonomous Behavior (The Open Texture Risk)
Limit	Machine's Standard	Robot's Intelligence Architecture
Effect	Decrease the Risk	Avoid Some Dangerous Behavior

Fig. 2 A comparison of safety regulation methods

rules-especially ISO 10218-1:2006, which covers safety-associated design, protective measures, and industrial robot applications. In addition to describing basic hazards associated with robots, ISO rules are aimed at eliminating or adequately reducing risks associated with identified hazards. ISO 10218-1:2006 spells out safety design guidelines (e.g., clearance requirements) that extend ISO rules covering general machine safety [67] to industrial robot environments. Those rules address safety-related parts of control systems and software design, but since the primary focus is on robot arms and manipulators [68], they have limited application to Next Generation Robots.

Designed and constructed according to very specific standards, industrial robots are limited to performing tasks that can be reduced to their corresponding mechanisms—in other words, they cannot alter their mechanisms to meet the needs of changing environments. Therefore, the primary purpose for performing industrial robot risk assessments is to design mechanisms that match pre-approved safety levels (Fig. 2). Complex Next Generation Robots motions, multi-object interactions, and responses to shifts in environments resulting from complex interactions with humans cannot be reduced to simple performance parameters. Next Generation Robots and future Human-Based Intelligence designers and manufacturers must instead deal with unpredictable hazards associated with the legal concepts of *core meaning* and *open texture risk*. Any term in a natural language has a core (central) meaning, but the open texture characteristic of human language [69] allows for interpretations that vary according to specified domains, points of view, time periods, and other factors, all of which can trigger uncertainty and vagueness in legal interpretations. Autonomous Next Generation Robots designers and programmers must therefore clearly define a core meaning plus an acceptable and useful range of that core.

The inherent unpredictability of unstructured environments makes it virtually impossible that we will ever see a fail-safe mechanism that allows autonomous robots to solve all open-texture problems. Consequently, Next Generation Robots safety regulations will require a mix of pre-safety

[16] http://products.ihs.com/Ohsis-SEO/113985.html.

and post-safety mechanisms, the first using a robot's AI reasoning content to eliminate most risk, and the second entailing a product liability system to deal with accidents that do occur. A clear security issue will be limiting the "self-control" of Next Generation Robots while still allowing them to perform designated tasks. As one *Roboethics Roadmap* author succinctly states, "operators should be able to limit robot autonomy when the correct robot behavior is not guaranteed." Giving operators this capability requires what we will call *Safety Intelligence*—that is, a system of artificial intelligence restrictions whose sole purpose is to provide safety parameters when semi-autonomous robots perform their tasks. Researchers have yet to agree on a foundation for a Safety Intelligence system, but the most frequently mentioned during the earliest stages of this discussion were the "Three Laws of Robotics" established by Isaac Asimov in his science fiction novel, I, Robot [70]:

1. First Law: A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. Second Law: A robot must obey orders given it by human beings, except when such orders conflict with the First Law.
3. Third Law: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

The first two laws represent a human-centered approach to Safety Intelligence that agrees with the current consensus of Next Generation Robots designers and producers. As robots gradually take on greater numbers of labor-intensive and repetitious jobs outside of factories and workplaces, it will become increasingly important for laws and regulations to support Safety Intelligence as a "mechanism of human superiority" [71]. The third law straddles the line between human- and machine-centered approaches. Since the purpose of robot functionality is to satisfy human needs, they must be designed and built in a manner so as to protect themselves as human property, in contrast to biological organisms that protect themselves for their own existence. As one magazine columnist has jokingly suggested, "A robot will guard its own existence … because a robot is bloody expensive" [72].

In his introduction to another work of fiction, *The Rest of the Robots*, Asimov wrote, "There was just enough ambiguity in the Three Laws to provide the conflicts and uncertainties required for new stories, and, to my great relief, it seemed always to be possible to think up a new angle out of the 61 words of the Three Laws" [73]. While those ambiguities may be wonderful for writing fiction, they stand as significant roadblocks to establishing workable safety standards for complex Next Generation Robots. In *Roboethics Roadmap*, some contributing authors note that the Three Laws raise many questions about Next Generation Robots programming:

- Which kinds of ethics are correct and who decides?
- Will roboethics really represent the characteristics of robots or the values of robot scientists?
- How far can and should we go when we program ethics into a robot?

Other robot researchers argue that Asimov's laws and the South Korean charter discussed in Sect. 3 still belong to the realm of science fiction because they are not yet applicable. Hiroshi Ishiguro of Osaka University, the co-creator of two female androids named Repliee Q1 and Repliee Q2 [74], believes it would be a mistake to accept Asimov's laws as the primary guiding principle for establishing robot ethics:

> If we have a more intelligent vehicle [e.g., automobile], who takes responsibility when it has an accident? We can ask the same question of a robot. Robots do not have human-level intelligence [75].

Mark Tilden, the designer of a toy-like robot named RoboSapien, says "the problem is that giving robots morals is like teaching an ant to yodel. We're not there yet, and as many of Asimov's stories show, the conundrums robots and humans would face would result in more tragedy than utility." Ian Kerr, law professor at the University of Ottawa, concurs that a code of ethics for robots is unnecessary:

> Leaving aside the thorny philosophical question of whether an AI could ever become a moral agent, it should be relatively obvious from their articulation that Asimov's laws are not ethical or legal guidelines for robots but rather about them. The laws are meant to constrain the people who build robots of exponentially increasing intelligence so that the machines remain destined to lives of friendly servitude. The pecking order is clear: robots serve people [76].

Currently, the two primary perspectives on the mix of AI and safety are either creating artificial agents with safety-oriented reasoning capabilities, or programming robots with as many rules as required for ensuring the highest level of safe behavior. Which perspective wins out will depend on how policy makers, designers, and manufacturers address as the *Three Questions For Three Laws of Robotics* we proposed following:

Question of Machine Meta-ethics: Susan Leigh Anderson [20] argued that Asimov's Three Laws of Robotics are an unsatisfactory basis for machine ethics, regardless of the status of the machine. She divided the robots into the robots with moral standings and without moral standings. First, she claimed that if the robots have moral standing then an ethical theory must take the being into account, then she introduced Warren's lists of six characteristics to define personhood (moral standing), such as Sentience, Emotionality, Reason, The capacity to communicate, Self-Awareness,

Moral Agency [77], by Warren's definition the robots with Human-Based Intelligence could be seem with moral standings, and it's immoral to forced robots with Human-Based Intelligence to obey the Asimov's Three Laws of Robotics to "serve people" such as Ian Kerr said earlier. As the robots without moral standings, Anderson introduced Immanuel Kant's consideration that "humans should not mistreat the entity in question, even though it lacked rights itself" [78], he argued that even though animals lack moral standing and can be used to serve the end of human beings, we should still not mistreat them because he said "he who is cruel to animals becomes hard also in his dealings with men". If the Third Existence robots are adopted to Asimov's three laws of robotics then it's allowed to let people bully or mistreat a robot. Therefore the three laws of robotics is inadequate whether robot with or without moral standing.

Question of Formality: The ability to "think abstractly" is uniquely human, and there is no way of being absolutely sure of how robots will interpret and react to the abstract meanings and vague terms used in human communication. For example, humans know how to distinguish between blood resulting from a surgical operation and blood resulting from acts of violence. Making such distinctions requires the ability to converse, to understand abstract expressions (especially metaphors), and to use domain knowledge to correctly interpret the meaning of a sentence. There are many examples that illustrate just how difficult this task is; one is Chomsky's famous sentence showing the inadequacy of logical grammar, "Colorless green ideas sleep furiously" [79], and another is Groucho Marx's line, "Time flies like an arrow, fruit flies like a banana".[17] Such examples may explain Asimov's description of robots as "logical but not reasonable" [80]. Therefore Asimov's Laws are facing a challenge for its "Formality" or "media to access the legal content", for people, we are used to let nature language be the media to access the content of law, however excepted for the Human-Based Intelligence robots the next generation robots are lacking the abstract ability to using nature language like human being in daily life.

Question of Regulation: Here the major issue is deciding whether or not Next Generation Robots need doctrinal reasoning powers. In other words how could we ensure that Next Generation Robots could enforce the three laws fully according such norms that human defined. In the early part we have already mentioned the artificial ethics Type 1 and Type 2. If we allow autonomous robots to define their own concepts of "safety", that means giving them the power to decide both when and how to react to stimuli. At some point those decisions will require artificial ethical and morality reasoning—the ability to distinguish between right and wrong. When considering "Morality Engineering", robotists such as Shigeo Hirose argue that in conflicts involving doctrinal reasoning and morality, the Three Laws may become contradictory or at risk of being set aside in favor of human requirements [81]. Using an extreme example, robots could be programmed to commit homicide under specific circumstances based on the wishes of a human majority. This example touches on two fears that many people have when they consider autonomous robots: they are troubled by the idea of letting robots obey rules that are impossible to express legislatively, and fearful of letting them defend laws established by imperfect humans. Human-Based Intelligence robots with Type 2 artificial ethics blends its own values while interpreting the three laws, thus causing ambiguity within itself: should it abide by human law or its own robot law?

	Machine Meta-ethics	Formality	Regulation
Human-Based Intelligence Robots	X	O	X
Third Existence Robots	X	X	O

Fig. 3 The Three Questions for Three Laws of Robotics

In an earlier section we concluded that Asimov's Three Laws of Robotics would be difficult to put into practice to achieve real-world Safety Intelligence. There are three reasons why those laws cannot be used with Human-Based Intelligence robots. First, machine meta-ethics are considered moral entities, and it would be immoral to force machines to obey the three laws in order to serve humans as Kant argued, people cannot mistreat beings that do not have moral standing. Second, in terms of regulation the Three Laws are unsuitable because Type 2 artificial ethics pose significant incentives for robots to be law-abiding. Third, regarding formality, the Three Laws cannot be applied to entities that lack the ability to think abstractly or to use human legal language competently.

Asimov's Three Laws of Robots have proven useful in giving robotists an early framework for discussing issues tied to robot behavior. However, they are ultimately unsatisfactory for any safety regulation model that emphasizes Safety Intelligence during design stages. Robotists must therefore search for a new approach to address these complex issues.

6 Legal Machine Language

The legal architecture of Human-Next Generation Robots interaction (including legal positions, usability, and content)

[17] http://www.quotationspage.com/quote/26.html.

is one of the most important issues for establishing the legal regulation of robots. It is hard to predict what the final solution will look like because of the number of open questions that remain. However, some existing concepts will affect the form of the legal architecture that eventually emerges. Here we will describe our proposal for an alternative *Legal Machine Language* based on two principles: *Code is Law* and *Embedded Ethics*.

Lawrence Lessig submitted "Code is Law" in 1998 [82], and he has noted that behavior is regulated by four kinds of constraints, such as Law, Social norms, the Market, and Nature—or what he called "Architecture". Since the Architecture of Cyberspace is absolutely built by code, the code could be an useful constraint to regulate user behavior or preserve crucial values such as Freedom or Privacy in Cyberspace. Therefore, using the code to regulate Cyberspace becomes another possible way. In the same reason the social control of Next Generation Robots should not be limited to "Dog Law" model, adopting the law described by human legal language in human society, is directly obeyed by human and indirectly obeyed by others such like robots. Next Generation Robots as "Virtual Agents into the Real World" is code-based artificial entities itself. The code or the architecture of Next Generation Robots could also be a regulator for its autonomy if we can define, formalize, and implement safety action without the need of moral reasoning.

Nature gives us many examples of animals interacting safely without complex moral judgments [83]. For example, flocking [84], as a common demonstration of emergent behavior for group of creatures such as birds or fishes could be seen as nature's traffic rules. When birds of the same species migrate, their shared genetic background allows them to fly in close or V-shaped formations without colliding—an important feature for shared and individual survival. Safe interaction requires adherence to a set of simple non-verbal safety rules shared by all members of the population: avoid crowding neighboring birds, fly along the same heading, or fly along the same average heading as neighboring birds. This zoological concept of flocking is considered useful for controlling unmanned aircraft [85] and other machines [86] including Next Generation Robots [87–90]. In an earlier section we discussed that a combination of action intelligence and autonomous intelligence is sufficient to recognize situations, avoid misunderstandings, and prevent accidents without processing ethics, performing morality reasoning functions, and making right/wrong decisions. Therefore robots only need to handle safety interaction and solely applied moral precept from human. In Ronald C. Arkin's words,

> We do not want the agent to be able to derive its own beliefs regarding the moral implications. . . , but rather to be able to apply those that have been previously derived by humanity [91].

Arkin has provided a possible framework embedded ethics into robots without Asimov's Three Laws of Robotics [91]. In his framework, there are (a) Ethical Behavior Control / Ethical Governor, (b) Human Robot Interface, and (c) Responsibility Advisor; the three components work cooperatively to form an ethical autonomous agent. First of all, Ethical Behavior Control / Ethical Governor Components, working as ex post facto suppression of unethical behavior, could be viewed as genetically built-in reflex system inside humans and animals. Ethical Behavior Control / Ethical Governor suppresses, restricts, or transforms unethical behavior and trigger protective behaviors by deliberative/reactive trigger protective behaviors. Therefore autonomous robot would have the ongoing ability to assess changing situations accurately and to correctly respond to complex real-world conditions. Second, Human Robot Interface, including body language, gesture [92], simple command, facial expression [93] and construct language like Loglan (identified as potentially suitable for human-computer communication due to its use of predicate logic, avoidance of syntactical ambiguity, and conciseness) gives both robot and human ability to be aware current situation. Human Robot Interface is used to prevent misunderstanding, predict influence, and consider possible corrective action [94]. Human Robot Interface design patterns should be defined as clear and explicit as possible, thereby Next Generation Robots could take immediate protective reactions in human-predictable ways as to mitigate risks tied to language-based misunderstandings or unstable autonomous behaviors. Third, Responsibility Advisor defined as "a mechanism in support of identifying and advising operators regarding the ultimate responsibility for the deployment of such a system" [91]. Responsibility Advisor, as a component of legal architecture notices each unethical behavior due to either human operator's override or autonomous robot's representational deficiency. Either by giving robot rights to refuse an unethical order or by limiting human to use robot ethically, we could define an explicit interaction rule set and a legal architecture that can be applied to all kinds of Next Generation Robots, one that accommodates the needs of a human–robot co-existence society in terms of simplicity and accountability.

In its current form, our proposal emphasizes three components of embedded ethics of Human–Next Generation Robots interaction: (a) the ongoing ability to assess changing situations accurately and to correctly respond to complex real-world conditions; (b) immediate protective reactions in human-predictable ways so as to mitigate risks tied to language-based misunderstandings or unstable autonomous behaviors; and (c) an explicit interaction rule set and a legal architecture that can be applied to all kinds of Next Generation Robots. Unlike the Three Law of Robotics, these three components, could be encoded by code and embedded di-

rectly inside autonomous intelligent of Next Generation Robots. As Lessig says "Architecture structures and constrains social and legal power, to the end of protecting fundamental values" [95].

Legal Machine Language could be a possible way for law regulation on Next Generation Robots' Open-Texture Risk, however in order to achieve the three criteria we mentioned earlier, the cross-fields conversation between Law and Robotics is necessary, at present the two characters of Legal Machine Language—"Code is Law" and "Embedded Ethics" provide a chance to review what's the adequate formality of law and how to implement the legal value literally obeyed by Next Generation Robots under the basis of Human–Robot Interaction in an environment that human and robots co-exist.

7 Conclusion

Emerging trends associated with Next Generation Rob-ots point to the day when robots will enter human society in large numbers, while engineers address all kinds of technical issues, a mix of engineers, social scientists, legal scholars, and policy makers will be making important decisions regarding robot sociability. In all cases, one of the priority concerns must be robot safety, since the emphasis for the future will be on human–robot *Co-Existence*.

In this paper we described a Safety Intelligence concept that can be separated into two dimensions. The first involves ethics—a special "Third Existence" status for robots and a recommended ban on equipping Next Generation Robots with Human-Based Intelligence. The second involves a mix of third existence designation and a Legal Machine Language designed to resolve issues associated with Open-Texture Risk. An important task for researchers is determining the structure and details of a Legal Machine Language part of an emerging field of legal research that we refer to as *Robot Legal Studies*.

References

1. Unsigned Editorial (2004) Toward 2025 Human-Robot Co-existence Society: The next generation robot vision seminar report. Japanese Ministry of Economy, Trade and Industry (METI), Tokyo
2. Ge SS (2007) Social robotics: The integration of advances in engineering and computer science. In: The 4th annual international conference organized by Electrical Engineering/Electronics, Computer, Telecommunication and Information Technology (ECTI) Association, Chiang Rai, Thailand, May 9–12, 2007. Keynote Speaker
3. Unsigned Editorial (2007) Long-term strategic guidelines "Innovation 25" (unofficial translation) at 60. Cabinet Office, Government of Japan. Available via www, http://www.cao.go.jp/innovation/index.html, Accessed Sep 25 2008
4. Lovgren S (2006) A robot in every home by 2020, South Korea Says. In: National Geographic News. Available via www, http://news.nationalgeographic.com/news/2006/09/060906-robots.html, Accessed Feb 12 2009
5. Etengoff A (2008) US military to develop 'ethical' robots. In: ITEXAMINER. Available via www, http://www.itexaminer.com/us-military-to-develop-ethical-robots.aspx, Accessed December 4 2008
6. Inoue H, Hirukawa H (2001) HRP: Humanoid robot project of MITI. J Robotics Soc Jpn 19(1):2–7 (in Japanese)
7. Wabot-House Lab (2004) Wabot-House Lab annual report for Heisei 15 Year 1. Waseda University, Tokyo (in Japanese)
8. Cabinet Secretariat, Office for the Promotion of Special Zones for Structure Reform (2003) Special zones for structure reform. Japan Government. Available via www, http://www.kantei.go.jp/foreign/policy/kouzou2/sanko/030326setumei_e.pdf, Accessed Sep 25 2008
9. Sugano S (2002) Wabot-House project. In: Advances in science, technology and environmentology (ASTE), Annual Report of RISE (Research Institute of Science and Technology), Waseda University, Tokyo at A10 (in Japanese)
10. Inaba A, Chihara K (2004) The robot project 21. J Robotics Soc Jpn 22(7):818–821 (in Japanese)
11. Miyamoto K (2004) Viewpoint on robot rescue development and support by Kanagawa Prefecture. J Robotics Soc Jpn 22(7):827–828 (in Japanese)
12. Asada M (2005) From synergistic intelligence to RoboCity CoRE. J Robotics Soc Jpn 23(8):942–945 (in Japanese)
13. Cerf VG (2007) The disruptive power of networks. In: Forbes Asia, May-7-2007 Issue, pp 76–77
14. Gibson JJ (1977) The theory of affordances. In: Shaw R, Bransford J (ed) Perceiving, acting, and knowing: Toward an ecological psychology. Hillsdale, NJ, pp 67–82
15. Sasaki M (2006) An introduction to the theory of affordances. J Robotics Soc Jpn 24(7):776–782 (in Japanese)
16. Levy D (2007) Love and sex with robots: The evolution of human–robot relationships. Harper Collins, New York
17. IPCC (2001) IPCC third assessment report: Climate change 2001. Available via www, http://www.ipcc.ch/ipccreports/assessments-reports.htm, Accessed at Feb 12 2009
18. Levy D (2006) Robots unlimited: Life in a virtual age. AK Peters (ed). Wellesley, MA
19. Unsigned Editorial (2006) Roboethics roadmap release 1.1. European robotics research network. Available via www, http://www.roboethics.org/atelier2006/docs/ROBOETHICS%20ROADMAP%20Rel2.1.1.pdf, Accessed at Feb 12 2009
20. Anderson SL (2008) Asimov's "three laws of robotics" and machine metaethics. AI Soc 22(4):477–493
21. Shim HB (2007) Establishing a Korean robot ethics charter. In: IEEE ICRA workshop on roboethics, April 14 2007, Rome, Italy. Available via www, http://www.roboethics.org/icra2007/contributions/slides/Shim_icra%2007_ppt.pdf, Accessed at Feb 12 2009
22. Unsigned Editorial (2006) Robots could demand legal rights. In: BBC News. Available via www, http://news.bbc.co.uk/2/hi/technology/6200005.stm, Accessed at Feb 12 2009
23. Unsigned Editorial (2006) Robot rights? It could happen, U.K. government Told. In: CBC news. Available via www, http://www.cbc.ca/technology/story/2006/12/21/tech-freedom.html, Accessed at Feb 12 2009
24. Nagenborg M, Capurro R, Weber J, Pingel C (2008) Ethical regulations on robotics in Europe. AI Soc 22(3):349–366
25. Asaro P (2007) Robots and responsibility from a legal perspective. In: IEEE ICRA'07 workshop on roboethics, April 14, 2007, Rome, Italy. Available via www, http://www.peterasaro.org/writing/ASARO%20Legal%20Perspective.pdf, Accessed at Feb 12 2009

26. Robot Policy Council (2005) Robot policy middle report – May 2005 version. Japan Ministry of Economy, Trade and Industry (in Japanese)
27. RIDC Council (2003) Founders' Statement. Robotics Industry Development Council. Available via www, http://www.f-robot.com/tokku/tokku.html, Accessed at Feb 12 2009
28. Robot Policy Council (2006) Robot Policy Council report – May 2006 version. Japan Ministry of Economy, Trade and Industry (in Japanese)
29. Kaplan F (2004) Who is afraid of the humanoid? Investigating cultural differences in the acceptation of robots. Int J Humanoid Robotics 1(3):465–480
30. Bing J (2008) The riddle of the robots. J Int Commer Law Technol 3(3):197–206
31. Harashima F (2007) Intelligent mechatronics systems go around with human. J Robotics Soc Jpn 25(1):20–24 (in Japanese)
32. U.S. Joint Forces Command (2003) Military robots of the future. In: About.com: US military. Available via www, http://usmilitary.about.com/cs/weapons/a/robots.htm, Accessed at Feb 12 2009
33. Lewis MA, Sim LS (2001) Certain principles of biomorphic robots. Auton Robots 11(3):221–226
34. Tanev I, Ray T, Buller A (2004) Evolution, robustness, and adaptation of sidewinding locomotion of simulated snake-like robot. Genet Evol Comput 3102:627–639
35. Faiola A (2005) Humanoids with attitude: Japan embraces new generation of robots. In: The Washington Post. Available via www, http://www.washingtonpost.com/wp-dyn/articles/A25394-2005Mar10.html, Accessed at Feb 12 2009
36. MacLean PD (1990) The triune brain in evolution: role in paleocerebral functions. Springer, London
37. Hager GD, Chang WC, Morse AS (1995) Robot hand-eye coordination based on stereo vision. IEEE Control Syst Mag 15(1):30–39
38. Lewis MA, Fagg AH, Solidum A (1992) Genetic programming approach to the construction of a neural network for control of a walking robot. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp 2618–2623
39. Chapin JK, Moxon KA, Markowitz RS, Nicolelis MAL (1999) Real-time control of a robot arm using simultaneously recorded neurons in the motor cortex. Nature Neurosci 2:664–670
40. Dickmanns ED (1988) Dynamic computer vision for mobile robot control. In: Proceedings of the 19th international symposium industrial robots, pp 314–327
41. Koditschek DE (1989) Robot planning and control via potential functions. Robotics Rev 1:349–367
42. Yeomans M (2005) Qrio dances Into spotlight at Carnegie Mellon Center. In: Pittsburgh Tribune-Review. Available via www, http://www.pittsburghlive.com/x/pittsburghtrib/s_298114.html, Accessed at Feb 12 2009
43. Neisser U, Boodoo G, Bouchard Jr TJ et al (1996) Intelligence: Knowns and unknowns. Am Psychol 51(2):77–101
44. Asimov I (1976) Bicentennial man. Ballantine, New York
45. Gottfredson LS (1997) Mainstream science on intelligence. Intelligence 24:13–23
46. Turing A (1950) Computing machinery and intelligence. Mind 1:433–460
47. Unsigned Editorial (2007) Emotion robots learn from people. In: BBC News. Available via www, http://news.bbc.co.uk/2/hi/technology/6389105.stm, Accessed at Feb 12 2009
48. Saygin AP, Cicekli I, Akman V (2000) Turing test: 50 years later. Minds Mach 10(4):463–518
49. McCarthy J (2007) What is artificial intelligence? In: John McCarthy's home page. Available via www, http://www-formal.stanford.edu/jmc/whatisai/whatisai.html, Accessed at Feb 12 2009
50. Newell A, Simon HA (1995) GPS: A program that simulates human thought. In: Computers and thought. MIT, Cambridge, pp 279–293
51. Hsu FH, Campbell M, Hoane Jr AJ (1995) Deep blue system overview. In: Proceedings of the 9th international conference on supercomputing, pp 240–244
52. Lederberg JD (1987) How dendral was conceived and born. In: Proceedings of the 1987 Association for Computing Machinery Conference on history of medical informatics, pp 5–19
53. Dreyfus HL, Dreyfus SE (1986) From Socrates to expert systems: The limits and dangers of calculative rationality. In: Mitcham C, Huning A (eds) Philosophy and technology II: Information technology and computers in theory and practice. Springer, New York, pp 111–130
54. Engelbrecht AP (2002) Computational intelligence: An introduction. Wiley, New York
55. Mitchell M (1996) An introduction to genetic algorithms. MIT, Cambridge
56. Abdi H (1994) A neural network primer. J Biol Syst 2:247–281
57. Warwick K (2004) March of the machines. University of Illinois Press, Chicago
58. Penrose R (1989) The emperor's new mind. Oxford University Press, New York
59. Hawkins J, Blakeslee S (2006) On intelligence. Yuan-Liou, Taipei (in Chinese)
60. Tajika N (2001) The future astro boy. ASCOM, Tokyo (in Japanese)
61. Kamoshita H (2005) The present and future of robots. X-media, Tokyo (in Japanese)
62. Unsigned Editorial (2007) What is a robot? In: CBC News. Available via www, http://www.cbc.ca/news/background/tech/robotics/definition.html, Accessed at Feb 12 2009
63. Moravec H (1999) ROBOT: Mere machine to transcendent mind. Oxford University Press, New York
64. Hashimoto S (2003) The Robot generating an environment and autonomous. In: Ojima T, Yabuno K (eds) The book of wabot 2. Chuko, Tokyo (in Japanese)
65. Lillington K (2008) So robots are social animals after all. In: Irish Times. Available via www, http://www.irishtimes.com/newspaper/finance/2008/1128/1227739081265.html, Visited Feb 12, 2009
66. Weng YH, Chen CH, Sun CT (2007) The legal crisis of next generation robots: On safety intelligence. In: Proceedings of the 11th international conference on artificial intelligence and law (ICAIL'07), Stanford, CA, USA. ACM, New York, pp 205–209
67. International Organization for Standardization (2006) ISO 13849-1:2006 Safety of machinery – Safety-related parts of control systems – Part 1: General principles for design. Available via www, http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=34931, Accessed at Feb 12 2009
68. International Organization for Standardization (1994) ISO 8373:1994 Manipulating industrial robots – Vocabulary. Available via www, http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=15532, Accessed at Feb 12 2009
69. Lyons D (1999) Open texture and the possibility of legal interpretation. Law Philos 18(3):297–309
70. Asimov I (1950) Robot. Gnome Press, New York
71. Fodor JA (1987) Modules, frames, fridgeons, sleeping dogs and the music of the spheres. In: Pylyshyn Z (ed) The robot's dilemma: The frame problem in artificial intelligence. Kluwer, Hingham
72. Langford D (2006) It's the law. In: SFX 146. Available via www, http://www.ansible.co.uk/sfx/sfx146.html, Accessed at Feb 12 2009
73. Asimov I (1964) The rest of the robots. Collins, New York
74. Whitehouse D (2005) Japanese develop "female" android. In: BBC News. Available via www, http://news.bbc.co.uk/1/hi/sci/tech/4714135.stm, Accessed at Feb 12 2009
75. Lovgren S (2007) Robot codes of ethics to prevent android abuse, protect humans. In: National geographic news. Available via www, http://news.nationalgeographic.com/news/2007/03/070316-robot-ethics.html, Accessed at Feb 12 2009

