# Role-Based Human-Robot Relationship Ethics