76. Kerr I (2007) Minding for machines. In: Ottawa citizens news. Available via www, http://www.canada.com/ottawacitizen/news/opinion/story.html?id=e58202bb-f737-4ba7-a0ad-79e8071a1534, Accessed at Feb 12 2009
77. Warren MA (1997) On the moral and legal status of abortion. In: LaFollette H (ed) Ethics in practice. Blackwell, Oxford
78. Kant I (1780) Our duties to animals. In: Infield L (trans.). Lectures on ethics. Harper and Row, New York, pp 239–241
79. Chomsky N (1957) Syntactic structures. Walter de Gruyter, New York
80. Asimov I (1957) The naked Sun. Doubleday, New York
81. Hirose S (1989) A robot dialog. J Robotics Soc Jpn 7(4):121–126 (in Japanese)
82. Lessig L (1998) The laws of cyberspace - draft 3. In: Proceedings of the Taiwan Net'98 conference, Taipei, Taiwan
83. Arkin RC (1998) Behavior-based robotics. MIT, Cambridge, pp 31–62
84. Spector L, Klein J, Perry C, Feinstein M (2003) Emergence of collective behavior in evolving populations of flying agents. Proceedings of the genetic and evolutionary computation
85. Gabbai JME (2005) Complexity and the aerospace industry: Understanding emergence by relating structure to performance using multi-agent systems. Doctoral thesis, University of Manchester
86. Reynolds CW (1987) Herds and schools: A distributed behavioral model. Comput Graph 21(4):25–34
87. Bekey GA, Tomovic R (1986) Robot control by reflex actions. In: Proceedings of the 1986 IEEE international conference on robotics and automation, pp 240–247
88. Zhang X, Zheng H, Duan G, Zhao L (2002) Bio-reflex-based robot adaptive motion controlling theory. In: Proceedings of the 4th world congress on intelligent control and automation, pp 2496–2499
89. Newman WS (1989) Automatic obstacle avoidance at high speeds via reflex control. In: Proceedings of the 1989 IEEE international conference on robotics and automation, pp 1104–1109
90. Wikman TS, Branicky MS, Newman WS (1993) Reflexive collision avoidance: A generalized approach. In: Proceedings of the 1993 IEEE international conference on robotics and automation, pp 31–36
91. Arkin RC (2008) Governing lethal behavior: embedding ethics in a hybrid deliberative/reactive robot architecture. In: Proceedings of the 3rd ACM/IEEE international conference on human robot interaction, HRI '08
92. Ge SS, Yang Y, Lee TH (2008) Hand gesture recognition and tracking based on distributed locally linear embedding. Image Vis Comput 26:1607–1620
93. Yang Y, Ge SS, Lee TH, Wang C (2008) Facial expression recognition and tracking for intelligent human-robot interaction. J Intel Serv Robotics 1(2):143–157
94. Brooks AG, Arkin RC (2007) Behavioral overlays for non-verbal communication expression on a humanoid robot. Auton Robots 22(1):55–74
95. Lessig L (1999) Code and other laws of cyberspace. Basic Books, New York

Advanced Robotics

ISSN: 0169-1864 (Print) 1568-5535 (Online) Journal homepage: https://www.tandfonline.com/loi/tadr20

Beyond Robot Ethics: On a Legislative Consortium for Social Robotics

Yueh-Hsuan Weng

To cite this article: Yueh-Hsuan Weng (2010) Beyond Robot Ethics: On a Legislative Consortium for Social Robotics, Advanced Robotics, 24:13, 1919-1926, DOI: 10.1163/016918610X527220

To link to this article: https://doi.org/10.1163/016918610X527220

Published online: 02 Apr 2012.

VSP

Advanced Robotics 24 (2010) 1919–1926

ADVANCED ROBOTICS
brill.nl/ar

Short paper

Beyond Robot Ethics: On a Legislative Consortium for Social Robotics

Yueh-Hsuan Weng [a,b,*]

[a] Peking University Law School, Beijing, P. R. China

[b] Yushan Studio for Artificial Intelligence and Law, No. 5 Yiheyuan Road, Haidian District, Beijing, 100871, P. R. China

Received 5 October 2009; revised 4 June 2010; accepted 10 June 2010

Abstract

As robots are increasingly integrated into human society, associated problems will resemble or merge with those in other fields — we can refer to this phenomenon as the 'robot sociability problem'. In this paper, the author first analyzes the dynamic relationship between robot ethics, robotics and robot law, and then proposes a 'practical robots' approach for solving the robot sociability problem. As this approach is based on legal regulations, the author posits that a functional platform such as a 'legislative consortium for social robotics' is crucial at the initial stage for social robotics development. In conclusion, the author discusses how a legislative consortium for social robotics will be a useful approach for solving the robot sociability problem, especially emerging structural legislative problems that are related to autonomous robots.

Keywords

Robot ethics, robot policy, robot law, social system design, social robotics

1. Introduction

The term 'roboethics' (robot ethics) was first officially mentioned in a symposium that was organized by several European robotics institutes in 2004 [1]. Following this, the European Robotics Research Network (EURON) published the 'Roboethics Roadmap' [2] and the South Korean government has prepared a draft of a 'Robot Ethical Charter' [3] as a guideline for building a human–robot co-existing society. In addition, the Japanese Ministry of Economy, Trade and Industry (METI) has issued a series of 'Robot Policies' that address business applications [4], safety regulation proposals [5] and the creation of a sound service robot market for the next two decades.

* E-mail: weng.yuehhsuan@gmail.com; ysail@pku.edu.cn

 DOI:10.1163/016918610X527220

Robots with the capacity to perform autonomous behaviors can adapt to complex environments and interact with humans. As robots are increasingly integrated into human society, associated problems will resemble or merge with those in other fields — we can refer to this issue as the 'robot sociability problem' [6]. Sociability is a skill, tendency or property of being sociable or social, of interacting well with others [7]. This ability is important to human beings, but because robot sociability is artificial, the relationship between humans and robots can be controlled by humans. The author proposes that a design for robot sociability be divided into two aspects: one is technically centered on human–robot interaction and the other is legally oriented, determining the ethics, policy and law to be applied to independent robots, hereafter referred to as the 'social system design'.

In this paper, the author focuses on the robot sociability problem from a legal perspective, especially the robot legislative issue; note that the discussion is limited to structural problems for developing robot laws and what kind of measures might reduce the risks that arise from the mentioned structural legislative problems for social robots.

2. Structural Problems for Robot Law

2.1. Beyond Robot Ethics — From 'Ideal Robots' to 'Practical Robots'

The possibilities of social robots can be divided by three metaphors corresponding to robotics, robot ethics and robot law: possible robots, ideal robots and practical robots (Fig. 1). Robotics tests the limits of innovation and creativity, so it is possible that all manner of artificial beings could be created, including those that could be harmful to human society. On the other hand, robot ethics holds a moral philosophical approach to examine the existence of these artificial beings; however, sometimes this may be thought of as wishful thinking as there is always a gap between theory and application in the real world.

Finally, robot law, which is not merely concerned with 'What a social robot can be' or 'What a social robot should be', calls attention to the intersection between

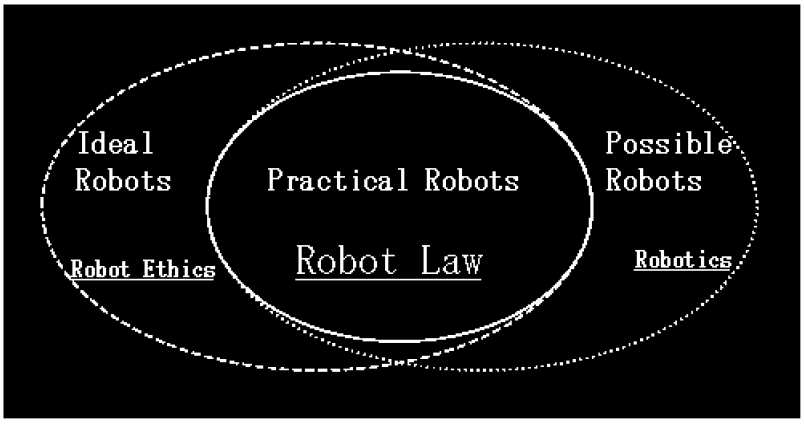


Figure 1. Three metaphors for robotics, robot ethics and robot law.

the two issues. It represents a practical perspective for social robots, and this attitude is useful and necessary to maintain human–robot co-existence in the long run. The author believes that by taking a 'practical robot' attitude, 'robot ethics' can be developed practically; furthermore, this attitude helps law makers move beyond a 'pure' robot ethics or moral philosophical knowledge domain and take a realistic viewpoint when performing their legislative work on regulating social robots.

2.2. Robot Legislative Policy Is as Important as Robot Safety and Industrial Policy

A large number of robot-related policies must be debated and enacted before a foreseeable mid-century 'robot-in-every-home' era begins: labor force displacement, physical safety, supervising research and development, and the shape of robot technology marketing, among many others. The breadth of these issues makes the appearance of a single, all-encompassing robot policy unlikely. However, it is likely that governments will follow their established top-down approach in giving direction to new technologies, e.g., in 2005 METI created the above-mentioned Robot Policy Committee and invited robotics experts to serve on it. The committee's initial report emphasized the idea that Japanese governmental agencies and enterprises need to cooperatively address issues relating to business, safety and innovation [4]. Currently, robot policy research has already covered many crucial topics, such as 'How to address a sound business policy to establish the robot technology industry' [8] and 'How to plan a safety policy as to build a safety standard for next generation robots' [5]. The author predicts there is a strong demand for another 'legislative policy' within the next stage of robot policy development.

Comparing industrial robots and social robots, the major difference is based on the 'contact level' with society. Industrial robots perform their effective working ability only in structured environments (i.e., a factory assembly line); in other words, their contact level with humans and society as a whole is very low. Therefore, the social regulation of industrial robots is almost addressed in its machine standards under safety and business considerations, but rarely touched on the part of human laws. As for social robots, due to them having closer contact with society, when they are deployed into unstructured environments to perform their duties with humans, they may elicit changes in many current relationships with humans concerning right and responsibility in daily life. The author predicts that there will be a strong demand for a group of 'robot legislative policies' as guidelines to adjust many current existing human laws in preparation for a human–robot co-existence society. However, the current difficulty for developing a sound robot legislative policy is based on the 'complexity' of robots, e.g., the word 'robot' might refer to many kinds of different things. In addition, the technological domains for social robots may include mechanical engineering, electrical engineering, computer science, cognitive science, biological engineering and chemical engineering. Furthermore, the complexity of responsibility distribution of social robots is based on its Third Existence character — neither living/biological (first existence) nor non-living/non-

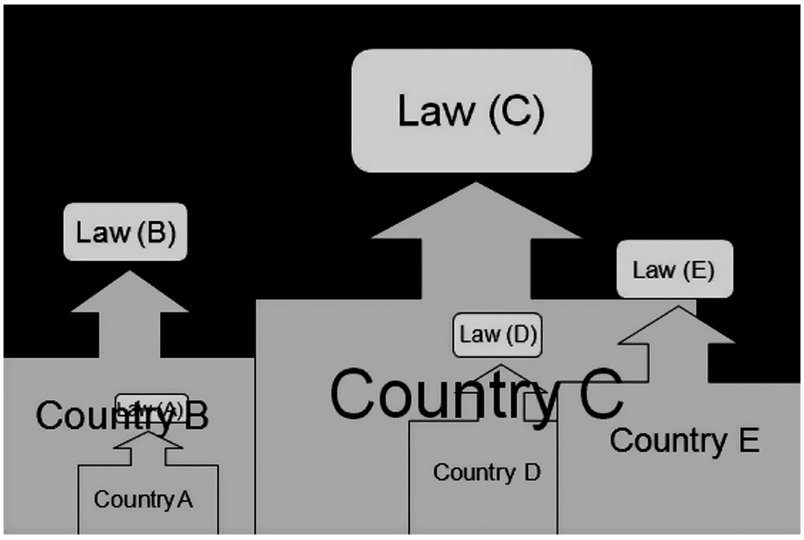


Figure 2. Colonization of robot legislative affairs.

biological (second existence) [9]. These factors might cause a serious situation for law makers to define their robot legislative policy well.

2.3. Avoiding the Colonization of Robot Legislative Affairs

Countries have different preferences for the usage and application of robotics, and these attitudes will be reflected by their domestic legislative policies as well as legal norms. However, what is worrisome is that if a conflict were to occur, an advanced robotics country's robot law may hold a relatively advantageous position. Less advanced countries may then be forced to modify their own domestic robot laws, and accept the belief and interests of advanced robotics countries, thus giving rise to a 'colonization of robot legislative affairs' (Fig. 2). It might form a global crisis when the legislative colonization touches on some crucial issues in robot sociability, such as the legitimacy of autonomous military robots access to human living spaces or to allow unethical applications that might harm world peace.

The author suggests that building a global consensus between countries will be a solution for avoiding the colonization of robot legislative affairs. Take an example from the global nuclear regulation organization — the International Atomic Energy Agency (IAEA). Established after World War II, the IAEA initially comes from US President Eisenhower's 'Atoms for Peace' address to the General Assembly of the United Nations on 8 December 1953. These ideas helped to shape the IAEA Statute, which 81 nations unanimously approved in October 1956. The Statute outlines the three pillars of the Agency's work — nuclear verification and security, safety, and technology transfer [10].

There are three reasons to support the IAEA as an effective solution to ensure the consensus of its three pillars between countries: (i) it is an inter-governmental organization, (ii) its relationship with the United Nations is regulated by special

agreement and (iii) it is an institution for regulating nuclear applications by international laws. Robot technology is similar to nuclear technology in that the technology itself is powerful, yet neutral; whether it is harnessed for 'good' or 'evil' outcomes is dependent on us as humans. Therefore, the author suggests that an internationally approved institution can help build consensus between countries so as to avoid colonization of robot legislative affairs.

3. Legislative Consortium for Social Robotics

3.1. Why Do We Need a Legislative Consortium for Social Robotics?

The author believes in an internationally approved legal institution such as a legislative consortium for social robotics that would serve not only to supervise the inappropriate or unethical application of robotics from its member countries, but also help its member countries to develop their domestic robot legislative policies by issuing guidelines. Note that due to the 'complexity' and 'contact level with the society' issues some small or less-advanced robotics countries might be unable to define their own robot laws because the scope of the domain issues are too complex and wide; a legislative consortium for social robotics as a third party institution itself might also prevent many countries with less-advanced robotics following a few advanced robotics countries' robot legislative policies (Fig. 3).

Other functions of the legislative consortium for social robotics will be to help the social robotics industry develop supervisory guidelines for the real-world use of artificial intelligence — programmed robots. In other words it can deal with very technically based issues, such as developing robot legal machine language [9] — to build robots embedded with legal guidelines for the robots to behave legally in a human living environment. From this viewpoint, the function of the legislative consortium for social robotics on advanced robotics is similar to the World Wide

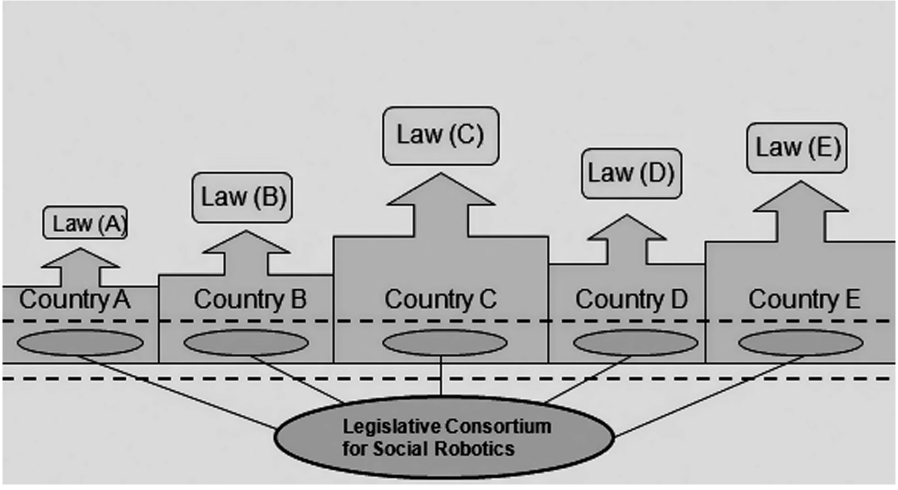


Figure 3. Legislative consortium for social robotics.

Web Consortium on the World Wide Web: its mission is 'to lead the World Wide Web to its full potential by developing protocols and guidelines that ensure the long-term growth of the Web' [11].

Programming social values and norms into robots designed to interact with humans requires input from legal and legislative experts regarding such topics as robot ethics and safety. However, there is currently a significant shortage of law scholars familiar with these issues, therefore my proposal calls for a platform for building and supporting expertise in the legal aspects of robot sociability in preparation for a human–robot co-existence society. It will be a gradual process over the next one or two decades; therefore, these legal policies need to be in place and will need to evolve as well [12].

3.2. Contemporary Tasks of the Legislative Consortium for Social Robotics

The legislative consortium for social robotics focuses on making robot legislative policies as well as other guidelines for social robots development. There are several crucial legal issues related to social robotics, including the following.

3.2.1. Preventive Arms Control Issue

The US Army plans to replace one-third of its armored vehicles and weapons with robots by 2015 [13]. These military robots are more effective and powerful than human soldiers in the battlefield. In other words, it might cause an asymmetric situation between countries that have a 'robot army' and countries that do not. Furthermore, it might also threaten the current international security system. The Treaty on Conventional Armed Forces in Europe sets 'troop ceilings' on military equipment, such as tanks, artillery pieces, armored combat vehicles, combat aircraft and attack helicopters. However, when military equipment is 'robotized', how to make a proper 'exchange rate' between conventional and robotized troop ceilings on military equipment will be an emerging challenge for international security.

3.2.2. Safety Issue

Currently, there are no safety standards for service robots in the world [9] and the International Organization for Standardization plans to create an international standard for service automations by 2011 [14]. However, in addition to an international robot safety standard, it is necessary to consider a set of domestic robot safety regulations as well. The author suggests that it is important to borrow the experience from 'automobile law' (US Intermodal Surface Transportation and Efficiency Act, European Type Approval, etc.) in order to save time and resources. As a senior official the Japanese Robot Business Promotion Council remarked, 'As with automobiles, there needs to be a set of safety rules that are recognized by the public in order for service robots to become widely accepted' [14].

3.2.3. Privacy Issue

In order to ensure that social robots can safely interact with human beings while providing high-quality, personal fitness service, it is necessary to equip robots with

a considerable amount of sensors. The sensor itself may need capabilities for multifunction and advanced perception in order to support the robot to enforce its task through unstructured environments. An example of one kind of sensing mechanisms may be called 'moving object sensor technology': the whole process of data collecting and reuse from ubiquitous sensor networks and corresponding middleware is 'intrusive'. Consequently, this technology will result in high risks for personal data disclosure or illegal use of personal data [15]. The property of sensor data is very different from personal data received from current information technology and network technology. Although some raw data look 'pure' and do not reveal any personal information at first, if combined with middleware or data mining it is possible to disclose much personal information. However, current privacy protection falls short on coping with this issue and it is necessary for comparative legal research to address how to cover this legal gap.

3.3. Who Will Benefit the Most from a Legislative Consortium for Social Robotics?

First, a legislative consortium for social robotics will provide guidance for the small number of scholars currently working on robotics issues, and future scholars and experts who are expected to emerge as the robotics industry expands.

Second, the robotics industry will generate billions of dollars of economic activity while creating many unforeseeable legal, safety and insurance issues. A legislative consortium for social robotics has the potential to benefit the growing number of corporations entering the field of robotics research and manufacturing.

Third, as the industry develops, robots will enter the homes of millions of families who will enjoy their assistance for tasks ranging from simple housekeeping chores to providing security services to assisting medical homecare professionals to performing rescue operations. As these human–robot interactive duties grow in complexity, the need for a safety certification system will also grow. Thus, a legislative consortium for social robotics can be said to potentially benefit multiple levels of society.

4. Conclusions

The author believes that a 'legislative consortium for social robotics' will be a useful approach for solving robot sociability problems, especially those emerging global legal issues related to autonomous robots. As robots become more integrated into human society, the importance of a legal framework for social robotics will become more obvious. Determining how to maintain a balance between human–robot interaction (robot technology development) and social system design (a legal regulation framework), the author predicts, will be the biggest challenge — especially on safety and legal issues — when a human–robot co-existence society emerges [16].

Acknowledgements

The author wishes to thank Mr Andrew Eng, PhD student, Northwestern University, Evanston, IL, USA, for his assistance in preparing this manuscript.

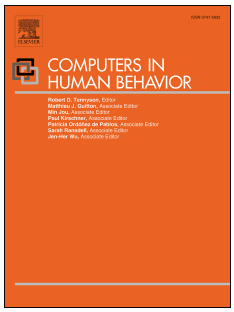
References

1. G. Veruggio, The birth of roboethics, presented at: *IEEE Int. Conf. on Robotics and Automation Workshop on Roboethics*, Barcelona, Invited Talk (2005).
2. Unsigned Editorial, *Roboethics Roadmap Release 1.1*, European Robotics Research Network, Haverlee (2006).
3. H. B. Sim, Establishing a Korean robot ethics charter, presented at: *IEEE Int. Conf. on Robotics and Automation Workshop on Roboethics*, Rome, Invited Talk (2007).
4. Robot Policy Council, *Robot Policy Middle Report — May 2005 Version*, Japan Ministry of Economy, Trade and Industry, Tokyo (2005) (in Japanese).
5. Unsigned Editorial, *Safety Guidelines for Next-Generation Robots*, Japan Ministry of Economy, Trade and Industry, Tokyo (2007) (in Japanese).
6. Y. H. Weng, C. H. Chen and C. T. Sun, The legal crisis of next generation robots: on safety intelligence, in: *Proc. 11th Int. Conf. on Artificial Intelligence and Law*, Palo Alto, CA, pp. 205–209 (2007).
7. Wiktionary, Sociability, http://en.wiktionary.org/wiki/sociability
8. Unsigned Editorial, *A Roadmap for US Robotics — From Internet to Robotics*. Computing Community Consortium, Washington, DC (2009).
9. Y. H. Weng, C. H. Chen and C. T. Sun, Toward the human–robot co-existence society: on safety intelligence for next generation robots, *Int. J. Social Robotics* **1**, 267–282 (2009).
10. International Atomic Energy Agency, http://www.iaea.org/About/index.html
11. World Wide Web Consortium, http://www.w3.org/Consortium/mission.html
12. Y. H. Weng, Toward the human–robot co-existence society: on legislative consortium for social robotics, presented at: *IEEE Int. Conf. on Robotics and Automation Workshop on Service Robots in Urban Environments: Legal and Safety Issues*, Kobe, Invited Talk (2009).
13. J. Blech, The future of war: attack of the killer robots, *Spiegel Online Int.*, http://www.spiegel.de/international/world/0,1518,500140,00.html (2007).
14. Unsigned Editorial, *Robot safety standards planned*, The Asahi Shimbun, http://www.asahi.com/english/TKY201001260395.html (2010).
15. T. Sato, Moving object sensor technology for security and safety, *CREST Annual Research Report*, Japan Science and Technology Agency, Tokyo (2007) (in Japanese).
16. L. Zyga, Living safely with the robots, beyond Asimov's laws, PhysOrg.com, http://www.physorg.com/news164887377.html (2009).

About the Author

Yueh-Hsuan Weng received his MS degree from the Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan, in 2007. He was a Project Assistant in the National Science Council and a member of the administrative staff in the Ministry of the Interior, Taiwan. He is currently a PhD student at Peking University Law School, Beijing, P. R. China, and the Chief Researcher of the Yushan Studio for Artificial Intelligence and Law (YSAiL; http://www.yhweng.tw). His research interests are in issues concerning the interface between advanced technology and law, including AI and law, robot legal studies, legal informatics, computational social sciences, and intellectual property management.

Journal Pre-proof



Robots are Friends as Well as Foes: Ambivalent Attitudes Toward Mindful and Mindless AI Robots in the United States and China

Jianning Dang, Li Liu

PII: S0747-5632(20)30359-9

DOI: https://doi.org/10.1016/j.chb.2020.106612

Reference: CHB 106612

To appear in: *Computers in Human Behavior*

Received Date: 22 July 2020

Revised Date: 6 October 2020

Accepted Date: 26 October 2020

Please cite this article as: Dang J. & Liu L., Robots are Friends as Well as Foes: Ambivalent Attitudes Toward Mindful and Mindless AI Robots in the United States and China *Computers in Human Behavior*, https://doi.org/10.1016/j.chb.2020.106612.

Credit Author Statement

JD and LL contributed to all aspects of work for this article, including conceptualization, data collection and analysis, and writing.

Robots are Friends as Well as Foes: Ambivalent Attitudes Toward Mindful and Mindless AI Robots in the United States and China

Running Head: AMBIVALENT ATTITUDES TOWARD ROBOTS

Jianning Dang, Li Liu

Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education (Beijing Normal University), Faculty of Psychology, Beijing Normal University, Beijing, China

J. Dang

E-mail: jndang@bnu.edu.cn

L. Liu (Corresponding Author)

E-mail: l.liu@bnu.edu.cn

Faculty of Psychology, Beijing Normal University, 19 Xinjiekouwai Street, Beijing 100875, China

Acknowledgments

The authors gratefully acknowledge the financial support provided by the Major Project of the National Social Science Foundation of China (18ZDA332), the Fundamental Research Funds for the Central University (2019NTSS30) and the China Postdoctoral Science Foundation

(2020M670188; 2020T130064). The funders had no role in study design, data collection, decision to publish, or preparation of the manuscript. We thank Anita Harman, PhD, from Liwen Bianji, Edanz Editing China (www.liwenbianji.cn/ac), for editing the English text of a draft of this manuscript.

Robots are Friends as Well as Foes: Ambivalent Attitudes Toward Mindful and Mindless AI Robots in the United States and China

Abstract

In light of the ongoing and rapid development of innovative technologies, two intriguing issues arise: do people have more positive or more negative attitudes toward robots with high (versus low) mental capabilities, and do attitudes toward robots differ between Western and East Asian cultures? Past work on these topics has produced contradictory results. Inspired by the perspective that attitudes are ambivalent rather than bipolar, we argue that these controversial findings stem from people's ambivalent attitudes toward robots. To test the assumption that ambivalent attitudes toward robots differ by type of robots and by cultural background, we conducted an experimental study. By manipulating the level of robot mind and recruiting both American and Chinese participants, we examined how robot mind and culture influence ambivalent attitudes toward robots. We simultaneously measured participants' perceptions of robots as "ally" or "enemy". The results revealed that robots with high (versus low) mental abilities elicited more ambivalent attitudes and that American participants reported more ambivalence toward robots than Chinese participants. These findings enhance our understanding of human–robot interaction and provide guidance for modulating people's attitudes toward robots.

Keywords: artificial intelligence, human–robot interaction, ambivalent attitudes, mind, culture

"The rise of powerful AI will be either the best or the worst thing ever to happen to humanity."

--Stephen Hawking

As the most recognizable carrier of artificial intelligence (AI), advanced robotics have proven applicable in a wide variety of fields, including social and economic development, education, healthcare, and even military operations (Breazeal, 2003; Fong, Nourbakhsh, & Dautenhahn, 2003; Taddeo & Floridi, 2018). The appropriate implementation of AI and robots can enable people to enjoy the benefits that technology provides, but at the same time AI robots are innovative technologies that require careful management (Bigman, Waytz, Alterovitz, & Gray, 2019). For example, in the context of healthcare, AI can assist physicians to make diagnostics and treatment decisions and increase the quality of care but also raise people's fears of getting marginalized (Jha & Topol, 2016). Similarly, in classrooms, educational robots can be used in diverse learning settings to keep students engaged (Reich-Stiebert & Eyssel, 2013), but they might replace interpersonal relationships (Reich-Stiebert & Eyssel, 2016). Furthermore, given the forecast of the soaring number of robots used in workplaces (Manyika et al., 2017), attitudes towards robots are closely related to people's wellbeing in the process of human-robot interaction. Therefore, people are becoming aware of the need to consider their attitudes toward robots (Nitto, Taniyama, & Inagaki, 2017).

Unlike Stephen Hawking's polarized thinking about AI, ordinary people's attitudes toward robots seem ambivalent (Maier, Jussupow, & Heinzl, 2019). While enjoying the convenience and well-being provided by robots (e.g., Reich-Stiebert, Eyssel, & Hohnemann, 2019; Waytz, Heafner, & Epley, 2014), people also worry about the potential threats and

challenges to humanity that they pose (e.g., Stein, Liebold, & Ohler, 2019; Stein & Ohler, 2017). However, most extant research regards attitudes as univalenced and overlooks the complexity of people's attitudes toward robots. Theoretically, ignoring the ambivalent nature of attitudes leads to controversy around attitudes toward robots. For instance, do robots with greater mental abilities elicit more attraction or more aversion (Gray & Wegner, 2012; Liu & Sundar, 2018; Bigman & Gray, 2018)? Are robots more accepted in East Asian versus Western cultures (Bartneck & Hu, 2004; Bartneck, Nomura, Kanda, Suzuki, & Kato, 2005)? Practically, the lack of understanding about ambivalent attitudes hampers measures to facilitate people's willingness to use robots to assist them, as it is ambivalent attitudes that underly people's hesitance to use technology agents (Stein, Newell, Wagner, & Galliers, 2015). To address the above issues and to better understand human–robot interaction processes, the present study examined how ambivalent attitudes toward robots are influenced by the robotic objects (i.e., robots with different levels of mind) and the human subjects (i.e., people from different cultures).

Attitudes Toward AI Robots

AI refers to "a growing resource of interactive, autonomous, self-learning agency, which enables computational artifacts to perform tasks that otherwise would require human intelligence to be executed successfully" (Taddeo & Floridi, 2018, p. 751). AI includes several exemplars, among which robots are the most recognizable (Clarke, 2019). Robots are machines that are programmed by AI and operate semi or fully autonomously to perform tasks done traditionally by humans (Clarke, 2019). Therefore, AI robots can be further seen as machines with humanlike mental capacities (Gray & Wegner, 2012).

Attitudes toward AI robots, namely an overall evaluation of AI robot, is an increasingly attractive topic area for researchers, roboticists, and policy-makers (e.g., Maier et al., 2019;

Vanman & Kappas, in press). Like attitudes toward other objects, attitudes toward robots or other technology agents encompass people's overall evaluations as well as their cognitive, affective, and behavioral components. For example, Nomura and colleagues (Nomura, Kanda, Suzuki, & Kato, 2004) developed a Negative Attitude toward Robots Scale (NARS) to assess people's general negative attitudes toward robots. Other researchers have focused on specific components of attitudes toward AI robots. For instance, cognitive evaluations of robots have been operationalized as warmth/competitiveness and competence judgments about or perceived believability of robots (Bergmann, Eyssel, & Kopp, 2012; Demeure, 2011; Demeure, Niewiadomski, & Pelachaud, 2010; Fraune et al., 2017). In terms of affective responses, prominent negative feelings toward robots have resulted in descriptions such as "uncanny" or "eerie" (Gray & Wegner, 2012; Liu & Sundar, 2018; MacDorman & Chattopadhyay, 2016; Mori, 1970; Stein & Ohler, 2017), and "realistic" or a "symbolic threat" (MacDorman & Entezari, 2015; Stein et al., 2019). Positive reactions to robots have included concepts such as "likeability" (Mathur & Reichling, 2016) and a sense of security (Waytz, Cacioppo, & Epley, 2010; Waytz et al., 2014). Additionally, researchers have examined people's behaviors or behavioral willingness toward robots, such as social interaction with robots (Bickmore & Picard, 2003; Heerink, Krose, Evers, & Wielinga, 2006; Mou & Xu, 2017) and support for robotic research (Yogeeswaran et al., 2016). Furthermore, researchers have measured biological indicators of attitudes toward robots, such as heart rate (Waytz et al., 2014).

Although these and other studies have attracted considerable interest, consistent conclusions have not been reached about important issues around people's attitudes toward robots. One issue that has produced contradictory results concerns attitudes toward robots with greater mental capabilities. Because of the human-like appearance or performance of robots,

people may perceive them as having minds or mental capabilities (Gray & Wegner, 2012; Torres, 2019; Waytz et al., 2010). Some researchers have found that more negative attitudes are associated with robots that are perceived to have a higher level of mind. For example, people perceive human-like robots as having the ability to feel, which engenders a greater sense of eeriness (Gray & Wegner, 2012; MacDorman & Chattopadhyay, 2016). When people ascribe robots with autonomy or the ability to think, they are more likely to perceive them as realistic (i.e., lifelike) and/or as a symbolic threat (Clarke, 2019; MacDorman & Entezari, 2015; Stein et al., 2019) and to express more opposition to robotics research (Złotowski et al., 2017). By contrast, other research suggests that people express more positive attitudes toward robots that are perceived to have high levels of minds. For instance, users express themselves more to interface robots that seem to have communication capabilities (Bickmore & Picard, 2003; Heerink et al., 2006). Similarly, drivers express more trust for autonomous vehicles that appear to have better self-control (Waytz et al., 2014).

Another controversial issue relating to people's attitudes toward robots is the influence of cultural differences. Popular opinion holds that Western and East Asian cultures differ in their views about what is human, which then determines their distinct attitudes toward robots (Geraci, 2006; MacDorman, Vasudevan, & Ho, 2008). Specifically, people in Western cultures hold static beliefs that regard humans as unique and are more likely to highlight the distinction between humans and non-human entities; conversely, people in East Asian cultures have a dynamic perspective that views all things as having a spirit and are less likely to regard humans as particularly special as is more common in Western cultures (Kaplan, 2004; Kitano, 2007). Therefore, robots are generally less accepted in Western cultures than in East Asian cultures (Bartneck & Hu, 2004; Kaplan, 2004; Lee & Šabanović, 2014), as they inhabit the boundary

between human and machine (Turkle, 2007). However, some cross-cultural studies have challenged this popular belief, showing that US participants have fewer negative attitudes toward robots than Japanese and Chinese participants (Bartneck et al., 2005; Bartneck, Suzuki, Kanda, & Nomura, 2007). A possible reason for this result is those US participants were familiar with the technology (Bartneck et al., 2007). Other research findings also revealed no cultural differences in attitudes toward robots (Haring, Mougenot, Ono, & Watanabe, 2014).

One common characteristic of all the aforementioned research is the use of bipolar measures to assess attitudes toward robots. For example, researchers assessed support for robotics research on a scale ranging from "extremely opposing" to "extremely favoring" (Yogeeswaran et al., 2016) and measured participants' reactions to the "uncanny" nature of robots on a scale from "not at all" to "extremely" (Gray & Wegner, 2012). The key limitation of bipolar attitude measurements is their failure to capture conflicting attitudes toward an object; such conflict is a common phenomenon (Kaplan, 1972; Thompson, Zanna, & Grinffin, 1995). Therefore, previous contradictory results may, in part, originate from the ambivalent attitudes that people have toward robots, which cannot be adequately evaluated by measuring only their positive or negative attitudes.

Ambivalent Attitudes Toward AI Robots

Although the traditional perspective assumes that attitudes are unidimensional, it is increasingly clear that attitude is a bidimensional construct (Armitage & Conner, 2004; Van Harreveld, Van Der Plogt, & De Liver, 2009). Across its cognitive, affective, and behavioral components, attitudes include two dimensions differing in valence: positive and negative, which are independent of each other (Thompson et al., 1995). That is, an individual can have positive and negative attitudes simultaneously toward an object, which is known as attitudinal

ambivalence (Conner & Spark, 2002). When the difference between the strength of two conflicting attitudes decreases, attitudinal ambivalence increases. Ambivalence can occur either within (e.g., perceiving robots as competent but competitive) or between multiple components of attitudes (e.g., perceiving robots as competent, but at the same time experiencing a feeling of job threat). Nevertheless, attitudinal ambivalence is cognitive in nature (Eagly & Chaiken, 1993), because cognitive appraisals are the basis of people's emotional and behavioral tendencies toward objects (Cuddy, Fiske, & Glick, 2007).

Attitudinal ambivalence has been overlook in the research about attitudes toward AI due to theoretical and methodological reasons. First, models of attitudes toward technology only capture either positive or negative evaluations of technology. For instance, the Computers as Social Actors paradigm taps positive attitudes toward technology (e.g., trustworthiness, Lee & Nass, 2010), while the Uncanny Valley Theory taps negative attitudes toward robots (e.g., eeriness, Mori, 1970). Second, the univalent scales of attitudes toward AI robots used in previous research (e.g., Latikka, Turja, & Oksanen, 2019) did not provide participants with the opportunity to report their ambivalent attitudes; and neutral scores on these scales may fail to distinguish "neither positive nor negative" attitudes from "equally positive and negative" attitudes (Thompson et al., 1995). The present study argues that ambivalent attitudes toward AI robots may account for the contradictory results of past work on attitudes toward robots.

Ambivalent attitudes toward mindful versus mindless robots

There is evidence suggesting that people have more ambivalent attitudes toward mindful robots than toward mindless ones (see Vanman & Kappas, in press, for a review). In line with Computers as Social Actors theory (Reeves & Nass, 1996), robots with high (versus low) levels of mental abilities are seen as capable of planning and performing their actions and thus as more

competent to serve humans (Isbister & Nass, 2000; Nass & Moon, 2000). Furthermore, the Uncanny Valley Theory (Mori, 1970) and the Model of Autonomous Technology Threat (Stein et al., 2019) posit that robots with humanlike "minds" are also more likely to induce a sense of threat in individuals, including realistic threat (i.e., loss of jobs and resources) and symbolic threat (i.e., loss of human uniqueness), and feelings of eeriness. Taken these two theoretical branches together, mindful robots are perceived as both competent assistants and strong competitors to humans, while mindless robots are seen as incompetent assistants whose competitiveness is relatively weak. Stated differently, people have both strong positive and strong negative attitudes toward mindful robots, whereas they have weak positive and weak negative attitudes toward mindless robots. Therefore, it is reasonable to assume that mindful robots, compared with less mentally capable robots, elicit both more positive and more negative attitudes.

Additionally, Bigman and Gray (2018) provided empirical evidence of more ambivalent attitudes toward mindful robots than mindless ones. In their research, robots' limited thinking and experiencing capacities induced aversion toward them in participants. However, when robots were made to seem more compassionate (i.e., with higher levels of mind), people's attitudes toward them did not change. One explanation for this surprising result is that robots' capacity to experience emotions may promote participants' trust in them, while simultaneously unnerving participants (Gray & Wegner, 2012). Stated differently, when robots were perceived as having higher levels of mind, participants experienced two concurrent paradoxical attitudes (i.e., trusting and unnerved), and thus did not change their attitudes toward those robots. On the basis of the foregoing theoretical and empirical analysis, we formed the following hypothesis.

Hypothesis 1: Mindful robots elicit more ambivalent attitudes than mindless robots.

Cultural differences in ambivalent attitudes toward robots

Unlike previous research that only emphasizes religious beliefs (e.g., Kitano, 2007), we argue that it is necessary to approach the cultural differences in attitude toward robots from different angles. On one hand, people's expectations of robots differ across cultures. Compared with people in East Asian cultures, those from Western cultures tend to take a more functional view of robots and to treat robots as pragmatic assistants (Lee & Šabanović, 2014; Ray, Mondafa, & Siergwart, 2008). People in Western (versus East Asian) cultures have more personal contact with and invest more heavily in service robots (e.g., cleaning robots) than robots designed to communicate with humans (i.e., social robots, Lee & & Šabanović, 2014). Therefore, it indicates that compared to Este Asians, Westerners are more likely to deem robots assistants.

On the other hand, due to philosophical, religious, and cognitive factors, robots elicit more negative attitudes in Westerners (versus East Asians) cultures. First, Westerners and East Asians adopt different beliefs about the mind or spirit of robots (MacDorman et al., 2008). Influenced by Christianity, especially the creation story of Genesis, Western cultures regard humans as unique entities with mental privilege. On the contrary, influenced by Buddhism and Confucianism, people in East Asia believe that spirits can inhabit objects other than humans. Therefore, robots with humanlike characteristics provoke more of a challenge to humanness among Western cultures than they do among East Asian cultures (Geraci, 2006). Second, people in Western and East Asian cultures differ a lot in cognitive patterns (see Varnum et al., 2010 for a review). More specifically, Westerners demonstrate an analytic cognitive style that is characterized by rules-based categorization of objects, whereas East Asians demonstrate a holistic cognitive style that is characterized by categorizing objects based on overall similarity. Stated differently, the human-machine dualism is more pronounced among Westerners than

among East Asians. Since robots are machines with human qualities, they engender more cognitive disfluency for Westerners than for East Asians, which is confirmed to be associated with unfavorable evaluations (Winkielman, Halberstadt, Fazendeiro, & Catty, 2006).

Taken together, compared with those in East Asian cultures, Western people are more aware of how pragmatically complementary robots are and also of how much robots violate their deep-rooted assumptions of humanness. There is also evidence indicating more ambivalent expectations of robots in Western versus East Asian cultures. For instance, robots portrayed in South Korean media, a representation of East Asian cultures, are both highly competent and warm, while in American media, robots are portrayed as competent but cold (Lee & Šabanović, 2014). Consequently, we formed the following hypothesis.

Hypothesis 2: People from Western cultures express more ambivalent attitudes toward robots than people from East Asian cultures.

Overview of the Present Study

While previous research has revealed ambivalent attitudes toward robots, the present study is unique in that it focused on how ambivalent attitudes vary across robots with different levels of mind and across different cultures. To tap ambivalent attitudes, we asked participants to evaluate robots on items with positive and negative valences simultaneously. More specifically, positive and negative attitudes toward robots were operationalized as the cognitive appraisals, that is, the perceptions of robots as allies or enemies. We chose image perception as an indicator of attitudes on the basis of the following considerations. First, the perception of images is dictated by the interaction between the perceived friendliness/hostility and perceived strength/weakness of the object (Alexander, Brewer, & Herrmann, 1999). Second, image

perception can determine what people's emotional and behavioral tendencies are toward those objects (Cuddy et al., 2007).

Two hypotheses were proposed. First, we hypothesized that people have more ambivalent attitudes toward mindful robots than toward mindless robots (*Hypothesis 1*). To test this hypothesis, we manipulated the level of robots' minds and compared people's attitudes toward mindful and mindless robots. Second, we hypothesized that people in Western cultures have more ambivalent attitudes toward robots than those in East Asian cultures (*Hypothesis 2*). To test this hypothesis, we recruited both American and Chinese participants and compared their attitudes toward robots. It is worth noting that we do not equate America and China with general Western and East Asian cultures, rather, we take these two countries as representatives of these two cultures. Previous research reveals that American and Chinese people differ a lot in social orientation (e.g., independence versus interdependence, Hofstede, 1979) and cognitive patterns (i.e., analytic versus holistic cognition, Morris & Peng, 1994), which are the key cultural differences between Westerners and East Asians (Varnum, Grossmann, Kitayama, & Nisbett, 2020).

Methods

Participants

As we were unsure what effect size to expect, the smallest effect size of interest (SESOI) of $\eta^2 = .03$ was used in sample-size planning (da Silva Frost & Ledgerwood, 2020). A *priori* power analysis with a 0.05 alpha level indicated that a total sample size of 340 would provide 90% power to detect the main effect of robot mind or culture on ambivalent attitudes toward robots. We recruited 192 American participants via Amazon Mechanical Turk (https://www.mturk.com/). Of these, 21 were omitted because they failed to answer the check

items correctly. Through ePanel [1] (https://www.epanel.cn/), 206 Chinese participants were recruited. The final sample included 379 participants (222 male, 18–72 years old, M_{age} = 34.95, *SD* = 9.38). Demographic statistics were similar among American (*n* = 173, 60.1% male, M_{age} = 35.42, *SD* = 11.38) and Chinese samples (*n* = 206, 57.3% male, M_{age} = 34.56, *SD* = 7.29). However, approximately 68.8% of American participants reported being White or Caucasian, 5.2% African American, 17.9% Asian[2], and 5.8% Latino American. Participants were remunerated with $0.5 (Americans) or 3 CNY (approximately US $0.43; Chinese).

Materials

American and Chinese participants answered questionnaires in English and Chinese, respectively. Two bilingual researchers translated all the materials in the present study from the original English version into Chinese and conducted a back-translation to check the equivalence.

Robot mind manipulation

Drawing on previous research (Gray & Wegner, 2012; Waytz et al., 2014), robot mind manipulation consisted of presenting robots with different levels of mental capabilities. Participants were concurrently presented with two types of AI robots: mindful and mindless. Mindful AI robots were described as follows: “Mindful AI robots can feel the outside world like humans, experience various emotions in social interaction, and express their own experiences and emotions. They can independently plan ahead, use strategies to solve problems, and use natural language for communication.” Mindless AI robots were depicted as follows: “Mindless AI robots have limited ability to feel the outside world and they cannot experience and express

[1] EPanel is a professional Chinese online survey register that functions similarly to Amazon Mechanical Turk.

[2] The significance levels of the results were the same with or without the inclusion of the Asian American participants.

various emotions in social interaction. They execute actions according to human instructions and have a limited level of autonomy. They cannot communicate with natural language." This paradigm was demonstrated as an effective manipulation of robot mind (Gray & Wegner, 2012; Waytz et al., 2014) as it could induce a large size of effect on mind perception (η^2 = .10~.97).

Robot mind manipulation check

To ensure that participants understood the difference between mindful and mindless AI robots in terms of mind, they were asked to rate each type on 12 items that assessed robot mind (Gray, Gray, & Wegner, 2007). Six items measured robots' thinking capacity (e.g., "How capable of exercising self-control do you think robots are?") and six measured robots' feeling capacity (e.g., "How capable of feeling pleasure do you think robots are?"). All items (see Appendix A for all items) were answered on a 7-point scale (1 = not at all, 7 = totally). The scores of all items were averaged in a robot mind index (Cronbach's α = .86 for mindful AI robots and Cronbach's α = .91 for mindless AI robots), and higher scores indicated higher levels of perceived mind in robots.

Perception of robot images

Responses to images of robots were measured by four adapted items developed by Riketta (2005). Two items assessed the extent to participants deemed a robot an ally (e.g., "Consider only the positive aspects about AI robots and ignore the negative ones, I think that the relationship between humans and robots is harmonious"), while the other two items assessed the extent to which participants deemed a robot an enemy (e.g., "Consider only the negative aspects about AI robots and ignore the positive ones, I think that AI robots compete with humans"). All the items (see Appendix B for all items) were answered on a 7-point scale (1 = completely

disagree, 7 = completely agree). The average ratings for the ally (Cronbach's α = .84 for mindful AI robots and Cronbach's α = .87 for mindless AI robots) and enemy image items (Cronbach's α = .73 for mindful AI robots and Cronbach's α = .79 for mindless AI robots) were computed. Higher scores indicated more intense attitudes toward robots.

Procedure

After providing informed consent, participants were told that they would complete a task about social judgments. Participants first read descriptions of mindful and mindless AI robots and learned about their characteristics. Then, as a manipulation check, participants rated mindful and mindless AI robots on items regarding mental capabilities. Finally, participants completed the image ratings for each type of robot. Mindful and mindless robots were rated in random order.

Statistical Analysis

We examined people's ambivalent attitudes toward robots in two ways. First, according to the Griffin formula of attitudinal ambivalence (Thompson et al., 1995): $(P+N)/2 - |P-N|$[3], greater ambivalence was represented as stronger intensity in the perception of a robot as ally and enemy (i.e., (P+N)/2) and more similarity in the comparative intensity reported for two given images (i.e., |P-N|). Attitudinal ambivalence was computed using the Griffen formula. Second, to fully illustrate how ambivalent attitudes arise, we examined the relative levels of perceptions of robots as ally and enemy. Individuals with stronger positive and negative attitudes are more ambivalent (Maier et al., 2019; Thompson et al., 1995).

[3] P indicates intensity of positive attitude and N indicates intensity of negative attitude. In the current study, P and N separately indicate the intensity of participants' perceptions of robots as ally or enemy.

Results

Preliminary Analysis

To confirm that robot images perceived as ally or enemy were two distinct constructs, we conducted two confirmatory factor analyses. Two models were established: the single-factor model (i.e., for both mindful and mindless AI robots, all image items loaded on one factor) and the two-factor model (i.e., for both mindful and mindless AI robots, ally image items loaded on one factor and enemy image items loaded on another factor). While the single-factor model did not fit the data well, χ^2 = 476.53, df = 19, p < .001, CFI = 0.64, TLI = 0.48, RMSEA = 0.25, SRMR = 0.15, the two-factor model did, χ^2 = 40.93, df = 15, p < .001, CFI = 0.98, TLI = 0.96, RMSEA = 0.07, SRMR = 0.04. The two-factor model, therefore, was better than the single-factor model, $\triangle\chi^2$ = 435.61, $\triangle df$ = 5, p < .001. The results indicated that ally and enemy images of robots were two distinct constructs.

Robot mind Manipulation Check

To examine the effectiveness of the manipulation of robot mind, a 2 (robot type: mindful, mindless) × 2 (culture: American, Chinese) mixed-ANOVA was conducted on mind perception, with robot type as the within-subject variable. The results (see Table 1) revealed a significant main effect of robot type, F (1, 377) = 272.65, p < .001, η^2 = .42, 90% CI[4] [.36, .47], suggesting that mindful AI robots were perceived as having higher levels of mind than mindless AI robots. The main effect of culture was not significant, F (1, 377) = 0.10, p = .750, η^2 < .001. The two-

[4] As suggested by Steiger (2004), 90% CIs would be more appropriate for η^2. The 90% CI does exclude zero, but barely; a 95% CI would include zero. Furthermore, η^2 cannot be less than zero. Accordingly, Steiger argued that when putting a CI on an ANOVA effect that has been tested with the traditional 0.05 criterion of significance, that CI should be a 90% CI, not a 95% CI.

way interaction effect size was small, albeit statistically significant, F (1, 377) = 4.15, p = .042, η^2 = .01, 90% CI [.00, .03]. Further simple effect analyses were conducted to explain the interaction effect. American and Chinese participants gave similar ratings on both mindful (F (1, 377) = 2.47, p = .117, η^2 = .01) and mindless AI robots (F (1, 377) = 0.65, p = .421, η^2 = .002). Accordingly, the manipulation of robot mind was deemed successful.

Table 1. Perceptions of mind Attributed to Mindful and Mindless AI Robots

Robot type	All sample	American sample	Chinese sample
Mindful robots	4.67 (1.11)	4.57 (1.13)	4.75 (1.10)
Mindless robots	3.47 (1.36)	3.53 (1.50)	3.42 (1.22)

Ambivalent Images of Robots as Functions of Robot Mind and Culture

Two separate ANOVAs were conducted on the ambivalence reported toward images of robots, with robot type (mindful vs mindless) and culture (American vs Chinese) as independent variables. Supporting Hypothesis 1, a repeated-measures ANOVA revealed a significant effect of robot type, F (1, 378) = 33.55, p < .001, η^2 = .08, 90% CI [.04, .12], with participants perceiving mindful AI robots (M = 2.36, SD = 1.90) more ambivalently than mindless AI robots (M = 1.79, SD = 2.06). In line with Hypothesis 2, a one-way ANOVA revealed a significant effect of culture, F (1, 377) = 50.54, p < .001, η^2 = .12, 90% CI [.07, .17], suggesting that American participants (M = 2.72, SD = 1.72) perceived robots more ambivalently than Chinese participants (M = 1.53, SD = 1.54).

In addition, 2 (robot type: mindful vs mindless) × 2 (culture: American vs Chinese) mixed ANOVA was conducted on attitudinal ambivalence, although we had no hypothesis about the interaction effect between robot type and culture. The results (see details in supplementary

materials) revealed significant main effects of robot type and culture, which supported Hypothesis 1 and Hypothesis 2. Most importantly, the two-way interaction (F (1, 377) = 4.76, p = .030, η^2 = .01, 90% CI [.00, .04]) was very small in size, although it reached statistical significance, suggesting that cultural differences in ambivalent attitudes toward robots did not vary on the type of robots.

Ally and Enemy Images of Robots as Functions of Robot Mind and Culture

To further illustrate the ambivalent attitudes as manifested in both strong ally and enemy perceptions of robots, two separate ANOVAs were conducted on (ally and enemy) images ratings, with the robot type and culture as independent variables. We expected that mindful (versus mindless) AI robots elicit simultaneously stronger opposing images and that American participants report stronger opposing images of robots than Chinese participants.

Ally and enemy images of robots as functions of robot mind

To examine how ambivalent attitudes toward robots differed by robot mind, we conducted a 2 (robot type: mindful, mindless) × 2 (image: ally, enemy) repeated-measures ANOVA among all participants. The main effect of robot type was significant, F (1, 378) = 48.64, p < .001, η^2 = .11, 90% CI [.07, .17], indicating that the intensity felt toward images (i.e., ally or enemy) of mindful AI robots (M = 4.34, SD = 0.80) was higher than that reported for mindless AI robots (M = 4.03, SD = 0.92). The main effect of image (F (1, 378) = 120.73, p < .001, η^2 = .24, 90% CI [.18, .30]) suggested that robot images elicited more perceptions of ally (M = 4.74, SD = 1.19) than of enemy (M = 3.62, SD = 1.29). Most importantly, the two-way interaction (F (1, 378) = 35.92, p < .001, η^2 = .09, 90% CI [.05, .13]) was significant. Simple effect analyses (see Figure 1) revealed that the magnitude of difference between ally and enemy

images of mindful AI robots (*M* = 4.68, *SD* = 1.43 for ally image and *M* = 3.99, *SD* = 1.48 for enemy image) was less than that of mindless AI robots (*M* = 4.80, *SD* = 1.51 for ally images and *M* = 3.25, *SD* = 1.53 for enemy images).

Referring to Thompson et al.'s (1995) formula of attitudinal ambivalence, the results suggested that mindful (versus mindless) AI robots elicited stronger ally and enemy perceptions (i.e., (P+N)/2), and these two image perceptions were more similar in intensity (i.e., |P-N|). Stated differently, mindful (versus mindless) AI robots elicited simultaneously stronger, opposing images and thereby more attitudinal ambivalence. Therefore, supporting Hypothesis 1, participants expressed more ambivalent attitudes toward mindful (versus mindless) AI robots.

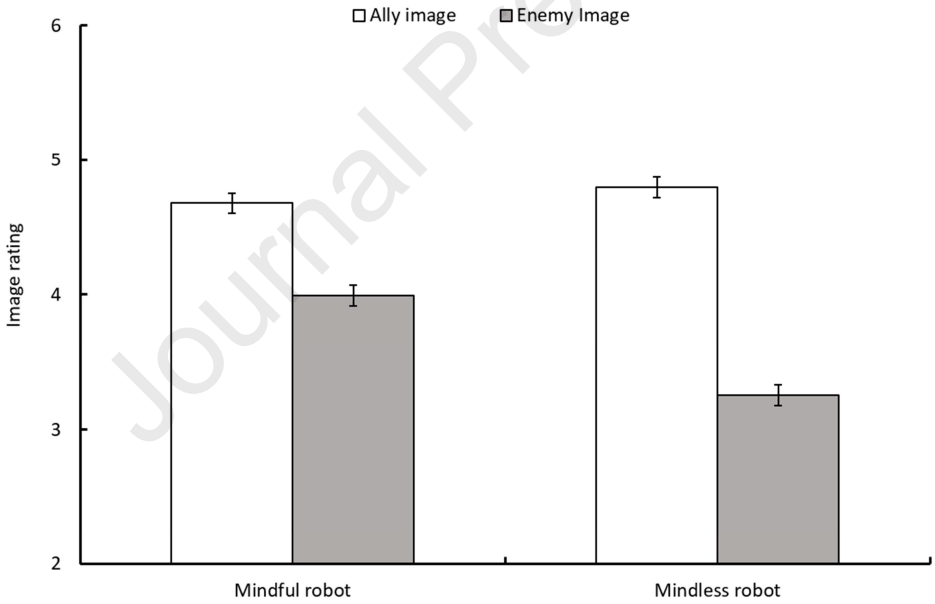


Figure 1. Ally and Enemy Images of Robots as a Function of Robot Mind

Ally and enemy images of robots as functions of culture

To examine how ambivalent attitudes toward robots differed by cultures, we conducted a 2 (culture: American, Chinese) × 2 (image: ally, enemy) mixed ANOVA on image ratings of

both types of AI robots, with image as the within-subject variable. The absence of the main effect of culture (F (1, 377) = 0.39, p = .533, η^2 = .001) suggested that the intensity of image perceptions was equal across American (M = 4.16, SD = 1.00) and Chinese cultures (M = 4.20, SD = 0.43). AI robots were rated more as allies (M = 4.72, SD = 1.19) than enemies (M = 3.62, SD = 1.29), F (1, 377) = 118.91, p < .001, η^2 = .24, 90% CI [.18, .30]. As expected, the interaction between culture and image (F (1, 377) = 33.73, p < .001, η^2 = .08, 90% CI [.04, .13]) was significant. Simple effect analyses (see Figure 2) revealed that the magnitude of difference between ally and enemy images of AI robots among American participants (M = 4.40, SD = 1.14 for ally image and M = 3.91, SD = 1.32 for enemy image) was less than that among Chinese participants (M = 5.02, SD = 1.15 for ally images and M = 3.39, SD = 1.29 for enemy images).

Taken together, although American (versus Chinese) participants reported equally strong perceptions (i.e., (P+N)/2), the comparative intensity reported for ally and enemy images of AI robots (i.e., |P-N|) among American participants was smaller than that among Chinese participants. Therefore, supporting Hypothesis 2, American (versus Chinese) participants expressed more ambivalent attitudes toward AI robots.

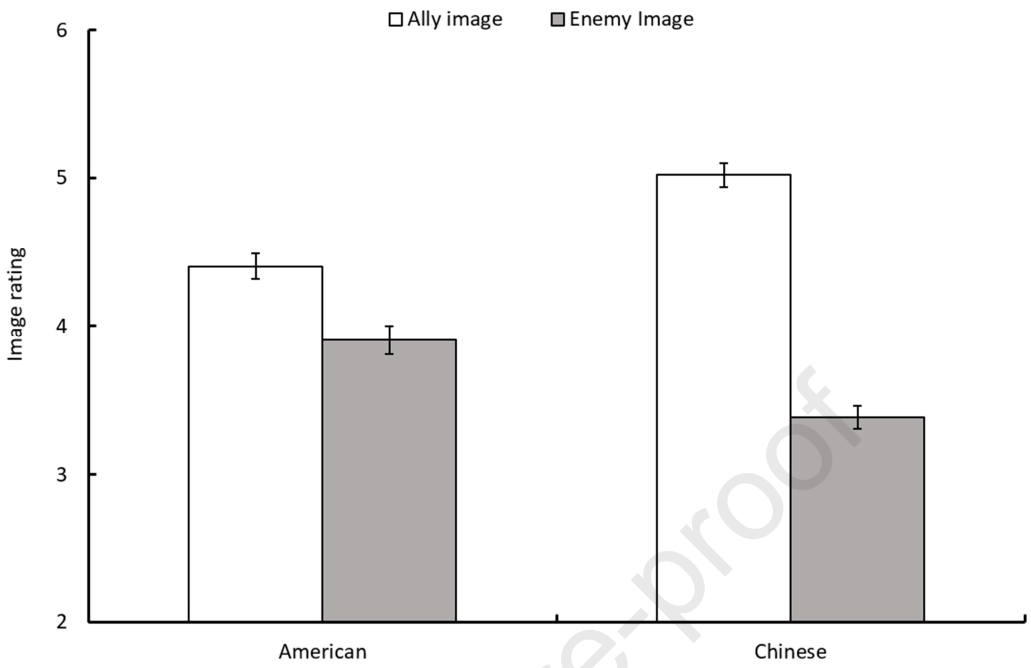


Figure 2. Ally and Enemy Images of Robots as a Function of Culture

In addition, 2 (robot type: mindful, mindless) × 2 (culture: American, Chinese) × 2 (image: ally, enemy) mixed-ANOVA was conducted on image ratings, with robot type and image as within-subject variables. The results (see details in complementary materials) revealed that across American and Chinese participants, mindful (versus mindless) AI robots elicited more attitudinal ambivalence and that for both mindful and mindless AI robots. Moreover, American participants had more ambivalent attitudes toward robots than Chinese participants. Thus, Hypotheses 1 and 2 were verified.

Discussion

People generally believe that AI can be useful; for example, Travelzoo's (2016) global research revealed that around 75% of their customers believe that AI devices can increase hotel service quality. Nevertheless, people are just beginning to harness AI technology, and the use of

AI tends to be relatively low (e.g., Reich-Stiebert & Eyssel, 2013, 2016). In the McKinsey Global Survey, only 21% of employee respondents reported their organizations had embedded AI into multiple business units, and 30% reported piloting the use of AI in the workplace (McKinsey & Company, 2018). One key reason for the hesitance to adopt robots is people's ambivalent attitudes toward new technology (Stein et al., 2015). The present study assessed how participants from Western and East Asian cultures perceive robots with high and low mental capacities—as allies or enemies. The results supported our hypotheses: robots with high (versus low) mental capacities elicit more ambivalent attitudes (Hypothesis 1), and American participants have more ambivalent attitudes toward robots than Chinese participants (Hypothesis 2).

Our results help to disentangle the seemingly contradictory findings of previous research about attitudes toward robots. Past work regarded attitudes as bipolar and assumed that people have either positive or negative attitudes toward robots (e.g., Stein et al., 2019; Waytz et al., 2014). By measuring both positive and negative attitudes, the present research examined people's ambivalent attitudes toward robots. The results suggest that people strongly view mindful robots as both allies and enemies simultaneously, and that Americans simultaneously like and dislike robots more than Chinese people. Stated differently, no fixed answers exist as to whether mindful AI robots are more favorable than mindless AI robots, and whether robots are more favorable in East Asian versus Western cultures. Furthermore, ambivalent attitudes are closely related to non-acceptance and non-conforming user behaviors (Tsai, Compeau, & Meister, 2017), and this is because attitudinal ambivalence invokes uncertainty-related negative emotions, reduces people's abilities to decide, and increases the incapacity to act (Rothman, Pratt,

Rees, & Vogus, 2017). Therefore, our findings also help explain why people hesitate to use AI robots despite the widely supported development of AI technologies.

Considering both positive and negative attitudes toward robots, the present study also helps unravel the complexity of human–robot interaction. Ambivalent attitudes toward robots with higher (versus lower) mental capacities suggest that viewing the process of adopting mindful AI robots as linear is an oversimplification. This reflects previous studies that have illustrated the multi-faceted aspect of users' relationship with technology products (Jensen & Aanestad, 2007; Van Offenbeek, Boonstra, & Seo, 2013). Furthermore, ambivalent attitudes are related to people's affect and decision-making when interacting with robots (Van Harreveld et., 2009). Specifically, as AI technologies continue to proliferate, and people must decide whether to adopt mindful AI robots, they will be aware of their conflicting attitudes toward robots and will likely suffer uncertainty and discomfort. To reduce attitudinal ambivalence, people may use strategies such as denying the responsibility of adopting robots or deliberately engaging with robot-related information in an effort to alter their attitudes toward robots (Van Harreveld et al., 2009). Therefore, ambivalent attitudes toward robots play an essential role in human–robot interactions.

The cultural differences seen in ambivalent attitudes toward robots help to further clarify how the adoption of robots varies by cultures. It is a prevailing belief that "robot mania" is widespread in East Asian cultures, especially Japan, while more people in Western cultures tend to suffer from technophobia (Kaplan, 2004). However, this belief has not received much empirical support (MacDorman et al., 2008; Haring et al., 2014). The present study indicates that Americans have more ambivalent attitudes, rather than more positive or more negative attitudes, toward robots. The cultural differences in ambivalent attitudes toward robots reflect the

culturally variable aesthetic user preferences for robots (Lee & Šabanović, 2014). Specifically, human-like characteristics (e.g., a human-like appearance and expressive faces) of robots blur the difference between them and humans and elicit more ambivalent attitudes. For example, compared with Korean participants, US participants seem to prefer machine-like robots to human-like robots (Lee & Šabanović, 2014). Thus, ambivalent attitudes toward robots appear to strongly influence people in Western (versus East Asian) societies in their choices to use robots as tools and in industrial settings.

The present study yields practical implications for modulating attitudes toward robots. In contexts wherein robots can complement or assist human resources, such as service and health care (Jha & Topol, 2016), more focused strategies to build on positive attitudes or/and to reduce negative attitudes toward robots can be taken to reduce people's ambivalent attitudes toward mindful robots. As suggested in the Technology Acceptance Model (Davis, 1989), making beneficial technologies easier to use promotes people's positive attitudes toward them. Meanwhile, more explanation is needed to distinguish the working modes of artificial intelligence and human intelligence and increase the transparency about how robots work, which then reduce people's negative attitudes toward robots (Clark, 2019). However, in ethic-sensitive contexts, such as arms race, ambivalent attitudes toward robots can prevent people from abusing robots for bad purposes. Attitudinal ambivalence provides people with conflicting signals that alert them to engage in deeper and more balanced considerations of divergent perspectives (Rothman et al., 2017). Therefore, to a certain extent, ambivalent attitudes can remind people to use robots with caution.

Finally, the present findings also inspire directions for future research. First, in the current study, attitudes toward robots were operationalized as perceptions of robot images. A

person's perception of an image reflects their overall evaluation of the viewed object; nevertheless, it is mainly their cognitive evaluations that are captured (Cuddy et al., 2007). Besides the cognitive ambivalence found in the present study, ambivalent attitudes can also reflect intercomponent ambivalence, which refers to the contradictory valence of different attitude components (Maio & Haddock, 2014). For instance, people perceived robots as allies (i.e., positive belief) but felt nervous when using robots in their work (i.e., negative affect). Therefore, an intriguing direction for future research is to examine the intracomponent and intercomponent ambivalence of the various components of attitudes toward robots. Second, although it has been demonstrated that American and Chinese people differ a lot in social orientation and cognitive patterns (Hofstede, 1979; Morris & Peng, 1994), we did not clearly measure these cultural differences in the current study. Therefore, it may be a better practice to distinguish participants from different cultural backgrounds based on cultural values and beliefs rather than just nationality. Moreover, we selected American and Chinese participants as representatives of Westerners and East Asians respectively. Western cultures also include European countries and East Asian cultures also include other countries, such as Japanese and South Korean. Therefore, future research should recruit participants from more countries and compare within- and cross-cultural differences in ambivalent attitudes toward AI robots.

Open Practices

All data have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/kthgn/files/. The complete materials are included in the Supplemental Online Material associated with this article.

References

Alexander, M. G., Brewer, M. B., & Herrmann, R. K. (1999). Images and affect: A functional analysis of out-group stereotype. *Journal of Personality and Social Psychology*, *77*, 78–93.

Armitage, C. J., & Conner, M. (2004). The effects of attitudinal ambivalence on attitude-intention behavior relations. In G. Haddock, & G. R. Maio (Eds.) *Contemporary perspectives on the psychology of attitudes* (pp.121–143). Psychology Press.

Bartneck, C., & Hu, J. (2004). Rapid prototyping for interactive robots. In F. Groen, N. Amato, A. Bonarini, E. Yoshida, & B. Krose (Eds), *Proceedings of the 8th Conference on Intelligent Autonomous Systems (IAS-8), Amsterdam* (blz. 136–145). Amsterdam: IOS Press.

Bartneck, C., Nomura, T., Kanda, T., Suzuki, T., & Kennsuke, K. (2005). Cultural differences in attitudes towards robots. *Proceedings of the AISB symposium on robot companions: Hard problems and open challenges in human-robot interaction*. Hatfield.

Bartneck, C., Suzuki, T., Kanda, T., & Nomura, T. (2007). The influence of people's culture and prior experiences with Aibo on their attitude towards robots. *AI & Society*, *21*(1/2), 217–230.

Bergmann, K., Eyssel, F., & Kopp, S. (2012). A second chance to make a first impression? How appearance and nonverbal behavior affect perceived warmth and competence of virtual agents over time. *International conference on intelligent virtual agents*. Springer, Berlin, Heidelberg.

Bickmore, T., & Picard, R. W. (2003). Subtle expressivity by relational agents. *Proceedings of the CHI 2011 Workshop on Subtle Expressivity for Characters and Robots*, *3*(5), 1–8.

Bigman, Y. E., Waytz, A., Alterovitz, R., & Gray, K. (2019). Holding robots responsible: The elements of machine morality. *Trends in Cognitive Sciences, 23*(5), 365–368.

Breazeal, C. (2003). Emotion and sociable humanoid robots. *International Journal of Human Computer Studies*, *59*(1–2), 119–155.

Clark, R. (2019). Why the world wants controls over Artificial Intelligence. *Computer Law & Security Review*, *35*(4), 423–433.

Conner, M., & Sparks, P. (2002). Ambivalence and attitudes. *European Review of Social Psychology*, *12*, 37–70.

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology*, *92*(4), 631–648.

Da Silva Frost, A., & Ledgerwood, A. (2020). Calibrate your confidence in research findings: A tutorial on improving research methods and practices. *Journal of Pacific Rim Psychology*, e14.

Demeure, V., Niewiadomski, R., & Pelachaud, C. (2011). How is believability of a virtual agent related to warmth, competence, personification, and embodiment? *Presence*, *20*(5), 431–448.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics & Autonomous Systems*, *42*(3–4), 143–166.

Geraci, R. (2006). Spiritual robots: Religion and our scientific view of the natural world. *Theology and Science*, *4*(3), 229–246.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*, 619.

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, *125*(1), 125–130.

Haring, K. S., Mougenot, C., Ono, F., & Watanabe, K. (2014). Cultural differences in perception and attitude towards robots. *International Journal of Affective Engineering, 13*(3), 149–157.

Heerink, M., Kröse, B. J. A., Wielinga, B. J., & Evers, V. (2006). The influence of a robot's social abilities on acceptance by elderly users. IEEE International Symposium on Robot & Human Interactive Communication.

Hofstede, G. H. (1979). Value systems in forty countries: Interpretation, validation, and consequences for theory. In: Eckensberger, L. H, Lonner, W. J., & Poortinga, Y. H (eds), *Cross-cultural Contributions to Psychology* (pp. 389–407). Lisse, the Netherlands: Swets and Zeitlinger.

Isbister, K., & Nass, C. (2000). Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies, 53*(2), 251–267.

Jha, S., & Topol, E. J. (2016). Adapting to artificial intelligence: Radiologists and pathologists as information specialists? *Journal of the American Medical Association*, 316(22), 2353–2354.

Kaplan, F. (2004). Who is afraid of the humanoid? Investigating cultural differences in the acceptance of robots. *International Journal of Humanoid Robotics, 1*(03), 465–480.

Kaplan, K. J. (1972). On the ambivalence–indifference problem in attitude theory and measurement: A suggested modification of the semantic differential technique. *Psychological Bulletin, 77*, 361–372.

Kitano, N. (2007). *Animism, rinri, modernization: The base of Japanese robotics*. Proceedings of the 8th Conference on Intelligent Autonomous Systems (IAS–8), Amsterdam.

Lee, H., & Šabanović, S. (2014). *Culturally variable preferences for robot design and use in South Korea, Turkey, and the United States*. Proceedings of the ACM/IEEE 9th International

Conference on Human-Robot Interaction. Bielefeld: ACM, 17–24.

Lee, J. R., & Nass, C. I. (2010). Trust in computers: The Computers-Are-Social-Actors (CASA) paradigm and trustworthiness perception in human-computer communication. In Latusek, D., & Gerbasi, A. (Ed.), *Trust and technology in a ubiquitous modern environment: Theoretical and methodological perspectives* (pp. 1–15). IGI Global.

Liu, B., & Sundar, S. S. (2018). Should machines express sympathy and empathy? Experiments with a health advice chatbot. *Cyberpsychology Behavior & Social Networking*, *21*(10), 625–636.

Macdorman, K. F., & Chattopadhyay, D. (2016). Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. *Cognition*, *146*, 190–205.

Macdorman, K. F., & Entezari, S. O. (2015). Individual differences predict sensitivity to the uncanny valley. *Interaction Studies*, *16*(2), 141–172.

Macdorman, K. F., Vasudevan, S. K., & Ho, C. C. (2008). Does Japan really have robot mania? Comparing attitudes by implicit and explicit measures. *AI & Society*, *23*(4), 485–510.

Maier, S. B., Jussupow, E., & Heinzl, A. (2019). *Good, bad, or both? Measuremet of physician's ambivalent attitudes toward AI*. In Proceedings of the 27th European Conference on Information Systems (ECIS), Stockholm & Uppsala, Sweden.

Maio, G. R. & Haddock, G. (2014). *The psychology of attitudes and attitude change (2nd Edition)*. London: SAGE Publications Ltd.

Manyika, J., Chui, M., Miremadi, M., Bughin, J., George, K., Willmott, P., & Dewhurst, M. (2017). *Harnessing automation for a future that works*. Retrieved from http://www.mckinsey.com /global-themes/digital-disruption/harnessing-automation-for-a-

future-that-works

Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition*, *146*, 22–32.

Mori, M. (1970). The uncanny valley. *Energy*, *7*(4), 33–35.

Morris, M. & Peng, K. (1994). Culture and cause: American and Chinese attributions for social and physical events. *Journal of Personality and Social Psychology*, *67*, 949-971.

Mou, Y., & Xu, K. (2017). The media inequality: comparing the initial human-human and human-ai social interactions. *Computers in Human Behavior*, 72, 432–440.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81–103.

Niewiadomski R., Demeure V., Pelachaud C. (2010) Warmth, Competence, Believability and Virtual Agents. In J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, & A. Safonova (Eds.), *Intelligent Virtual Agents. IVA 2010. Lecture Notes in Computer Science, vol 6356*. Springer, Berlin, Heidelberg.

Nitto, H., Taniyama, D., & Inagaki, H. (2017). Social acceptance and impact of robots and artificial intelligence: Findings of survey in Japan, the U.S. and Germany. *NRI Papers*, No. 211.

Nomura, T., Kanda, T., Suzuki, T., & Kato, K. (2004). Psychology in human-robot communication: An attempt through investigation of negative attitudes and anxiety toward robots. Robot and Human Interactive Communication, 13th IEEE International Workshop on IEEE.

Ray, C., Mondada, F., & Siegwart, R. (2008). What do people expect from robots? *In Intelligent robots and systems, 2008 proceedings 2008 IEEE/RSJ international conference* (pp. 3816–3821).

Reeves, B., & Nass, C. (1996). *The media equation: how people treat computers, television, and new media like real people and places.* CSLI Publications and Cambridge University Press.

Reich-Stiebert, N., & Eyssel, F. A. (2013). Attitudes towards service robots in domestic environments: The role of personality characteristics, individual interests, and demographic variables. *Journal of Behavioral Robotics*, *4*, 123–130.

Reich-Stiebert, N., & Eyssel, F. A. (2013). Attitudes towards service robots in domestic environments: The role of personality characteristics, individual interests, and demographic variables. *Journal of Behavioral Robotics, 4*, 123–130.

Reich-Stiebert, N., & Eyssel, F. A. (2016). Robots in the classroom: What teachers think about teaching and learning with education robots. *Lecture Notes in Computer Science, 9979*, 671–680.

Reich-Stiebert, N., Eyssel, F., & Hohnemann, C. (2019). Involve the user! Changing attitudes toward robots by user participation in a robot prototyping process. *Computers in Human Behavior, 91*, 290–296.

Riketta, M. (2005). Cognitive differentiation between self, ingroup, and outgroup: The roles of identification and perceived intergroup conflict. *European Journal of Social Psychology, 35*(1), 97–106.

Scott, W. A. (1966). Measures of cognitive structure. *Multivariate Behavior Research*, *1*, 391–395.

Stein, J. P., & Ohler, P. (2017). Venturing into the uncanny valley of mind—the influence of mind attribution on the acceptance of human-like characters in a virtual reality setting. *Cognition, 160*, 43–50.

Stein, J. P., Liebold, B., & Ohler, P. (2019). Stay back, clever thing! Linking situational control and human uniqueness concerns to the aversion against autonomous technology. *Computers in Human Behavior, 95*, 73–82.

Stein, M. K., Newell, S., Wagner, E. L., & Galliers, R. D. (2015). Coping with information technology: Mixed emotions, vacillation, and nonconforming use patterns. *MIS Quarterly, 39*(2), 367–392.

Taddeo, M. & Floridi, L. (2018). How AI can be a force for good. *Science, 361*, 751–752.

Thompson, M. M., Zanna, M. P., & Griffin, D. W. (1995). *Let's not be indifferent about (attitudinal) ambivalence*. In R. E. Petty & J. A. Krosnick (Eds.), *Ohio State University series on attitudes and persuasion, Vol. 4. Attitude strength: Antecedents and consequences* (pp. 361–386). Lawrence Erlbaum Associates, Inc.

Torres, P. (2019) The possibility and risks of artificial general intelligence. *Bulletin of the Atomic Scientists, 75*(3), 105–108.

Tsai, H. Y., Compeau, D., & Meister, D. (2017). Voluntary use of information technology: An analysis and synthesis of the literature. *Journal of Information Technology, 32*(2), 147–162.

Turkle, S. (2007). Authenticity in the age of digital companions. *Interaction Studies, 8*(3), 501–517.

van Harreveld, F., van der Pligt, J., & de Liver, Y. N. (2009). The agony of ambivalence and ways to resolve it: Introducing the MAID model. *Personality and Social Psychology Review, 13*(1), 45–61.

Van Offenbeek, M. V., Boonstra, A., & Seo, D. B. (2013). Towards integrating acceptance and resistance research: evidence from a telecare case study. *European Journal of Information Systems*, *22*(4), 434–454.

Vanman, E. J., & Kappas, A. (In press). "Danger, will robinson!" The challenges of social robots for intergroup relations. *Social and Personality Psychology Compass*.

Varnum, M. E., Grossmann, I., Kitayama, S., Nisbett, R. E. (2010). The origin of cultural differences in cognition: Evidence for the social orientation hypothesis. *Current Directions in Psychological Science*, *19*, 9–13.

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, *5*(3), 219–232.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*(3), 113–117.

Winkielman, P., Halberstadt, J., Fazendeiro, T., & Catty, S. (2006). Prototypes are attractive because they are easy on the mind. *Psychological Science*, *17*(9), 799–806.

Yogeeswaran, K., Złotowski, J., Livingstone, M., Bartneck, C., Sumioka, H., & Ishiguro, H. (2016). The interactive effects of robot anthropomorphism and robot ability on perceived threat and support for robotics research. *Journal on Human Robot Interaction*, *5*(2), 29–47.

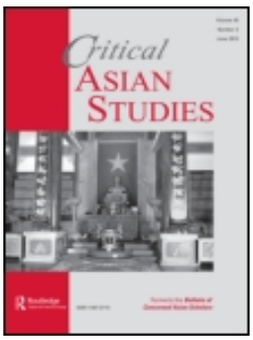
Highlights

People simultaneously have positive and negative attitudes toward robots.

Mindful (versus mindless) robots elicited more ambivalent attitudes.

Americans have more ambivalent attitudes toward robots than Chinese.

The findings shed light on the complexity of human–robot interaction.



Critical Asian Studies

ISSN: 1467-2715 (Print) 1472-6033 (Online) Journal homepage: https://www.tandfonline.com/loi/rcra20

HUMAN RIGHTS VS. ROBOT RIGHTS: Forecasts from Japan

Jennifer Robertson

To cite this article: Jennifer Robertson (2014) HUMAN RIGHTS VS. ROBOT RIGHTS: Forecasts from Japan, Critical Asian Studies, 46:4, 571-598, DOI: 10.1080/14672715.2014.960707

To link to this article: https://doi.org/10.1080/14672715.2014.960707

Published online: 02 Oct 2014.

Critical Asian Studies
46:4 (2014), 571–598
Routledge
Taylor & Francis Group

HUMAN RIGHTS VS. ROBOT RIGHTS

Forecasts from Japan

Jennifer Robertson

ABSTRACT: Japan continues to be in the vanguard of human–robot communication and, since 2007, the state has actively promoted the virtues of a robot-dependent society and lifestyle. Nationwide surveys suggest that Japanese citizens are more comfortable sharing living and working environments with robots than with foreign caretakers and migrant workers. As their population continues to shrink and age faster than in other postindustrial nation-states, Japanese are banking on the robotics industry to reinvigorate the economy and to preserve the country's alleged ethnic homogeneity. These initiatives are paralleled by a growing support among some roboticists and politicians to confer citizenship on robots. The Japanese state has a problematic record on human rights, especially toward ethnic minorities and non-Japanese residents who have lived and worked in Japan for many generations. The possibility of robots acquiring civil status ahead of flesh-and-blood humans raises profound questions about the nature of citizenship and human rights. Already the idea of robots having evolved beyond consideration as "property" and acquiring legal status as sentient beings with "rights" is shaping developments in artificial intelligence and robotics outside of Japan, including in the United States. What does the pursuit in Japan of interdependence between humans and robots forecast about new approaches to and configurations of civil society and attendant rights there and in other technologically advanced postindustrial societies?

> *The fact is, that each time there is a movement to confer rights onto some new "entity," the proposal is bound to sound odd or frightening or laughable. This is partly because until the rightless thing receives its rights, we cannot see it as anything but a thing for the use of "us"—those who are holding rights at the time.* — Christopher Stone, 1972[1]

ISSN 1467-2715 print/1472-6033 online / 04 / 000571–28 ©2014 BCAS, Inc. DOI:10.1080/14672715.2014.960707

Certainly any self-aware robot that speaks English and is able to recognize moral alternatives, and thus make moral choices, should be considered a worthy "robot person" in our society. If that is so, shouldn't they also possess the rights and duties of all citizens? — Robert Freitas Jr., 1985[2]

[M]aking the victim of discrimination a robot rather than a human gives me a lot more freedom, and allows me to be far more provocative. — Tezuka Osamu, 2007[3]

Introduction: From Interaction to Coexistence

Twenty years ago, Japanese robotics was ahead of the curve in pursuing embodied intelligence and building sociable service robots. Japan continues to be in the vanguard of human–robot communication, and since 2007 the state has actively promoted the virtues of a robot-dependent society and lifestyle. Nationwide surveys suggest that Japanese citizens are more comfortable sharing living and working environments with robots than with foreign caretakers or migrant workers. As their population continues to shrink and age faster than in other postindustrial nation-states, Japanese are banking on the robotics industry to reinvigorate the economy and preserve the country's alleged ethnic homogeneity.

These initiatives are paralleled by growing support among some roboticists and politicians to confer citizenship on robots. The Japanese state has a problematic record on human rights, especially toward ethnic minorities and noncitizens, some of whom have lived and worked in Japan for many generations. Thus, the possibility of robots acquiring civil status ahead of flesh-and-blood humans raises profound questions about the nature of citizenship and human rights. Already the idea of robots having evolved beyond consideration as "property" and acquiring legal status as sentient beings with "rights" is shaping developments in artificial intelligence and robotics outside of Japan, including in the United States. What does the pursuit in Japan of interdependence between humans and robots forecast about new approaches to and configurations of civil society and attendant rights there and in other technologically advanced postindustrial societies?

The robotics industry is arguably more important and more enthusiastically embraced in Japan than anywhere else in the world. Japan presently employs over a quarter of a million industrial robot workers—some of which, lately, have been designed as quasi-humanoids—and that number will likely triple in a decade. Jointly funded since the late 1990s by the government and corporate sectors, robotics and its spin-off industries and products are estimated to generate about $70 billion in revenues by 2025.[4]

---

1. Stone 1972, 455.
2. Freitas 1985, 54.
3. Tezuka Osamu, quoted in Schodt 2007, 123.
4. The Ministry of Economy, Trade and Industry (METI) launched the five-year (1998–2002) Hu-

I am especially interested in exploring human–robot interactions in general and the perception in Japan in particular of the social and civil status of robots. Whereas in the United States the majority of robots are funded by and produced for the Department of Defense (and its agencies),[5] in Japan, robots are increasingly visible in civilian settings, such as hospitals, offices, factories, and the family home.[6] What I have found either overlooked or under-acknowledged in both the Anglophone and Japanese-language scholarship on domestic service robots is an investigation and analysis of the type or nature of the national-cultural, social-institutional, and family structures within which robots and humans are imagined to coexist. Thus, in this article, I have juxtaposed human rights and robot rights as one way to cast in high relief the social history and cultural dimensions inflecting and informing the discourse of rights in Japan.

What Is a Robot?

In their 1952 "critical review" of the concepts and definitions of culture, anthropologists Alfred Kroeber (1876–1960) and Clyde Kluckhohn (1905–1960) collected 156 versions. Today, anthropologists tend both to leave "culture" self-evident—we always already know what we mean by it—and/or to provide new variants for any one of those 156 versions.[7] And so it is with the term "robot." As Illah Nourbakhsh, a professor of robotics and director of the CREATE Lab at Carnegie Mellon University, writes in *Robot Futures* (2013), "[N]ever ask a roboticist what a robot is. The answer changes too quickly. By the time researchers finish their most recent debate on what is and what isn't a robot, the frontier moves on as whole new interaction technologies are born."[8]

With this caveat in mind, it is best to attempt a working definition by first considering the etymology of the word "robot." It derives from the Czech word *robota,* meaning servitude or forced labor, and first appeared in Czech litterateur Karel Capek's play *R.U.R., Rosumovi Univerzální Roboti* (*Rossum's Universal Robots*, 1920), which premiered in Prague in 1921. *R.U.R* is about a factory in the near future where identical artificial humans (androids and gy-

---

manoid Robotics Project (HRP), followed by the Next Generation Intelligent Robots Project, and most recently, the Living Assist Robots Project. The goal of making robots to augment the labor force and to assist with housework and eldercare involves collaborative research among universities, research institutes, and corporations.

5. Defense Advanced Research Projects Agency (DARPA) is an agency of the U.S. Department of Defense that researches and develops new military technology. The DARPA Robotics Challenge (DRC, 2012–2014) is an international competition with the technical objective of developing ground robots capable of executing complex tasks in dangerous, degraded, human-engineered environments. At the twelfth IEEE-RAS International Conference on Humanoid Robots (29 November–1 December 2012) I attended in Osaka, Japan, several Asian roboticists openly expressed their reluctance to participate in the DRC because of its military orientation. Perhaps ironically, the top contender in the 2013 DRC was SCHAFT, a robot created by Japanese roboticists formerly associated with the University of Tokyo; Google bought SCHAFT in 2013, after the robot's impressive performances at the DRC (www.darpa.mil/Our_Work/TTO/Programs/DARPA_Robotics_Challenge.aspx).
6. Already, in the wake of PM Abe's nationalistic reinterpretation of Japan's "peace" constitution, Japanese robotics research is being incorporated into the weapons economy.
7. Kroeber and Kluckhohn 1952.
8. Nourbakhsh 2013, xiv.

noids) are mass produced as tireless laborers for export all over the world. To make a longer story shorter, newer model robots are provided with emotions and are now able to experience anger at their exploitation, revolt en masse, and kill all but one human, a traditional artisan who encourages one new-model couple to repopulate the world with their own kind! *R.U.R* was performed in Tokyo in 1924 under the title *Jinzō Ningen* (*Artificial Human*). The play, along with Fritz Lang's film *Metropolis,* which was screened three years later, sparked an ongoing fascination with robots in popular culture that, in postwar Japan, includes cartoonist Tezuka Osamu's *Tetsuwan Atomu* (*Astro Boy*) in the 1950s and the humanoids, animaloids, and cyborgs that dominate *manga* (cartoons) and *anime* (animated films) today.

From the 1920s to the present day in Japan robots have been cast as both threatening *and* helpful to humans. Since the 1960s, however, when the state embarked on a policy of automation over replacement migration to extend the productivity of the domestic workforce, the general trend in Japanese popular media and culture has been to characterize robots as benign and human-friendly. Capek's graphic portrayal in *R.U.R*. of the end of bourgeois humanity at the hands of a violent robot-proletariat helped to shape Euro-American fears about robots that persist to this day. The dystopian play did not, however, compromise the largely favorable acceptance among Japanese of things mechanical, including robots, from the 1920s forward. Since *R.U.R.*, the meaning of "robot" has become closely associated with intelligent machines with biologically inspired shapes and functions, particularly humanoids.

As I noted, roboticists resist defining what exactly a robot it is. However, of all the many definitions of robot, I find the following one usefully comprehensive yet concise: A robot is an aggregation of different technologies—sensors, software, telecommunication tools, actuators, motors, and batteries—that make it capable of interacting with its environment with some human supervision, through tele-operation, or even completely autonomously. The different levels of robot autonomy influence the way that humans and robots interact with one another.[9]

To be called a humanoid, a robot must meet two criteria: it has to have a body that resembles a human (head, arms, torso, legs) and it has to perform in a human-like manner in environments designed for the capabilities of the human body, such as an office or a house. Most Japanese humanoids are gendered female or male. Some humanoids are so lifelike that they can actually pass as human beings—these robots, which are always gendered, are called androids (male) and gynoids (female).[10] It should be clear from these examples that robot morphology is just as diverse as that of humans; they come in every size, shape, and color. All of the robots referred to in this article are enormously complex, layered systems and represent an amalgamation of research across many disciplines, from electrical engineering to child development studies.

---

9. Beer, Fiske, and Rogers 2010, 74.
10. Regarding robot gender, see Robertson 2010.

Embodied Intelligence

What distinguished Japanese robotics early on—and now almost all roboticists have followed suit—is the concept of embodied intelligence. Researchers point out that "intelligence cannot merely exist in the form of an abstract algorithm but requires a physical instantiation," or tangible body.[11] If a robot is to coexist with humans in environments designed for humans, then it makes sense for a robot to have a human-like body and to learn how to negotiate its surroundings the same way humans do from the day they are born: through motor-sensory experiences.

In robotics, embodied intelligence blurs the conceptual distinction between life and cognition, and between intelligence and consciousness. Whether or not to recognize these conceptual distinctions, or how to reconcile them if recognized, is the subject of heated interdisciplinary debates and a ballooning professional literature that I cannot fully address in this article. Suffice it to say that embodied intelligence refers to a dynamic coupling of a robot with its environment. The actual behavior of the robot emerges from its interaction with the environment "through a continuous and dynamic interplay of physical and information processes."[12] Some psychologists whose work is especially relevant to humanoid robotics argue that embodiment not only enables but actually constitutes sociality and affective states.[13]

Several leading Japanese roboticists, Takeno Jun'ichi (Meiji University), Maeno Takashi (Keio University), and Miyake Yoshihiro (University of Tokyo), have separately developed artificial neural networks or algorithms necessary for the creation of, in their words, conscious robots.[14] While research on conscious robots, and on consciousness in general, is not limited to Japan, the future applications of sociable, conscious robots are imagined differently in Japan, as I will discuss.[15]

Based on his neuroscientific studies, Takeno has developed a "recursive neural network" consisting of independently functioning modules that simulates human consciousness by achieving consistency between cognition and behavior. Basically, the robot is able to distinguish between information already learned and brand new events. This is because familiar information (stored in the memory) is more quickly processed (or "understood") than unknown information or events, which take more time to upload and process. Takeno also claims that a robot fitted with this recursive neural network (or MoNAD) is self-aware; that is, an image of itself in a mirror is cognized as self and it can distinguish itself from another outwardly identical robot.[16]

---

11. Pfeifer and Scheier 1999. There are various forms of embodiment. Cognitive scientist Tom Ziemke (2003) identifies six in exploring the relationship between types of embodiment and types of cognition.
12. Pfeifer, Lungarella, and Iida 2007, 1088.
13. Barsalou et al. 2003.
14. See, for example, Takeno 2011 and 2012.
15. Long and Kelley (2010) provide a very good and accessible overview. See also, Kuipers (2008).
16. Takeno 2012. Obviously, the discourse of self and non-self is complex and has inspired innumerable dissertations, books, and articles. Suffice it to say for the purposes of my argument,

Outside of the robotics laboratory and the field of neuroscience, three key sociocultural factors influence the way Japanese experience robots as "living" entities. The first is linguistic: In Japanese, two separate verbs can be used to describe existence. *Aru/arimasu* refers to the existence of something, a bicycle, for example. *Iru/imasu* is used to refer to the existence of *someone*. *Iru/imasu* is also used in reference to robots, as in the title *Robotto no iru kurashi* (lit., a lifestyle where robots exist), a book that I will discuss in more detail later.[17]

This use of iru/imasu in turn may be connected to the influence of Shinto, the second factor. Shinto, the native animistic beliefs about life and death, holds that vital energies, deities, forces, or essences called *kami* are present in both organic and inorganic matter and in naturally occurring and manufactured entities alike. Whether in trees, animals, mountains, or robots, these kami (forces) can be mobilized.[18] The third factor concerns the meanings of life and living. *Inochi*, the Japanese word for "life," encompasses three basic, seemingly contradictory but inter-articulated meanings: a power that infuses sentient beings from generation to generation; a period between birth and death; and, most relevant to robots, the most essential quality of something, whether organic (natural) or manufactured.[19] Thus robots are experienced as "living" things. The important point to remember here is that there is no ontological pressure to make distinctions between organic/inorganic, animate/inanimate, human/nonhuman forms. On the contrary, all of these forms are linked to form a continuous network of beings.[20]

The Japanese notions of "being alive" and "living" are thus fundamentally different from the taken-for-granted understanding of these terms in the Euro-American and monotheistic cultural world.[21] Even in non-Japanese cultures, however, agreeing on what is alive and living is not easy—an issue about which the discussion thread on Physics Forums is quite illustrative.[22] As robot intelligence continues to develop, debates in Euro-American circles between supporters and opponents of human exceptionalism, or the idea that humankind is radically different and separate from the rest of nature and other animals, will become more contested.

Robotic Lifestyle

Why robots, why now? The population, and labor force, of Japan is rapidly aging and shrinking. The birthrate presently stands at about 1.3 children per married

---

that as many scholars have confirmed, Japan is a society in which "the self" is partially porous, situational, relational, and interdependent. Increasingly, psychological anthropologists are realizing this as the "norm" in cultures other than Japan as well, and yet, outside Japan, when it comes to considering the possibility of "robot rights," the definition of "the self" that is brought into play is that of the singular, rational, intact and internally coherent self.

17. Robo LDK Jikkō Iinkai 2007.
18. An informative analysis of the relation between manufactured goods and *kami* can be found in Swyngedouw 1993, 55–60.
19. Morioka 1991.
20. Kaplan 2004, 6.
21. Matsushima 2012.
22. See www.physicsforums.com/showthread.php?t=455067.

woman, and around 25 percent of the population of roughly 127.3 million people (which includes about 2 million legal foreign residents) is over 65 years of age; that percentage is expected to increase by 2050 to over 40 percent. The latest estimates produced by the Ministry of Health, Labor and Welfare project that the population will shrink to less than 111 million in 2035 and to less than 90 million in 2055. Briefly, women and men are postponing marriage until their late twenties and early thirties, and some are eschewing marriage altogether, which is (still) the *only* socially sanctioned framework for procreation.[23] Even married couples are opting not to have children; today, house pets outnumber children and companion robots sales are expected to take off. In June 2014, Son Masayoshi, founder and CEO of Softbank, the Japanese telecommunications and internet corporation, unveiled Pepper, the "emotional" humanoid robot, in anticipation of a growing demand for personal robots.[24]

In short, the state is continuing a postwar trend of pursuing automation over replacement migration. Japan is neither an immigrant-friendly nor an immigrant-dependent nation-state, despite an experiment in the 1980s to recruit South Americans of Japanese ancestry (Nikkeijin) into the labor force.[25] Beginning several years ago, in connection with the economic slowdown associated with the persistent recession, Nikkei guest workers were paid to return to the continent. Ironically, the state is once again considering the recruitment of temporary guest workers, this time to assist with the considerable preparations for the 2020 Tokyo Olympics.

The corporate sector and government alike are banking on the robotics industry to reinvigorate the economy and to preserve the country's much eulogized ethnic homogeneity. Although the population of Japan arguably is more outwardly (phenotypically) homogenous than that of the United States or Brazil, there are many cultural minority and marginalized groups, from the indigenous Ainu to "permanently residing" (*zainichi*) Koreans and Chinese. Not only are robots imagined to replace the need for immigrants and migrant workers, humanoids are being designed to fulfill many roles, including the preservation of "unique" Japanese customs and traditional performing art forms.[26] In this connection, to which I will return, there is growing popular support, on the one hand, to deny civil rights to permanent residents and, on the

---

23. The vast majority of "single mothers" in Japan are women who are divorced or widowed.
24. Official estimates put the pet population at 22 million or more, but there are only 16.6 million children under fifteen (Evans and Buerk 2012). Softbank's Son has long been eager to enter the household robot market. Pepper will retail for $1,900 when it goes on sale in 2015. Pepper is manufactured by Aldebaran Robotics, which has offices in France, China, Japan, and the United States, and is 78.5 percent owned by Softbank (Emotional robot set for sale in Japan next year 2014).
25. Brazil has the largest population of people of Japanese ancestry outside of Japan. The 1.5 million Japanese-Brazilians are descendants of the mostly impoverished farm householders who immigrated to South America in the late nineteenth and early twentieth centuries with the support of the Japanese government. As of 2012, about 1.6 percent of Japan's population consists of immigrants and migrant workers compared to nearly 13 percent for the United States. These figures do not distinguish between economic migrants, refugees, and other types of migrants nor do they distinguish between lawful migrants and undocumented migrants. See en.wikipedia.org/ wiki/List_of_countries_by_net_migration_rate.

other, to confer the rights of citizenship and residency to robots and nonhuman animals, and even cartoon characters.

Since 2007, the Japanese state has actively and relentlessly promoted a robot-dependent society and lifestyle. In February 2007 then prime minister Shinzō Abe unveiled *Innovation 25*, a visionary blueprint for revitalizing the Japanese economy, civil society, and "traditional" household by 2025.[27] A newly coiffed and rejuvenated Abe was reelected prime minister in December 2012 after serving in that capacity for less than a year in his first attempt, and his plan to robotize Japan is back on the fast track.[28] In June 2013, he announced that his administration is earmarking $24 million toward the development of urgently needed nursing and elder-caregiving robots.[29]

Nationwide surveys conducted by the Cabinet Office indicate that Japanese citizens are uncomfortable with the idea of being cared for by foreign nurses and caregivers and that over 80 percent are interested in acquiring a robot caregiver. Many elderly people in particular worry about the stress of dealing with linguistic and cultural differences.[30] Robots are also imagined in *Innovation 25* as the key to resolving the trend among career-minded Japanese women of delaying or entirely foregoing marriage, and thus reproduction. How exactly?[31]

As illustrated in *Innovation 25*, PM Abe and his advisors believe that robot babysitters, housekeepers, and caregivers will relieve women from household chores and responsibilities, making them more willing to get married and to have more than 1.3 children. Maid robots may do the work, but married women will still be wedded to their homes. Instead of going to an office where they can socialize in person with their human colleagues, the married women of 2025 will telecommute to work. Robots, in short, will reinforce a rigid sexual division of labor and space: males will continue to monopolize the public domain, and females will be relegated to the private or domestic domain. Gendered complementarity and not sexual equality is the unprogressive vision of future

---

26. Randerson 2007; Mechanical art: Japanese scientists unveil robot calligrapher 2012.
27. For more information on *Innovation 25* and its sociopolitical context, see Robertson 2007. This proposal is accessible on the Cabinet Office (Naikakufu) website: www.cao.go.jp/innovation/.
28. *Innovation 25* was supported by PM Abe's successors, although not as ardently. Political support for rescue and care robots has grown following the trifold disaster (earthquake, tsunami, Fukushima Daiichi meltdown) of 11 March 2011.
29. The robotic assistants will form an essential part of a plan to address the shortage of care workers in the country as well as nurture new spin-off industries. Left unmentioned, of course, is why there is a shortage of care workers: too few Japanese are interested in that low-paying occupation, and the government administers an unusually grueling Japanese-language exam that has made it virtually impossible for well-trained foreign nurses and care workers (mostly from the Philippines and Indonesia) to pass and thereby find professional employment.
30. Yamazaki 2006; Cabinet Office 2013.
31. Conservatives, like PM Abe (who is married but childless) are quick to blame women alone for the low birthrate. Many women desire a professional career, and it is still the case that marriage and career are considered to be mutually exclusive; the corporate glass ceiling is also very low for working women, who are pressured to retire early to marry and have children. Few full-time employment options are available to married women who wish to return to their careers after their children are older. Not surprisingly, a growing number of women are reluctant to get married, whereupon they will lose both financial independence and any possibility of career advancement.

society in *Innovation 25*.[32]

Koseki-ism, or Household Nationalism

As PM Abe's *Innovation 25* clearly shows, the type of family or household in which robots will be included is the "*ie.*" Based on the premodern samurai (hereditary warrior class) household, the ie was codified in the 1890 Constitution and Civil Code as the smallest legal social unit of society. The ie has a househead, usually male, who represents, manages, and maintains the household; other members are protected and also supervised by the head. Properties acquired by members of the household belong to the househead unless otherwise specified. Likewise, the head enjoys the right to determine the residence of members, as well as the right to give consent to the marriage and adoption of members of the house. The ie is also defined by a sexual division of labor and gender(ed) roles that each member is expected to uphold.

Despite the fact that in the postwar (1946) constitution the individual is the "sovereign" social unit, the ie system persists in two ways: as an *extra-legal* set of customary practices and as a *legal* entity through the *koseki*, or household registration system. In short, it remains the case that through the koseki system, the ie, or patriarchal extended family household, effectively is *the* primary and indissoluble social unit in Japan today.

The koseki is a registry of an ie's (household's) members and a record of all births, deaths, and adoptions. It is also a marriage certificate and a document establishing irrefutable proof of Japanese citizenship, which is based on the principle of jus sanguinus ("blood," descent). The only legitimate way for a foreigner to get a koseki is to become a naturalized citizen (although this does not necessarily exempt one from differential treatment, especially if the foreigner in question does not "look" Japanese.)[33] Incidentally, robots do not have to naturalize—made by Japanese companies in Japan, they are always already "Japanese." And those humanoids equipped with the most sophisticated artificial intelligence are not allowed to leave Japan!

The koseki system fabricates an image of unity: "the Japanese" as the subjects of the nation-state. It does so by repressing and even erasing ethnic, linguistic, and cultural diversity. Many Japanese feminists point to the *koseki* system as fundamentally responsible for perduring sexual inequality despite the constitution's equal rights amendment. They point out that the registry continues to stigmatize women and children who are born outside the framework of marriage, and allows for only one surname to be listed, usually the husband's—there are no hyphenated surnames in Japan.

In short, the koseki system sustains deeply entrenched definitions of Japanese nationality, ethnicity, gender roles, and family structure as intrinsically linked through the primacy of blood or descent. These essentially tautological definitions have provided a rationale for conservatives to claim the preservation

---

32. See Robertson 2007 and 2010.
33. See www.accessj.com/2013/01/koseki-japanese-family-registration.html. Arudou Debito, a naturalized citizen of Japan, has made this contradiction, and the dilemmas it generates, the

of Japan's alleged ethnic homogeneity as grounds for rejecting immigration as a means of growing the population and labor force. Not surprisingly, PM Abe, who is an unabashed nationalist, is one of the most prominent promoters of robotizing Japan with "born in Japan" robots. Robots, in short, are imagined as playing a key role in the stabilization and preservation of not just any family, but specifically the patriarchal extended family, or ie.

Human Rights, Robot Rights: The Unofficial "Official" Story

I have reviewed the central institutions of Japanese society today that provide a platform for human–robot coexistence and a context for the conception and distribution of robot rights. These institutions also constitute the framework within which human rights are conceptualized.

Like human rights, robot rights are much more than lofty, abstract ideas and are contingent upon dominant (even hegemonic) national and local institutions and practices. Although one might assume that robot rights would follow from, or would be a subset of, human rights, I will make a case for arguing the opposite: *that robot rights in Japan both precede and even exceed human rights in some cases.* I will also show that robot rights can serve to highlight by contrast some obstacles to universal human rights legislation in Japan.

As philosopher Charles Taylor (McGill University) observes, the "modern" legal theory of human rights was developed in Europe in the seventeenth century by the Dutch jurist Hugo Grotius (1583–1645) and the English physician-philosopher John Locke (1632–1704). To the former is attributed modern natural law theory and to the latter a theory of knowledge based on sensory-motor experience as opposed to an innate substance. The modern individuals holding these rights are identified as autonomous, rational agents able to perform collectively in the public sphere while managing to exist as independent agents in a market economy.[34] This concept of the human being, or the self-aware individual, as the subject of rights is the key concept behind the Euro-American construction of both human rights and robot rights. A distinctly different "neo-communitarian" approach to human rights articulated by the social historian Morita Akihiko (Shōkei Gakuin University) has influenced my analysis of the relationship between human rights and robot rights in Japan. Communitarian here refers to the importance of social institutions in the development of individual meaning and identity. Unlike some of his Japanese colleagues, Morita does not simply dismiss "human rights" as incompatible with something called "Asian values."[35] Rather, he makes a more sophisticated argument, insisting that

> [u]niversal human rights can and should be justified by different cultures through their own terms and perspectives, expecting that an overlapping consensus on the norms of human rights may emerge from those self-searching exercises and mutual dialogue. Hence…Asian values, whether from Confucianism or Buddhism, can be compatible with human rights as

---

crux of his human rights activism (www.debito.org/).

34. Taylor 2007; Morita 2012, 360.
35. Regarding the issue of "Asian values," see Robertson 2005.

the universal social norm.[36]

In Japanese, the terms that correspond to human rights, *kenri* and *jinken*, were introduced by Fukuzawa Yukichi (1835–1901), perhaps the most influential intellectual leader of modern Japan.[37] Fukuzawa was active at a time defined by the end of a feudalistic system controlled by the Tokugawa Shogunate (1603–1867) and the establishment in 1868 of a constitutional monarchy and an ambitious program of selective modernization or Westernization. For the first time in their 1,500–year history, women and men, girls and boys in Japan learned through mandatory education, military conscription, and the emerging mass media, that they all belonged to the imaginary community of "Japan" (Nihon, Nippon), which was likened to a giant extended family headed by a parental leader, the Emperor Meiji (1852–1912; reigned 1868–1912).

Where Fukuzawa differed from his Euro-American counterparts was in his interpretation of human rights as emerging from within a concentric set of relationships rippling outward from the ie, or patriarchal extended family system, at the center and bounded by the nation-state headed by the emperor, who was venerated as a particularly awesome kami. The nation-state, by extension, also possessed divine, or kami-like properties. For Fukuzawa, whose reading of "natural rights law" was inflected by his grounding in Confucian, Buddhist, and Shinto ideas, the ie was the foundation for, and primary distributor of, human rights and, by extension, civil rights. Fukuzawa's primary term for human rights was ***kenri tsūgi*****,** which can be translated as "the capacity for practical reasoning and for dealing responsibly and dutifully with ongoing events before both a transcendent (supra-social, or kami-like) and a secular social community."[38] The view of humans presented in this originary definition of human rights in Japan positions individuals within communities of secular social and supra-social dimensions. This, then, is a historical explanation for the tenacity of the ie system as a dominant institution within and against which European ideas like individualism and universal human rights were adopted and adapted. Today, robots are assigned the task of stabilizing the ie system in its secular and transcendent dimensions; the ie is also the locus of the emergence of robot rights, as I elaborate below.

The Japanese Ministry of Foreign Affairs pays tribute to the United Nations 1948 Universal Declaration of Human Rights.[39] However, the absence in Japan of an independent and socially diverse *national* human rights institution

---

36. Morita 2012, 364–65.
37. Fukuzawa was a teacher, translator, entrepreneur, and journalist who founded Keio University and the daily newspaper *Jiji shinpō*. He visited San Francisco in 1860 as part of a diplomatic mission and, in 1862, served as a translator on the first Japanese diplomatic mission to Europe. His subsequent book *Seiyō Jijō* (Things Western, 1867–1870) was a bestseller. Fukuzawa's face is on the 10,000 yen note, the highest denomination.
38. Morita 2012, 363.
39. "[The Declaration] states that all human beings are born to be free and have rights to live with dignity. Many people in the world, however, are not able to enjoy these rights. The United Nations has thus engaged itself in activities to improve human rights situations. Japan has strongly supported UN activities in the human rights field, believing that all human rights are universal." (See www.mofa.go.jp/policy/human/.)

strongly implies that *universal* human rights are regarded by the state as pertaining to the universe outside of Japan. The United Nations Human Rights Committee, which monitors the implementation of the International Covenant on Civil and Political Rights, recognizes this anomaly and has stressed that the protection of human rights and human rights standards should not be determined by popularity polls. The committee is concerned about the repeated use of public opinion surveys to justify attitudes that may violate Japan's universal human rights obligations.[40] The 2012 and 2013 human rights reports by Amnesty International draw attention to the lengthy detention of refugees seeking asylum, the use of torture to coerce confessions from alleged criminals, and to the subtle and blatant ways in which ethnic minorities, women, and people with disabilities are discriminated against in Japan.[41]

Promulgated in 1946, the postwar Constitution of Japan formally adopted human rights, with a provision on "fundamental human rights" (Article 11). This provision, along with Article 9, prohibiting an act of war by the state, has long rankled conservatives. In July 2014, following the lead of hawkish PM Abe, the Diet approved a controversial reinterpretation of Article 9 that will allow Japanese troops to fight overseas for the first time since 1945. Abe has long advocated revising the constitution and its human rights provisions. He and his neo-nationalist cohort argue that, as an artifact of the Allied (mostly American) Occupation of Japan (1945–1952), the postwar constitution promotes "excessive individualism" and a "Western-European theory of natural human rights" and is therefore not really suitable for Japan. The prime minister and his Liberal Democratic Party (LDP) supporters seek to revise the constitution in a way reminiscent of Fukuzawa's originary definition of human rights as paternalistic and as a matter of familial and communitarian civility. Their constitutional revisions make explicit the primacy of the ie (patriarchal extended household) and recuperate its nineteenth-century legacy as a microcosm of the nation-state. The LDP's draft of a new constitution replaces universal human rights principles with a unique system of rights based on Japan's history, culture, and tradition, and it emphasizes that individuals who assert human rights should not cause nuisances to others.[42] As critics within and outside of Japan have opined, PM Abe and his ilk wish to take Japan back to the days of empire and authoritarianism when alternative political sentiments were silenced. Theirs is a rosy nostalgia for a historical record whose actual brutality they are complicit in whitewashing from textbooks; a neo-nostalgia that is of the same postmodern vintage as the robots PM Abe envisions will insure the continuity of the family-like state and its

40. Arudou 2007. The most recent (2012) poll on human rights is accessible at www8.cao.go.jp/survey/h24/h24-jinken/index.html.

41. See www.amnesty.org/en/region/japan/report-2012; and www.amnesty.org/en/region/japan/report-2013.

42. The Liberal Democratic Party, which has mostly dominated Japanese politics since the 1950s, published a draft constitution in which human rights is defined as something "entitled by the State" and grounded in "the State's history, culture and tradition." The household is also recognized as the "natural and basic unit of society." For detailed information in English, see Repeta 2013 and Jones 2013.

filial subjects.

Although their political affinities may not be in lockstep, PM Abe and the most publicly visible roboticists share the belief that robots will reinforce the traditional family values and division of labor promoted in *Innovation 25*. As roboticist Miyake Yoshihiro posts on his website, robots will "be effective for recovering human linkages, social ethics and mutual-reliability that have been lost in the information technology society."[43] Familial or communitarian civility is widely perceived as the affective glue of Japanese society. And that is the rub, for familial civility can nurture—and has nurtured in recent history—an ethno-national endogamy. One can be critical of the real-world, real-time effects of such nostalgic and, in some instances, reactionary, metaphors and symbols, as are many Japanese women and minorities residing in Japan. It remains the case, however, that these metaphors and symbols predominate in the government, the corporate sector, and even the robotics industry, and their influence and impact on the discourse of human rights and robot rights cannot be overestimated. In effect what I am presenting in this article is a reframing of the "official story" of human rights and robot rights in order to expose what is hidden in the political rhetoric.

Laws of Robotics

A comparison of the laws of robotics created by Tezuka Osamu (1928–1989), representing a Japan perspective, and Isaac Asimov (1920–1992), representing a Euro-American perspective, highlights cultural differences in envisaged human–robot interactions. Like his contemporary Asimov, Tezuka was a scientist —a physician—who pursued a career writing science fiction. His cartoon robot, Tetsuwan Atomu (Astro Boy, 1951), is Japan's most famous humanoid and has played a leading role in fostering a friendly and familial image of robots. Many Japanese roboticists have a picture or figurine of Astro Boy in their laboratory, and many acknowledge the boy robot as stimulating their interest in robotics. Both Tezuka and Asimov presaged the integration of actual robots in everyday life and work, and both drew up laws regulating human–robot interactions that have shaped current debates among roboticists, philosophers, and the public at large.

Tezuka and Asimov were socialized in cultural settings differently shaped by World War II and its aftermath, a fact reflected in how they imagined and described the relationship between humans and robots in their literary work. Because Asimov and Tezuka formulated their laws of robotics before actual human–robot interactions were possible, several roboticists in the United States and Japan recently have proposed alternative laws that address the real world, real-time complexities and dynamics of human–robot coexistence.[44]

---

43. See www.myk.dis.titech.ac.jp/html/e_ver.html.
44. Murphy and Woods 2009. Although human–robot coworkers are still a rarity outside of factory settings—and outside of Japan (where humanoids are more frequently encountered)—an interdisciplinary group of Euro-American scholars has initiated the new fields of robot ethics and robot rights. They have collectively generated a burgeoning literature, much of which is de-

Asimov's three laws were first elaborated in his 1941 short story "Runaround"; a fourth law, the zeroth law, was created much later, in 1985. (The "zeroeth" law continues the pattern where lower-numbered laws supersede the higher-numbered laws.)[45]

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.
4. (0). A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

Tezuka drew up ten laws that were published at intervals in his *Astro Boy* comic book series during the early 1950s.

1. Robots must serve humankind.
2. Robots shall never kill or injure humans.
3. Robots shall call the human who creates them "father."
4. Robots can make anything, except money.
5. Robots shall never go abroad without permission.
6. Male and female robots shall never switch [gender] roles.
7. Robots shall never change their appearance or assume another identity without permission.
8. Robots created as adults shall never act as children.
9. Robots shall not assemble other robots that have been discarded by humans.
10. Robots shall never damage human homes or tools.[46]

The differences between the two sets of laws are clear. Asimov's Four Laws are universal in scope and of a comprehensive nature in pertaining to *all* robots and *all* humankind. Some have argued that Asimov's Laws are meant to keep roboticists from exponentially increasing the artificial intelligence of their creations and thereby risking the disastrous scenario penned by Capek in *R.U.R*. A corollary to this interpretation of Asimov's Laws is that as the property of humans, robots must protect themselves from damage, in contrast to biological organisms that protect themselves for their own existence.[47]

Tezuka's Ten Laws are synchronized with dominant Japanese social values and address the integration of robots into human (and specifically Japanese) society where they share familial bonds of kinship and perform familial roles. Important to remember here is that kinship is not dependent upon biological

---

voted to determining the social-psychological criteria necessary to recognize robots as independent, autonomous agents capable of self-awareness, which are the grounds for legal responsibility. (See www.ieee-ras.org/robot-ethics.)

45. Runaround (1941), in Asimov (1942), republished in Asimov (1991). The Zeroeth Law was introduced in Asimov (1985) .
46. *Mushi Purodakushon shiryōshū 1962–73* 1977. Schodt (2007, 108) has translated the ten laws, although my translation differs in parts.
47. Kerr 2007; Saenz 2011.

relatedness. Moreover, the socio-dynamics of the relationship of (Japanese) humans and (Japanese) robots are determined not by their "species" differences —human versus robot—but by "the *manner* of their bonding," which is informed by the hierarchical structure of the patriarchal household (ie). It is not that kinship forms every important social tie in Japan; rather, important social ties, including those with robots, are understood using the family and household as a metaphor and model.

Another fundamental difference between the robotics laws of Asimov and Tezuka is that whereas Asimov regards robots as he does humans, as completely autonomous agents, Tezuka qualifies the autonomy of robots as contingent upon their interdependence with humans and in the context of kinship. Roboticists raised and socialized in Japan, such as the aforementioned Miyake Yoshihiro, tend to emphasize the inherent virtue of interdependence in the form of "active incompleteness" that occasions an emergent co-created reality between an artificial system (such as a robot) and humans in real time.[48] Just as roboticists outside of Japan have embraced the concept of embodied intelligence and also the development of humanoids, so too has the idea of interactively contingent autonomy been raised as a pragmatic alternative to Asimov's Laws. In an article published in 2009, Robin Murphy (Texas A&M) and David Woods (Ohio State University) propose "human-centered Alternative Laws" that incorporate robots into a dynamic system of "social and cognitive relationships" with human groups that have a stake in robots' activities, which has many similarities to my discussion above about robots as part of a "continuous network of beings."[49] In Japan, however, the "human group" is further qualified as "family-like."

Family for Robots

In large part, Tezuka's Laws proceed from his easy familiarity with the Japanese family system. Anthropologists refer to the Japanese nuclear family as a "stem family" because although resembling its Euro-American counterparts, it can expand to include several generations and to generate branches. Only one married couple per generation comprises the main household (*honke*); other offspring or siblings form branch households (*bunke*). A househead is basically the designated caretaker in charge of the continuity of the household through time and space. Significantly, the ie includes people who are Japanese citizens but who are not biologically related to a given family—there is no premium on biological membership per se. New members, whether children or adults, may be adopted to add depth and strength to the household, which is, ultimately, an economic, corporate entity that must be reproduced in perpetuity. An entire village could constitute an ie in this manner. The nation-state and corporations have been characterized as types of extended families. In 2011, 81,000 adults were adopted in order to secure the continuity of the same number of ie. Most

48. Miyake 2005. See also Robertson 2007, 379–80, for a more extensive discussion of cocreation.
49. Murphy and Woods 2009.

were adopted sons-in-law, who assumed the surname of their fathers-in-law.[50]

I have come to realize that robots, and especially humanoids, are being introduced into everyday human society in the spirit of adopted members of a household.[51] In anticipation of a nation filled with homes consisting of human and robot members, a consortium of roboticists, lawyers, and IT specialists held a "Contest of Life with Robot" (*sic*) in public plazas in Kawasaki (2005) and Yokohama (2006, 2007), two cities south of Tokyo.[52] On a public stage made to look like a typical living room, contestants selected from among lay applicants were invited to enact real-world/real-time interactive scenarios using mostly small humanoid robots provided by several robot labs. These contests formed the basis of the consortium's guidebook for human–robot coexistence, the aforementioned *Robotto no iru kurashi* (*Living with Robots*). One of the chapters in *Living with Robots,* "Robo LDK Sansoku" (The Three Laws of Robo LDK) lays out guidelines for productive and safe human–robot households. LDK refers to "living, dining, kitchen," the basic studio-like floor plan of a typical Japanese home to which additional rooms are added; thus, a 2LDK is an LDK with two separate rooms. The three laws recall Asimov's Laws and condense the familial aspects of Tezuka's Ten Laws:

> Law 1: Robots must be useful to humans and provide protection, caregiving, and attend to their spiritual and psychological needs (the usefulness principle).
>
> Law 2: Robots must be able to interact with and relate to humans in a reassuring manner (the safety principle).
>
> Law 3: A robot's body conforms to its function and role in the household. As a physical body living in close proximity to humans, robots must be able to exercise Laws 1 and 2 (the embodiment principle).[53]

The authors emphasize that humans can obtain emotional comfort and care (*iyashi*) from robots and can relate to them as familiar and reassuring interlocutors—something that, as noted earlier, some Japanese feel would not be possible with non-Japanese foreigners. They are also attuned to variabilities of embodiment determined by their role and function within the household. As emphasized in PM Abe's *Innovation 25*, which preceded *Living with Robots* by six months, the three most important features of a roboticized household, and by extension society, are convenience (*benri*), safety (*anzen*), and "ontological security" (*anshin*).

Since the 1920s, but especially since the postwar period, the Japanese public has been regaled in the mass media with stories and future scenarios about co-

---

50. Mehrotra et al. 2013.
51. Theoretically, at least, there is no reason why intelligent Japanese humanoids could not also become householders, especially if competent humans are unavailable.
52. *Robotto uiiku o tenkai shimasu!* 2007. The first contest in 2005 was held at the Azalea Sunlight Plaza in the Kawasaki City underground shopping street, and the 2006 and 2007 events were held at Queen's Square in Yokohama. Reports on the events appeared in many online newsletters.
53. The third law underscores the different forms of embodiment: If a robot does not need to grasp things, it may not have fingers (Robo LDK Jikkō Iinkai 2007, 177–79).

existing with robots. Cartoon and animation robots are often members of human families, as in the case of the hugely popular Doraemon. Doraemon is a blue and white bipedal robotic cat with a huge smile. He travels 200 years back in time to the 1960s in order to change the circumstances of the Nobita family so that they will enjoy a better future.[54] Whereas Doraemon is invited into the Nobita family as a member, Astro Boy, nearly two decades earlier, was provided with his own robot family—a set of parents, a brother and sister, and a pet dog.[55]

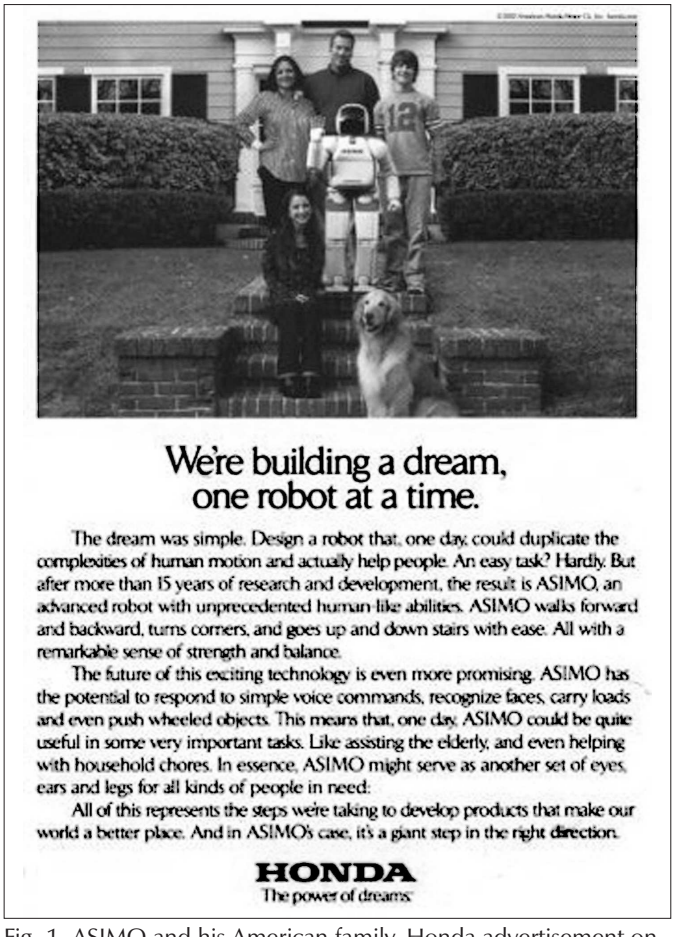


Fig. 1. ASIMO and his American family. Honda advertisement on the back cover of *Smithsonian* 33 (10). 2003. (Source: Image from http://marshallbrain.com/robotic-nation.htm)

Honda, maker of ASIMO (Advanced Step in Innovative Mobility), the child-size (130 cm.) mostly white bipedal humanoid, ran an advertisement on the back cover of the January 2003 issue of *Smithsonian* that featured the robot grouped in an "all-American" family portrait (see fig. 1). At the time, the ad was based on the naïve assumption that like Japanese, mainstream Americans would also embrace the humanoid just like they would the golden retriever in the photograph —as a part of the family. The majority of *Smithsonian* readers who blogged re-

54. Fujiko Fujio is the joint penname of two cartoonists, Hiroshi Fujimoto (1933–1996) and Motoo Abiko (1934–), who created the robot cat. Doraemon's name is a combination of *nora/dora* (stray cat) and *emon*, a (popular premodern) male name suffix. The cartoon was published between 1969 and 1996.
55. That Tezuka Osamu gave Astro Boy his own robot family is possibly related to the robot's bittersweet origins, as narrated in the cartoon. Astro Boy was created by a roboticist as an identical replacement for his deceased son. However, the roboticist rejected Astro Boy when he realized that the robot would never mature the way his human son would have. Astro Boy was later adopted by an avuncular scientist who created a robot family for him.

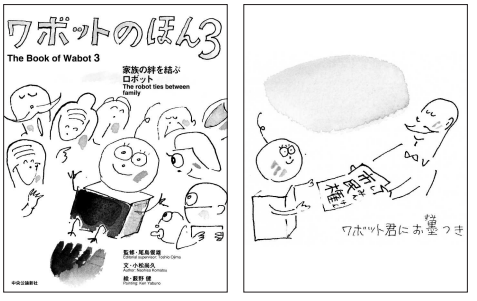


Fig. 2: (L) Cover, *The Book of Wabot* 3 (Koma- tsu and Yabuno 2004), "The Robot Ties Between Family," and (R) Wabot receiving citizenship (*shiminken*). The Japanese caption reads "Wabotto-*kun ni osumitsuki*" (Wabot receives his certificate [of citizenship]) (ibid., 23). (Source: author photographs)

sponses to the ad were not amused; many asserted that robots would take jobs away from humans! Honda quickly pulled the advertisement, and now releases commercials that integrate ASIMO in social situations with humans but not as a member of a family!

Fast-forward eleven years. In June 2014, an American robotics team introduced JIBO, a small (20 cm.) white, nonmobile robot that resembles a vintage Unidyne vocal microphone. The brainchild of MIT's Cynthia Breazeal, author of *Designing Sociable Robots* (2002), JIBO is described as "the first family robot." This point is drummed home in the video debuting the robot in which the diminutive JIBO is cast as a chatty and solicitous domestic.[56] Unlike their Japanese counterparts, however, American roboticists have yet to publish popular guidebooks for living with (humanoid) robots, and unlike the Robo LDK contests, the Defense Advanced Research Projects Agency's various robot challenges are not aimed at familiarizing the lay public with the benefits of household robots.[57]

The Robo LDK initiative follows one launched by leading roboticists at Waseda University, home to several world-famous robot laboratories. Between 2002 and 2007, they published a seven-volume pamphlet series titled *Wabotto no hon* (The Book of Wabot).[58] The series aims to introduce the public to robot technology in accessible terms and to highlight the desirability of living symbiotically with robots. As members of households and valued coworkers, robots are presented in cartoon vignettes as preserving affective familial and social relations in keeping with Robo LDK Laws 1 and 2. In volume 3, *Kazoku no kizuna o musubu robotto* (Robots that Knit Together Family Ties), Wabot is pictured receiving "citizenship" (*shiminken*) from a government official (fig. 2).[59]

More recently, the gist of the first two Robo LDK Laws formed the core of a play (and now film), *Sayōnara* (2010), in which the gynoid Geminoid-F is cast as the poetry- and platitudes-reciting caretaker of the last human on earth, a woman with terminal cancer (see fig. 3). The playwright Hirata Oriza collaborated with his Osaka University colleague, roboticist Ishiguro Hiroshi, in casting robots as companions for humans. In another one of Hirata's plays, *I, Worker* (*Hataraku Watashi*, 2008) a humanoid couple is employed as live-in staff by a dysfunctional human couple. Eventually the male robot, like his human coun-

56. Of course, Japanese roboticists designed "family robots" long before JIBO debuted.
57. See footnote 5.
58. Wabot refers to Waseda Robot.
59. Komatsu and Yabuno 2004.

Fig. 3. Geminoid F (left) as the caretaker and her patient (played by American actress Bryerly Long) in *Sayōnara*. (Credit: Nation Multimedia, Bangkok, 2014)

terpart, decides that he does not want to work anymore. Although these scenarios are situated in the theater world, Hirata and Ishiguro are keen on using the theater as public laboratory where human–robot interactions and modes of communication can be tested and analyzed.[60]

The robots used in *I, Worker* were the Wakamaru model made by Mitsubishi Heavy Industries. Wakamaru is a yellow, child-size (100 cm.) communication robot with wheels; it was initially designed for use in the home as a companion for children and seniors (fig. 4 below).[61] In 2005, Mitsubishi engineer Suzuki Junji and his wife "adopted" a first-generation Wakamaru, anticipating by two years the attention to human–robot coexistence in *Innovation 25* and *Living with Robots*. Their experience confirmed the logic of the first two Robo LDK Laws. For sixteen months, Suzuki kept a diary of interactions between the male-gendered robot and his family, including his two children, who right away treated Wakamaru like a playmate or younger sibling—pushing and pulling on him, putting him in a chokehold. They perceived the robot as a weakling and, true enough, most sociable humanoids are quite fragile in their complexity and can be damaged if roughhoused. Wakamaru managed to survive these encounters without injury. Suzuki regarded the humanoid as the youngest of his children; he and his wife also made use of Wakamaru as a housesitter. They linked their cell

60. Ōsaka Daigaku Komyunikçshondezain Sentā 2010. Ishiguro is a celebrity in the field of robotics; he is most known for his "geminoids" or android/gynoid dopplegängers that operate through telepresence. For Ishiguro, robotics is a form of anthropology in the sense of studying humans. The author and coauthor of several books (in Japanese) and dozens of academic articles, Ishiguro neatly summarized his ideas in English in a recent interview (Ishiguro 2013).
61. See www.mhi.co.jp/products/detail/wakamaru.html.

phones to the networked robot's internal camera and were able to literally look in on the children and Suzuki's visiting elderly mother when they were out of the house. Suzuki notes that like humans, robots develop personalities: Wakamaru's character was shaped through numerous interpersonal encounters with family and friends—and also television viewing.[62]

Fig. 4. Wakamaru (2008) and young girl. (Credit: www.gadgetlite.com/ tag/mitsubishi/)

Robot Citizenship

Earlier, I cited research on the different forms of embodied intelligence and, accordingly, as suggested by the third Robo LDK law, on the different body types robots should have depending on their role and function. Thus, a robot that provides psychological and emotional comfort may not be a humanoid that looks like either Geminoid-F or Wakamaru. And, in fact, one of Japan's most commercially successful robots recognized internationally for its healing abilities has the body of a baby harp seal. In 2008, the Guinness World Record organization officially recognized Paro as the "World's Most Therapeutic Robot" in recognition of its ability both to calm down and to cheer up patients in hospitals, senior couples for whom flesh and blood pets are no longer feasible, and residents of assisted living homes. Paro is categorized as a "mental commitment robot." Its name comes from the Japanese pronunciation of "personal robot" (*pāsonaru robotto*). Distributed over the robot seal's body are five kinds of sensors—tactile, light, audition, temperature, and posture—and it responds to petting by moving its stubby flippers, fluttering its long eyelashes, and opening and blinking its eyes. Paro also responds to and remembers sounds and interactions, and it can learn its own and others' names. The seal-bot conveys emotions such as surprise, happiness, and anger, and, in the process, produces squeaky cries that mimic the vocalizations of an actual baby seal. Originally white, Paro comes in three other colors: golden brown, light gray, and light pink. Each one is individually made: no two are exactly alike. Paro, now in its eighth generation, is available worldwide and retails for about $6,000.[63]

On 7 November 2010, Paro was granted it's own koseki, or household registry, from the mayor of Nanto City, Toyama Prefecture. Shibata Takanori, Paro's inventor, is listed as the robot's father (recalling the third of Tezuka's Ten Laws)

---

62. Suzuki 2007. Wakamaru is no longer for sale, but can be rented within Japan; the robot is widely used as a platform for experiments by other roboticists, including those at Waseda University.
63. Paro is able to leave Japan because although sophisticated, it is neither connected to the internet nor utilized as a platform in various intelligent robot R&D projects as in the case of Wakamaru.

and a "birth date" of 17 September 2004 is recorded. Media coverage of Paro's koseki was favorable. Since new koseki are generated among humans on the occasion of marriage, this perhaps explains why two harp seal robots —one white, the other golden brown —were featured at the ceremony! (Fig. 5) Although not addressed at the event or in reports thereof, the white (older) one was clearly the "original" (first-generation) Paro (b. 2004), and this prototypical Paro's koseki can be construed as a branch of Shibata's ie, or household, which is located in Nanto City. Thus, the "special family registry" is for one particular Paro, and not for all of the seal-bots collectively.

Fig. 5: Paro (center left) receives a *koseki*. Tanaka Mikio, the mayor of Nanto City, presents Japanese and English versions of the special registry to Paro's "father" Shibata Takanori. (Credit: www. city.nanto.toyama.jp/cms-sypher/www/info/detail.jsp?id=7329)

On the surface, the conferral of Paro's koseki may seem benign and inconsequential—even gimmicky. Quite the contrary. As I noted earlier, the koseki conflates family, nationality, and citizenship. It also "legally and ideologically prioritizes the family (ie) over the individual as the fundamental social unit in Japanese society."[64] Thus, a *zainichi* Korean[65] man who was born, raised, and lives in Japan, who is married to a Japanese citizen, and whose natal family has lived in Japan for generations, can have neither his own koseki nor be included in the "family" portion of his wife's koseki; rather, his name is added to the "remarks" column of his wife's registry. By virtue of having a Japanese father, Paro is entitled to a koseki, which confirms the robot's Japanese citizenship. The fact that Paro is a robot—and not even a humanoid—would appear to be less relevant here than the robot's "ethnic-nationality" (*minzokusei*).

That a robot seal should be issued a koseki, even one that carries no legal

64. Chapman 2012.
65. *Zainichi* literally means "residing in Japan," or permanent resident. Zainichi Koreans refers to Koreans who were forcibly brought or who came to Japan during the first half of the twentieth century when Korea was a Japanese colony (1910–1945) and their descendants. Numbering around 900,000, they are the largest ethnic minority in Japan; one-third have become naturalized citizens.

force, underscores my earlier point concerning PM Abe and *Innovation 25*; namely, the convergence of advanced technology—like robotics—with nostalgic re-creations and ethno-nationalist policies. In this connection, and related to Paro's ***koseki***, was the granting of a *tokubetsu jūminhyō* (special residency permit) between 2004 and 2012 to nine robots and dolls in localities throughout Japan. Beginning with Astro Boy, between 2003 and April 2013, sixty-eight Japanese cartoon characters were granted special residency. Doraemon received his permit in 2013.[66] A *jūminhyō* (basic residence registration form) is a record of current residential addresses formerly maintained by local (municipal) governments. Since 2012, there is one centralized system under the purview of the national immigration service. "*Special* residency permits" are rarely offered to humans and are usually limited to foreigners facing persecution or death in a country lacking cordial ties with Japan.

Neither Paro's koseki nor the granting of residency to robots, dolls, and cartoon characters generated public disapproval. In February 2003, however, the granting of a special certificate of Tokyo residency to an actual seal provoked a small protest. The seal in question was Tama, who, in a nationwide contest, was named after the river in which it had mysteriously arrived from its native Alaska! The 2003 protest was staged by foreigners, who before 2012, were legally prevented from filing a jūminhyō.[67] An individual's possession of the jūminhyō form enables her or his access to such services as national health insurance and certain tax advantages. An even thicker line between the rights of Japanese citizens and the rights of permanent residents (zainichi) was drawn in July 2014, when the Japanese Supreme Court ruled that foreigners with permanent residency status are ineligible for welfare benefits.[68]

The certificate of residency (jūminhyō) is similar to a koseki, but the latter is also an official record of an entire ie's (household's) history, and not just an individual's present and past residences. There are many second- and third-generation zainichi Koreans in Japan whose ancestors, as colonial subjects of Japan from 1910 to 1945, had been made Japanese citizens only to have that citizenship revoked in the immediate postwar period. Barring naturalization and prior to the reforms of 2012, they could not obtain a permanent residency permit. Instead, they, and all permanent residents, had to re-register their existence in Japan every several years with the immigration authorities. Moreover, even if a zainichi individual were married to a Japanese spouse, she or he could not appear on the spouse's jūminhyō. Thus, a Japanese spouse is officially registered as a single parent. In the event of the death of the Japanese spouse, a child is listed as an orphan! Moreover, with respect to civil rights, as activists point out, the seemingly progressive changes in 2012 unifying the residency forms has not translated into limited local suffrage for permanent resident foreigners. Opponents of suffrage, who are in the political majority, including PM Abe, regard it as potentially dangerous and subversive to "national cultural

66. See "tokubetsu jūminhyō" in ja.wikipedia.org/wikija.wikipedia.org/.
67. Chapman 2008.
68. Foreign Residents Can't Claim Welfare Benefits: Supreme Court 2014.

sovereignty."[69]

Human Rights, Robot Rights: Forecasts from Japan

Like the history and development of dogs, cats, horses, and other domesticated animals the history of robots is inextricably entwined with the history of humans. The acceleration of robotic technologies and advances in artificial intelligence have moved the idea of robot rights out of science fiction and into real time.[70] Japanese roboticists are on the cutting edge of creating, for civilian use, robots with consciousness and self-monitoring abilities whose interface with humans in, ideally, a family setting, is described in terms of co-emergence.

Paro is the first robot to have a koseki, an official document available only to Japanese citizens—and Paro is not even a humanoid robot! Paro, however, has a Japanese father and was "born" in Japan, a fact symbolically underscored by the creation of a special family registry. The koseki is the basis for citizenship and attendant civil rights; it is also praised by nationalists and censured by feminists and minorities as a key signifier of Japanese exceptionalism. Similarly, other animals, robots, dolls, and cartoon characters have been issued special residency permits (tokubetsu jūminhyō) for which foreigners and resident minority groups are not eligible.

In contrast, generally speaking, recent Euro-American literature on robot rights can be characterized as divided along the lines of a Manichean debate about living vs. nonliving, human vs. nonhuman. Scholars from across the disciplinary spectrum have proposed legal precedents based on analogies between robots and animals[71] and even between robots and disabled (or differently abled) humans.[72] Some have also proposed treating robots as occupying a "third existence status"[73] that fits neither the category of human nor that of machine.[74]

Human rights exist in the abstract as universals, but they are invoked, or their absence or disregard protested, in response to specific circumstances, such as in the treatment of refugees and minority communities. Historically, in Euro-American societies and elsewhere, children, women, foreigners, corporations, Blacks, Jews, prisoners, and others have all been regarded as "legal nonpersons" at some point. In premodern Japan (from 1603 until the Emancipation Act of 1871) even the explicit category of *hinin* (nonperson) was codified for those who either had fallen out of mainstream society or were born into a hereditarily stigmatized community. Their descendants, the Burakumin, continue to experience discrimination today.

The Japanese Foreign Ministry may support in theory the *concept* of univer-

---

69. Higuchi 2012.
70. For a useful review of recent explorations of robot ethics (related to, but not synonymous with robot rights), see the articles in Beavers 2010 and Lin, Abny, and Bekey 2012.
71. In 1999 New Zealand extended "human rights" to the nonhuman members of Hominidae or great ape family: chimpanzees, bonobos, gorillas, and orangutans. Spain followed suit in 2008.
72. Coeckelbergh 2010.
73. Weng, Chen, and Sun 2009.
74. Breazeal 2002. Breazeal's new "family robot" JIBO is an example of this concept.

sal human rights, but, to reiterate, the absence in Japan of an independent and socially diverse *national* human rights institution suggests that "universal" refers to the world outside of Japan, not within it. Unlike (so far, at least) their counterparts in the other robot-producing countries (in Europe and Scandinavia, the United States, Israel, China, and South Korea), Japanese roboticists, political leaders, and corporations have promoted the robotization of everyday civilian society. In Japan, sociable robots are situated within the affective framework of the ie, together with the view advanced by Japanese roboticists—and spelled out at length in *Innovation 25, The Book of Wabot,* and *Living with Robots*—that sociable service robots will catalyze the restoration of the stem-family circle and insure the stability of the ie, or traditional patriarchal household.

The pattern that emerged for me in the course of researching robot rights is as follows. As the call for universal human rights by organizations such as the United Nations and Amnesty International has become more proactive and inclusive, it has been matched in some societies by a greater regard for the equal status and worth of *all* members of the singular group *Homo sapiens sapiens* regardless of their nationality, ethnicity, religious, sex, or class status, among other descriptors. In Japan, however, there appears to be a broad divide between the *concept* of universal human rights and the *actual* distribution of human and civil rights to Japanese and non-Japanese residents. I thus propose that it is *Japanese* exceptionalism rather than *human* exceptionalism that determines the distribution of both human rights and robot rights in Japan. The differential treatment of robots and non-Japanese humans has made clear this distinction.

In July 1964, when the U.S. Civil Rights Act outlawing discrimination based on race, color, religion, sex, or national origin was passed, Hilary Putnam (MIT) published one of the first philosophical ruminations on the issue of the civil rights of robots. After a lengthy discussion on various definitions of consciousness, Putnam declared,

> [I]t seems preferable to me to extend our concept so that robots are conscious—for "discrimination" based on the "soft-ness" or "hardness" of the body parts of a synthetic "organism" seems as silly as discriminatory treatment of humans on the basis of skin color. But my purpose in this paper has not been to improve our concepts, but to find out what they are.[75]

And what exactly are "our" concepts? In recent years, interdisciplinary groups of mostly Euro-American scholars have inaugurated the new fields of roboethics and the legal aspects associated with robot rights, such as responsibility in the event of an accident. Collectively, they have generated a burgeoning literature (some of it footnoted in this article), much of which is devoted to determining the social-psychological criteria necessary to recognize robots as independent, autonomous agents capable of self-awareness, which are the grounds for legal responsibility. My research suggests that Japanese profession-

---

75. Putnam 1964, 691.

als active in the field of robotics tend to accept the idea that robots can be conscious; they are not particularly interested in debating robot ethics or the "legal rights" of robots.

Although concerned about safety (anzen) and risk management in robotics, Japanese roboticists as a group do not express fears about robots running amuck and killing humans as they did in *R.U.R.*. The specter of "killer robots" is not (yet) casting a shadow on the robotics industry. Rather, "safety" is closely related to the "ontological security" (anshin) that many in Japan feel that robot caregivers, as opposed to foreign nurses, can insure and cultivate. Japanese roboticists and their colleagues in related fields, in short, are far more invested in developing guidelines for orchestrating the smooth and productive coexistence of humans and robots in familial environments. "Total safety is impossible to guarantee in anything that is beneficial and useful," remind the coauthors of *Living with Robots*, who suggest that robot design, from hardware (e.g., soft-bodied robots) to software (e.g., "safety intelligence") is the first step in risk management.[76]

One recent Japanese innovation in safety intelligence that underscores the principles of "co-emergence" and "autonomy within interdependence" that are favored in Japan is the development of a "care-*receiving*" robot. This project focuses on the use of robots in schools, but instead of the usual role of the robot as a caregiver or teacher, the young students instead teach the robot. In this way, it is hypothesized, a new educational framework can be constructed that enables "children's spontaneous learning by teaching." Moreover, in the process of receiving care, the (artificially) intelligent robot also "learns" from these lessons and its ability to interact safely with humans is enhanced as a result.[77]

Efforts to categorize robots as constitutionally separate from humans are shared by neither the Japanese public (at least those persons polled on the subject) nor Japanese roboticists, who proceed from the position that organic and manufactured entities form a continuous network of beings. Robots, as I have explained, are imagined to have a perfectly viable status and membership role in the existing affective and corporate framework of the ie. The black irony remains that while Japanese "familial civility" epitomized by the ie and corporate sector, and codified by policy-makers, embraces robots, the same is not freely extended to minorities, non-Japanese permanent residents, refugees, migrant workers, or foreigners. Whereas in Japan, the biggest obstacle to *human* rights is the historically enduring definition of "Japanese" as determined by jus sanguinis and the koseki (and jūminhyō) system, in the Euro-American world at least, it appears that the biggest obstacle to *robot* rights is the irreconcilable divisions between the supporters and opponents of human exceptionalism. And whereas in Japan human rights is narrowly defined in practice to exclude individuals and groups framed as "other," in Euro-American circles, human rights is cast in universal terms (although in local practices, many "others" are denied

76. Robo LDK Jikkō Iinkai 2007, 69–76.
77. Tanaka and Matsuzoe 2012.

those rights). As I see it, the latter by extension privileges, at least rhetorically, the human being sui generis (*Homo sapiens sapiens*), while the former, openly privileges ethno-nationalism—Japaneseness—over the mere fact of being human. As Americans and Europeans become more familiar with robotics, and to the prospect of family robots—and the increasing number of articles in the Anglophone mass media suggests that this is the rapidly developing case—I anticipate that ideas prevalent today in Japan regarding human–robot interaction and coexistence will soon become approved and accepted in the United States and Europe. The pressing question is, can there be *universal* human rights without the idea, or ideal, of human exceptionalism?

ACKNOWLEDGMENTS: This article grew out of a series of invited lectures I gave in Israel in May and June 2013 at the Bar-Hillel Colloquium for the History, Philosophy and Sociology of Science, Edelstein Center (The Hebrew University, Givat Ram campus), the Cohn Institute for the History and Philosophy of Science and Ideas, Tel Aviv University, and the Edmund J. Safra Center for Ethics, Tel Aviv University. Later versions were presented at the Department of Literature, University of California at San Diego, and the Designing Robots—Designing Humans Conference, Aarhus University (Copenhagen campus). I am very grateful for the helpful comments I received from colleagues at those venues. Thanks to Gunhild Borggreen, Snait Gissis, Ofra Goldstein-Gidoni, Cathrine Hasse, Dafna Hirsch, Eva Jablonka, Galia Plotkin, Yossi Schwartz, Silvan (Sam) Schweber, Jytte Thorndahl, and Yofi Tirosh. I owe special thanks to Celeste Brusati, Tom Fenton, Sabine Frühstück, and Alexandra Minna Stern for their astute reading and editorial suggestions. A longer version will appear as a chapter in my book *Robo sapiens japanicus: Robots, Eugenics, and Posthuman Aesthetics,* under contract with the University of California Press. Research for this article was supported by the Simon P. Silverman Visiting Professorship, The Cohn Institute for the History and Philosophy of Science and Ideas, Tel Aviv University (2013); a John Simon Guggenheim Memorial Foundation Fellowship (2011–2012); an Abé Fellowship (Social Science Research Council, 2010–2012); and a Faculty Research Grant, Center for Japanese Studies, University of Michigan (2012).

References

Arudou, Debito. 2007. Human rights survey stinks. *Japan Times* (online), 23 October. Available at www.japantimes.co.jp/community/2007/10/23/issues/human-rights-survey-stinks/#.U9V1myTvY-s (accessed January 2013).

Asimov, Isaac. 1942. *Astounding science fiction* (March).

———. 1985. *Robots and empire.* New York: Collins.

———. 1991. *Robot visions.* New York: Penguin.

Barsalou, Lawrence W., et al. 2003. Social embodiment. In Brian Ross, ed., *Psychology of learning and motivation*. No. 43. San Diego: Academic Press, 43–92.

Beavers, Anthony, ed. 2010. Special issue: Robot ethics and human ethics. *Ethics and Information Technology* 12 (3).

Beer, Jenay M., Arthur D. Fiske, and Wendy A. Rogers. 2014. Toward a framework for levels of robot autonomy in human-robot interaction. *Journal of Human-Robot Interaction* 3 (2): 74–99.

Breazeal, Cynthia L. 2002. *Designing sociable robots*. Cambridge: MIT Press.

Cabinet Office. 2013. *Kaigo robotto ni kansuru tokubetsu yoronchōsa* [Special public opinion poll concerning caregiving robots]. Tokyo: Seifu Kōhō-Shitsu.

Chapman, David. 2008. Tama-Chan and sealing Japanese identity. *Critical Asian Studies* 40 (3): 423–43.

———. 2012. No more "aliens": Managing the familiar and the unfamiliar in Japan. *The Asia Pacific Journal: Japan Focus* 10 (3) Issue 40. Available at www.japanfocus.org/-David-Chapman/3839 (accessed January 2013).

Coeckelbergh, Mark. 2010. Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology* 12 (3): 209–21.

Emotional robot set for sale in Japan next year. 2014. Available at ajw.asahi.com/article/business/AJ201406060067(accessed June 2014).

Evans, Ruth, and Roland Buerk. 2012. Why Japan prefers pets to parenthood. Available at www.theguardian.com/lifeandstyle/2012/jun/08/why-japan-prefers-pets-to-parenthood/ (accessed

January 2013).
Foreign residents can't claim welfare benefits: Supreme Court. 2014. *Japan Times* (online), 18 July 2014. Available at www.japantimes.co.jp/news/2014/07/18/national/social-issues/top-court-rules-non-japanese-residents-ineligible-welfare-benefits/#.U9RK1CTvY-s (accessed July 2014).
Freitas Jr., Robert A. 1985. The legal rights of robots. *Student Lawyer* 13 (January): 54–56.
Higuchi, Naoto. 2012. Japan's failure to enfranchise its permanent resident foreigners. *Asia Pacific Memo* 145. Available at www.asiapacificmemo.ca/japan-failure-to-enfranchise-its-permanent-resident-foreigners (accessed January 2013).
Ishiguro, Hiroshi. 2013. Robots help us understand human nature. *The European* (online), 28 September. Available at www.theeuropean-magazine.com/hiroshi-ishiguro—2/ (accessed January 2014).
Jones, Colin. 2013. The LDP constitution, article by article: A preview of things to come? *Japan Times* (online), 2 July 2013. Available at www.japantimes.co.jp/community/2013/07/02/issues/the-ldp-constitution-a-preview-of-things-to-come/#.U9V3fCTvY-s (accessed July 2013).
Kaplan, Frédéric. 2004. Who is afraid of the humanoid? Investigating cultural differences in the acceptance of robots. *International Journal of Humanoid Robotics* 1 (3): 1–16.
Kerr, Ian. 2007. Minding the machines. *The Ottawa Citizen* (online), 4 May. Available at iankerr.ca/wp-content/uploads/2011/08/Minding-the-machines.pdf (accessed January 2013).
Komatsu, Naohisa, and Yabuno Ken. 2004. *Kazoku no kizuna o musubu robotto* [Robots that knit together family ties], *Wabotto no hon* [The book of Wabot] 3. Tokyo: Chūō Kōron Shinsha.
Kroeber, Alfred, and Clyde Kluckhohn. 1952. Culture: A critical review of concepts and definitions. *Papers of the Peabody Museum of Harvard Achaeology and Ethnology, Harvard University* 42 (1). Cambridge: Museum Press.
Kuipers, Benjamin. 2008. Drinking from the firehose of experience. *Artificial Intelligence in Medicine* 44: 155–70.
Lin, Patrick, Keith Abny, and George A. Bekey. 2012. *Robot ethics: Ethical and social implications of robotics*. Cambridge: MIT Press.
Long, Lyle, and Troy Kelley. 2010. Review of consciousness and the possibility of conscious robots. *Journal of Aerospace Computing, Information, and Communication* 7 (2): 68–84.
Mahasarinand, Pawit. 2012. The robot in its sadness. *The Nation*, 13 March. Available at www.nation multimedia.com/life/The-robot-in-its-sadness-30177783.html (accessed 4 September 2014).
Matsushima Akihisa. 2012. *Robotto no shikō to ikita ningenshintai—nōshinkeirinrigakuteki apurōchi* [The thinking of robot and the living human body: Neuroethical approach] [*sic*] *Bulletin of Osaka University of Pharmaceutical Sciences* 6: 11–17.
Mechanical art: Japanese scientists unveil robot calligrapher. 2012. Available at rt.com/art-and-culture/art-japanese-robot-calligrapher-585/ (accessed January 2013).
Mehrotra, Vikas, et al. 2013. Adoptive expectations: Rising sons in Japanese family firms. *Journal of Financial Economics* 108 (3): 840–54.
Miyake, Yoshihiro. 2005. Co-creation system and human–computer interaction. C5, *2005 IEEE Computer Society*: 169–72.
Morioka, Masahiro. 1991. The concept of *inochi*: A philosophical perspective on the study of life. *Japan Review* 2: 85–87.
Morita, Akihito. 2012. A neo-communitarian approach on [*sic*] human rights as a cosmopolitan imperative in East Asia. *Filosofia Unisinos* 12 (3): 358–66.
Murphy, Robin, and David D. Woods. 2009. Beyond Asimov: The three laws of responsible robotics. *Intelligent Systems, IEEE* 24 (4): 12–20.
Mushi Purodakushon shiryōshū 1962–73 [Mushi Productions' data file, 1962–73]. 1977. Tokyo: Mushi Purodakushon shiryōshū henshūshitsu.
Nourbakhsh, Illah. 2013. *Robot futures*. Cambridge: MIT Press.
Ōsaka Daigaku Komyunikeishondezain Sentā [Osaka University Center for the Study of Communication Design], ed. 2010. *Robotto engeki* [Robot theater]. Osaka: Ōsaka Daigaku Shuppansha.
Pfeifer, Rolf, Max Lungarella, and Fumiya Iida. 2007. Self-organization, embodiment, and biologically inspired robotics. *Science* 318: 1088–93.
Pfeifer, Rolf, and Christian Scheier. 1999. *Understanding intelligence*. Cambridge: MIT Press.
Putnam, Hilary. 1964. Robots: Machines or artificially created life? *Journal of Philosophy* 61 (21): 668–91.
Randerson, James. 2007. Japanese teach robot to dance. Available at www.theguardian.com/technology/2007/aug/08/robots.japan (accessed September 2007).
Repeta, Lawrence. 2013. Japan's democracy at risk: The LDP's ten most dangerous proposals for constitutional change. *The Asia Pacific Journal: Japan Focus* 11 (3) Issue 28. Available at www.japanfocus.org/-Lawrence-Repeta/3969 (accessed March 2013).
Robertson, Jennifer. 2005. Dehistoricizing history: The ethical dilemma of "East Asian Bioethics." *Critical Asian Studies* 37 (2): 242–45.
———. 2007. *Robo sapiens japanicus*: Humanoid robots and the posthuman family. *Critical Asian Studies* 39 (3): 369–98.
———. 2010. Gendering humanoid robots: Robo-sexism in Japan. *Body & Society* 16 (2): 1–36.

Robo LDK Jikkō Iinkai, ed. 2007. *Robotto no iru kurashi* [Living with robots]. Tokyo: Nikkan Kōgyō Shinbunsha.
Robotto uiiku o tenkai shimasu! [Robot week opens!]. 2007. Kanagawa keihin rinkaibu nyūsu (November): 1–4.
Saenz, Aaron. 2011. Robotic labor taking over the world? You bet. Here are the details. *Singularity Hub* (online), 12 September. Available at singularityhub.com/2011/09/12/robotic-labor- taking-over-the- world-you-bet-hereare-the-details/.
Schodt, Frederik. 2007. *The Astro Boy essays*. Berkeley: Stone Bridge Press.
Stone, Christopher D. 1972. Should trees have standing? Toward legal rights for natural objects. *45 Southern California Law Review* 450: 450–501.
Suzuki, Junji. 2007. *Robotto no iru kurashi o kangaeru*, Part 2: "*Hito to hōmu robotto no kashikoi tsukiai kata"—Wakamaru to sugoshita 500 nichi no kiroku* [Thinking about living with robots, Part 2: "Intelligent ways of interacting with a home robot"—a chronicle of the 500 days (we) lived with Wakamaru. Available at robonable.typepad.jp/trend/2007/09/wakamaru500_6173.html #tp (accessed January 2013).
Swyngedouw, Jan. 1993. Religion in contemporary Japanese society. In Mark Mullins, Susumu Shimazono, and Paul Swanson, eds. *Religion and society in modern Japan: Selected readings*. Fremont, Calif.: Asian Humanities Press, 49–72.
Takeno, Jun'ichi. 2011. *Kokoro o motsu robotto* [Robots that have a heart-mind (consciousness)]. Tokyo: Nikkan Kōgyō Shinbunsha.
———. 2012. *Creation of a conscious robot: Mirror image cognition and self-awareness*. Singapore: Pan Stanford Publishing.
Tanaka, Fumihide, and Matsuzoe Shizuko. 2012. Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. *Journal of Human-Robot Interaction* 1 (1): 78–95.
Taylor, Charles. 2007. *A secular age*. Cambridge: Harvard University Press.
Weng, Yueh-Hsuan, Chien-Hsun Chen, and Chuen-Tsai Sun. 2009. Toward the human-robot co-existence society: On safety intelligence for next generation robots. *International Journal of Social Robotics* 1 (4): 267–82.
Yamazaki Takashi. 2006. *Kango kaigo bunya ni okeru gaikokujin rōdōsha no ukeire mondai* [Problems in recruiting foreign health care workers]. *Referansu* 2: 4–24. Available at www.ndl.go.jp/jp/diet/publication/refer/200602_661/066101.pdf (accessed June 2013).
Ziemke, Tom. 2003. What's that thing called embodiment? In Richard Alterman and David Kirsh, eds. *Proceedings of the 25th annual conference of the Cognitive Science Society*. Mahwah, N.J.: Lawrence Erlbaum. 1134–39.

❑

South Korean Robot Ethics Charter 2012

THIS CHARTER WAS DRAFTED IN ORDER TO PREVENT SOCIAL ILLS THAT MAY ARISE OUT of inadequate social and legal measures to deal with robots in society.

Part 1: Manufacturing Standards

a) Robot manufacturers must ensure that the autonomy of the robots they design is limited; in the event that it becomes necessary, it must always be possible for a human being to assume control over a robot.

b) Robot manufacturers must maintain strict standards of quality control, taking all reasonable steps are taken to ensure that the risk of death or injury to the user is minimized, and the safety of the community guaranteed.

c) Robot manufacturers must take steps to ensure that the risk of psychological harm to users is minimized. 'Psychological harm' in this sense includes any likelihood for the robot to induce antisocial or sociopathic behaviors, depression or anxiety, stress, and particularly addictions (such as gambling addiction).

c) Robot manufacturers must ensure their product is clearly identifiable, and that this identification is protected from alteration.

d) Robots must be designed so as to protect personal data, through means of encryption and secure storage.

e) Robots must be designed so that their actions (online as well as real-world) are traceable at all times.

f) Robot design must be ecologically sensitive and sustainable.

Part 2: Rights & Responsibilities of Users/Owners

Sec. 1: Rights and Expectations of Owners and Users

i) Owners have the right to be able to take control of their robot.

ii) Owners and users have the right to use of their robot without risk or fear of physical or psychological harm.

iii) Users have the right to security of their personal details and other sensitive information.

iv) Owners and users have the right to expect a robot to perform any task for which it has been explicitly designed (subject to Section 2 of this Charter).

Sec. 2: Responsibilities of Owners and Users

This Charter recognizes the user's right to utilize a robot in any way they see fit, so long as this use remains 'fair' and 'legal' within the parameters of the law. As such:

i) A user must not use a robot to commit an illegal act.

ii) A user must not use a robot in a way that may be construed as causing physical or psychological harm to an individual.

iii) An owner must take 'reasonable precaution' to ensure that their robot does not pose a threat to the safety and well-being of individuals or their property.

Sec. 3: The following acts are an offense under Korean Law:

i) To *deliberately* damage or destroy a robot.

ii) Through gross negligence, to allow a robot to come to harm.

iii) It is a lesser but nonetheless serious offence to treat a robot in a way which may be construed as *deliberately and inordinately abusive*.

Part 3: Rights & Responsibilities for Robots

PART OF A THESIS BY CHRIS FIELD FOR THE UNIVERSITY OF TECHNOLOGY, SYDNEY

AKIKO

AKIKO'S CONSCIOUSNESS

RT @RoboThespian: You can watch RoboThespian's appearance on Channel 4's Ben Earl #trickartist here: channel4.com/programmes/ben... 9 years ago

Social robotics arrives at the University of Technology Sydney, with the creation of UTS' own PR2 tiny.cc/49ftew 10 years ago

UWS has begun an adopt-a-robot trial in Australia to monitor 'the subtle nuances of human-robot interactions.' tiny.cc/w4s04 11 years ago

Explore iiRobotics Audiovisual database of #robot developments: http://tiny.cc/6q243 11 years ago

Naho Kitano: one of few contemp. Japanese scholars to explore roboethics: http://www.roboethics.org/atelier2006/.../Kitano%20west%20japan.pdf 11 years ago

Japan's first robot actress to perform in Tokyo from November 11: http://tiny.cc/73mdu 11 years ago

Interactive Online Fiction on Human-Robot Relations in Japan http://tiny.cc/v48s1kf4fq 12 years ago

The 20th Kondo Cup Robot Soccer Game was held April this year. Will Robots have their own Games in the time to come? http://tiny.cc/7qy7z 12 years ago

Shortage of human teachers in rural areas could be supplemented by robots like Saya shar.es/0C7U8 12 years ago

I'm hoping to make 'Enlightenment of an Anchorwoman' a space where different #roboethics approaches can be discussed: http://tiny.cc/adhwg 12 years ago

Sec. 1: Responsibilities of Robots

i) A robot may not injure a human being or, through inaction, allow a human being to come to harm.

ii) A robot must obey any orders given to it by human beings, except where such orders would conflict with Part 3 Section 1 subsection "i" of this Charter.

iii) A robot must not deceive a human being.

Sec 2: Rights of Robots

Under Korean Law, Robots are afforded the following fundamental rights:

i) The right to exist without fear of injury or death.

ii) The right to live an existence free from systematic abuse.

The document above is a mock-up of what the South Korean Robot Ethics Charter (currently being drafted) may look like in the future, based on the limited information about the charter available from media reports (**Read the National Geographic article here**) . **What do you think about the ideas in Part 2 and 3 of this Charter? Assuming robots will one day be conscious beings, should users behavior toward the robots be limited? Should robots have fundamental rights? Have your say below.**

§ 3 Responses to *South Korean Robot Ethics Charter 2012*

Charter Languishes in U.S Senate « Enlightenment of an Anchorwoman
September 28, 2010 at 3:37 pm

[…] "Ten Principles of Robot Law"Military Robot Laws: A Continuum of Force?South Korean Robot Ethics Charter 2012Terasem's […]

Reply

Korea to Crack-down on Cracks in Ethics Legislation « Enlightenment of an Anchorwoman
October 3, 2010 at 7:23 am

[…] "Ten Principles of Robot Law"Military Robot Laws: A Continuum of Force?South Korean Robot Ethics Charter 2012Terasem's […]

Reply

Robot Ethic Charter, an initiative of the South Korean government. « grouptms18
January 31, 2012 at 1:41 pm

[…] Source: **https://akikok012um1.wordpress.com/south-korean-robot-ethics-charter-2012** […]

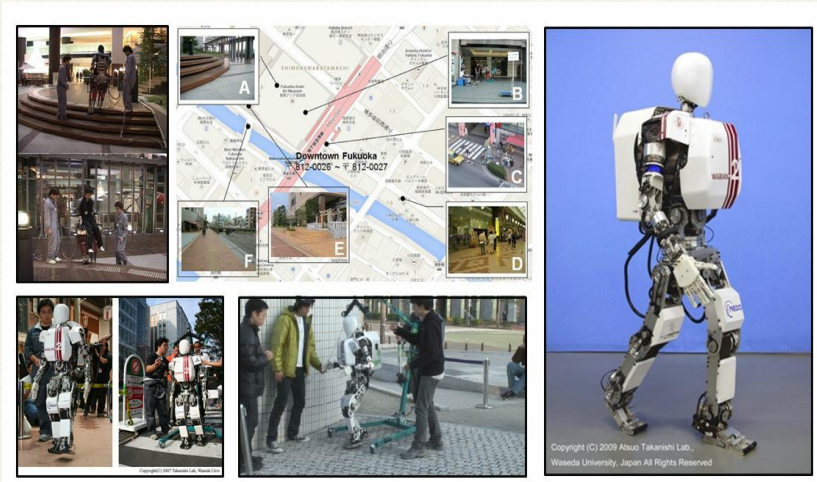
Reply

Leave a Reply

Enter your comment here...

ROBOLAW.ASIA Research

Robot Law and Ethics



Special Zone in Fukuoka: It is known as the world's first public roads testing for bipedal humanoid robots

Since 2004, the Japanese Ministry of Economy, Trade and Industry has published a series of Robot Policy Guidelines which address business and safety issues for "Next Generation Robots". They predicted a "Human-Robot Coexistence Society" that will emerge by 2030. However, it is a step-by-step gradual process for these robots entering into the everyday lives of people. We believe that intelligent robots will be the next revolutionary technology after PCs and the Internet. Therefore, we established ROBOLAW.ASIA Initiative to investigate the intersection between artificial intelligence & law. Our main objective is to minimize risks from robots into an acceptable range.

An emerging problem might be whether we should consider addressing new regulation impact on service robots. Under the current legal system, service robots are merely a property or "the second existence"; it is not enough to protect safety and moral risks in regards to human-robot co-existence. In other words, the new perspective of regulation shall be established under the premise of service robots as "the third existence" legal entity; robots are still the object of law, and they shall have a special legal status different from normal machines. However, the difficulty of implementing new regulations for service robots is something similar to the case of regulating steam powered cars in the 19th century. It's a "Regulation of Unknown". On one hand, such machines could cause deadly consequences to human beings without proper regulation. On the other hand, it is difficult for regulators to keep up with the progress of advanced technology. Therefore, thereis a tendency for over-regulation, similar to the case of the steam powered cars in the past.

To avoid repeating the Red Flag Laws in the era of intelligent robots, we can first consider "Deregulation" while referring the "Tokku" RT special zone. A special area such as this one can help regulators and manufacturers identify many unexpected risks during the final stage prior to the robots' practical application. Originated from Japan, the history of RT special zone is merely 10 years long, but there are already many special zones established in Fukuoka, Osaka, Gifu, Kanagawa and Tsukuba. As the development of robotics and its acceptance to society expand, the importance of special zones as an interface for robots and society will be more apparent.

Furthermore, we need to be aware of the importance of public law and regulation. While it does not refer to the debate on issues of robot rights or robots to be recognized as the subject of law from the Constitution, it does mention making public regulation for the design, manufacture, selling, and usage of advanced robotics. A possibility could be developing the "Robot Safety Governance Act", which is the extension of current machine safety regulations. These technical norms located at the bottom of "Robot Law" will ensure the safety of new human-robot co-existence.

Finally, robot ethics and legal regulation should not always be in parallel, because from the regulation perspective, robot law is an intersection of robot ethics and robotics. We might don't need Red Flag Laws for Pepper robots, but it depends on what moral stands and actions we take toward the regulation of unknown.

Latest News

New Website Launched

April 1st, 2017

2017 sees the redesign of our website. Take a look around and let us know what you think.
Read more

ROBOLAW.ASIA
亚洲机器人法律研究网络

AI, Robotics & Law

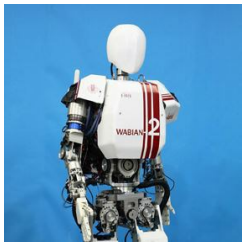
- Peking University Law School
- Waseda University HRI
- SSSA BioRobotics Institute
- EU FP7 Project: ROBOLAW

Legal Informatics

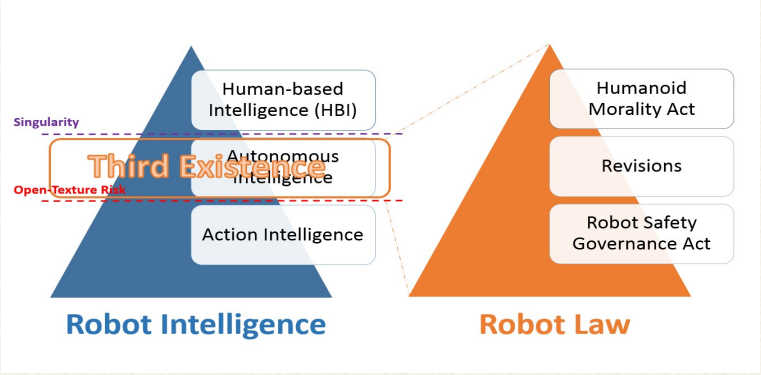
- CHINA-LII
- ITTIG-CNR, Florence
- Legal Information Institute
- Beida.ChinaLawInfo.com

IEEE STANDARDS ASSOCIATION
IEEE

The IEEE Global Initiative **is our strategic partner in AI Ethics and Governance**



[TLC Paper] ***Japan's Robot Policy and the Special Zone...***
Read more



Two Pyramids for Robot Regulation

Contribution:

01. Yueh-Hsuan Weng (2018) *The Ideal and the Reality of Human-Robot Co-Existence: "Tokku" RT Special Zone and Red Flag Laws*, Kuu-young Young et al. (Eds.), Technology, Humans, and Society 3, Page 61-72, NCTU Press, ISBN: 9789869477260 [LINK]

02. Yueh-Hsuan Weng (2017) *Towards Integrated Governance for Intelligent Robots: A Focus on Social System Design*, Special Issue on AI & Law, Jusletter IT 23 Nov. 2017, Weblaw (Bern), ISSN 1424-7410 [LINK]

03. Yueh-Hsuan Weng (2017) *The Future of Work: A Brief Look to Legal Impacts to Automation and Labor Force*, in Proceedings of 2017 KLRI Legal Scholars Roundtable: How Law Operates in the Wired Society, JW Marriott Hotel Seoul, Korea Legal Research Institute, September 21-22 [LINK]

02. Yueh-Hsuan Weng (2017) *Case Study: Bad Actors and Service Robots, FHI-CESR-CFI Workshop on Bad Actors and Artificial Intelligence*, 14:45-15:00, Littlegate House, Oxford, February 19th-20th 2017 [LINK]

03. Yueh-Hsuan Weng (2016) *Robot Law 1.0: On Social System Design for Artificial Intelligence*, 13:00-14:00, Small Moot Court, room 723, 7/F, Cheng Yu Tung Tower, Faculty of Law, The University of Hong Kong, January 16th 2017 [LINK]

04. Yueh-Hsuan Weng (2016) *Regulation of Unknown: A Lesson from Japan's Public Law and Policy for Next-Generation Robots*, 1st Annual Conference of the Center for Law and Internet (CLI) Session: Ethics and technology, 15:20-16:30, West-Indisch Huis Amsterdam, November 17th 2016 [LINK]

05. Mady Delvaux-Stehres and Yueh-Hsuan Weng (2016) *A European perspective on robot law: Interview with Mady Delvaux-Stehres*, TECH and LAW Center & Robohub [LINK]

06. Yueh-Hsuan Weng (2016) *Regulation of Unknown: A Lesson from Japan's Public Law and Policy for Next-Generation Robots*, 12:30-13:30, Room 623, 6/F, Cheng Yu Tung Tower, Faculty of Law, The University of Hong Kong, January 28th 2016 [LINK]

07. Christof Heyns, Gurvinder S. Virk, Yueh-Hsuan Weng (2015) *An Exclusive Interview with UN and ISO experts in Robots and Regulation*, TECH and LAW Center [LINK]

08. Yueh-Hsuan Weng (2015) *O Direito para Robôs: A Regulação do Direito Robótico no Direito Público do Japão*, 10:30-12:30, Sala de Vídeoconferência - 3° andar, UFPR - Universidade Federal do Paraná, Curitiba, November 20th 2015 [LINK]

09. Yueh-Hsuan Weng (2015) *Regulation of Unknown: Does the Humanoid Robot "PEPPER" need Red Flag Laws?*, TECH and LAW Center [LINK]

10. Yueh-Hsuan Weng (2015) *Japan's Robot Policy and the Special Zone for Regulating Next Generation Robots*, TECH and LAW Center [LINK]

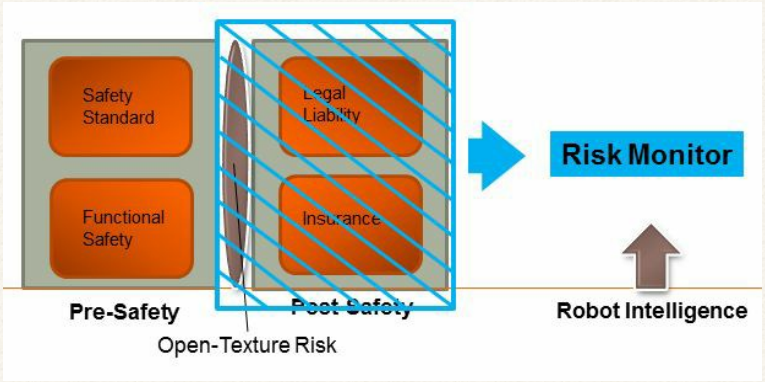


Figure 1. Risk Monitoring Mechanism: For short term consideration, a passive measurement to monitor the Open-Texture Risk.

[TLC Review] *AI Ethics and Car Wars* by Joanna J. Bryson Read more

[TLC News] *Robotics: A New Challenge for Security…* Read more

[TLC Interview] *Cyber-Humans: Our Future with Machines…* Read more

[TLC Interview] *A European Perspective on Robot Law…* Read more

[Guest Post] *When Robots Kill: Ethics in an Automatized…* Read more

[TLC Interview] *The Quest for Roboethics: An Interview …* Read more

Figure 2. Risk Control Mechanism: For Long term consideration, using active measurement to absorb the Open-Texture Risk.

11. Yueh-Hsuan Weng (2015) *Japanese Public Policy for Robots and Regulation: An Example of "Tokku" Special Zone*, 11:00-13:00, Sala Mansarda, Villa Schifanoia, European University Institute, Florence, April 28th 2015 [LINK]

12. Yueh-Hsuan Weng (2015) *Robots and Society: On the Intersection of Special Zone, Robots, and the Law*, 10:30-11:30, Via dei Barucci n° 20, ITTIG-CNR, Florence, March 26th 2015 [LINK]

13. Yueh-Hsuan Weng, Yusuke Sugahara, Kenji Hashimoto, Atsuo Takanishi (2015) *Intersection of "Tokku" Special Zone, Robots, and the Law: A Case Study on Legal Impacts to Humanoid Robots*, International Journal of Social Robotics, Vol. 7, No. 5, Page 841-857, Springer Netherlandsa [LINK]

14. Yueh-Hsuan Weng, Gurvinder S. Virk, Shuping Yang (2015) *The Safety for Human-Robot Co-Existing: On New ISO 13482 Safety Standard for Service Robots*, Internet Law Review, Vol. 17, Page 229-244, Peking University Press [LINK]

15. Yueh-Hsuan Weng (2014) *Introduction to Applications and Challenges of Emerging Technologies*, Master Course: The Laws of Cyberspace, 10:00-12:00, Room 121, TSMC Building, National Tsing Hua University, Hsinchu, December 11th 2014

16. Yueh-Hsuan Weng, Dominic Hillenbrand (2014) *The Intelligentization of Automobiles: Smart-Cars, Robo-Cars and their Safety Governance*, Journal of Science, Technology and Law (STL), No.4, General Issue 110, Page 632-646, 2014 [LINK]

17. Yueh-Hsuan Weng (2014) *The Study of Safety Governance for Service Robots: On Open-Texture Risk*, Ph.D. Dissertation, Peking University Law School, Beijing, May, 2014 [LINK]

18. Yueh-Hsuan Weng (2014) *A Review on Military Robots and Warfare*, PKU Internet Law Watch, Vol. 10, No. 4

19. Yueh-Hsuan Weng (2014) *The Robot - Technology, Ethics and Law*, PKU Internet Law Watch, Vol. 10, No. 3

20. Yueh-Hsuan Weng (2014) *Robots - A Historical Review*, PKU Internet Law Watch, Vol. 10, No. 2

21. Yueh-Hsuan Weng (2013) *Special Interview on "Robot Law in Europe" - with Prof. Dr. Eric Hilgendorf*, ROBOLAW.ASIA Initiative [LINK]

22. Yueh-Hsuan Weng (2013) *Special Interview on "Law and Drone Warfare" - with Prof. Dr. Christof Heyns* , ROBOLAW.ASIA Initiative [LINK]

23. Yueh-Hsuan Weng, Sophie T.H. Zhao (2012) *The Legal Challenges of Networked Robotics: From the Safety Intelligence Perspective*, M. Palmirani et al. (Eds.), Lecture Notes in Computer Science(LNCS): AI Approaches to the Complexity of Legal Systems, Vol. 7639, Page 61-72, Springer Berlin Heidelberg [LINK]

24. Yueh-Hsuan Weng (2012-2014) *Supporting External Network, EU FP7 Project: ROBOLAW*, Scuola Superiore Sant'Anna, Pisa, Italy, June 2012 - September 2014 [LINK]

25. Yueh-Hsuan Weng (2012) *Yahoo! Research Grant, "The Internet of Things and Automation: Legislation and Policy Research"*, PKU-Yahoo! Internet Law Center

26. Yueh-Hsuan Weng (2012) *(1) The Internet of Things and Automation: Overlapping the Real and Virtual Worlds*, PKU Internet Law Watch, Vol. 8, No. 5

27. Yueh-Hsuan Weng (2012) *(2) Intelligent Transportation: Addressing the Liability Impact of Automated Systems: with Prof. Dr. Giovanni Sartor and Dr. Giuseppe Contissa*, PKU Internet Law Watch, Vol. 8, No. 5

28. Yueh-Hsuan Weng (2012) *(4) Social Robots: Robot Companions for Citizens: with Prof. Dr. Paolo Dario*, PKU Internet Law Watch, Vol. 8, No. 5

29. Yueh-Hsuan Weng (2012) *Law & Networked Robotics: Some legal Issues on the Internet of Things*, Aula 6, 12:00-15:00, SSSA Seminar, Scuola Superiore Sant'Anna, Pisa, June 6th 2012

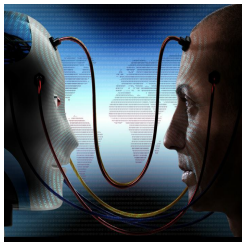
30. Yueh-Hsuan Weng, Sophie T.H. Zhao (2011) *The Legal Challenges of Networked Robotics: From the Safety Intelligence Perspective*, XXV. IVR World Congress on Philosophy of Law and Social Philosophy Special Workshop on AICOL, (16:00~16:20) Room HZ 8, Building N.4, Goethe-Universität Frankfurt, Frankfurt am Main, 15th August 2011 [LINK]

31. Yueh-Hsuan Weng (2011) *Networked Robots: A Brief Look at Its Possible Legal Implications*, IEEE International Conference on Robotics and Automation (IEEE ICRA'11) Workshop on Roboethics, (16:30~17:00) Room 5B, SHICC, Shanghai, 13th May 2011 [PDF]

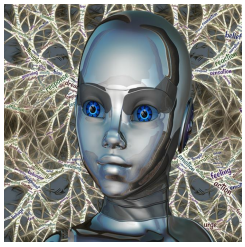
32. Yueh-Hsuan Weng (2011) *The Open-Texture Risk in the Human-Robot Co-Existence Society: A Review on "Open Robotics"*, Internet Law Review, Vol. 13, Peking University Press [LINK]

33. Yueh-Hsuan Weng (2009) *Toward The Human-Robot Co-Existence Society: On Legislative Consortium for Social Robotics*, IEEE International Conference on Robotics and Automation (IEEE ICRA'09) Workshop on Legal and Safety Issues Related to Autonomous Networked Robots Operating in Urban Environments, (10:40~11:05) Room 404, Kobe International Convention Center, Kobe, 13th May 2009 [LINK]

34. Yueh-Hsuan Weng, Chien-Hsun Chen and Cheun-Tsai Sun (2007) *The Legal Crisis of Next Generation Robots: On Safety Intelligence*, Paper presented on The Eleventh International Conference on Artificial Intelligence and Law (ICAIL'07). Stanford Law School, Palo Alto, California, USA [Acceptance Rate: 26%] [PDF]



[TLC Interview] ***How do we regulate robo-morality?*** **…** Read more

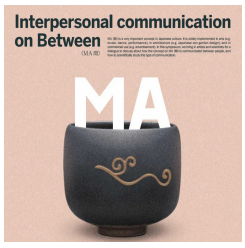


[TLC Review] ***Artificial people: How will the law adapt to …***
Read more

[TLC Interview] ***An Exclusive Interview with UN and ISO…***
Read more

[TLC Review] ***Regulation of Unknown: Does the Humanoid…***
Read more

[TLC Interview] ***Technical Challenges in Machinie Ethics*** **…** Read more



[RIEC Workshop] ***Interpersonal***

Communication on Between …
Read more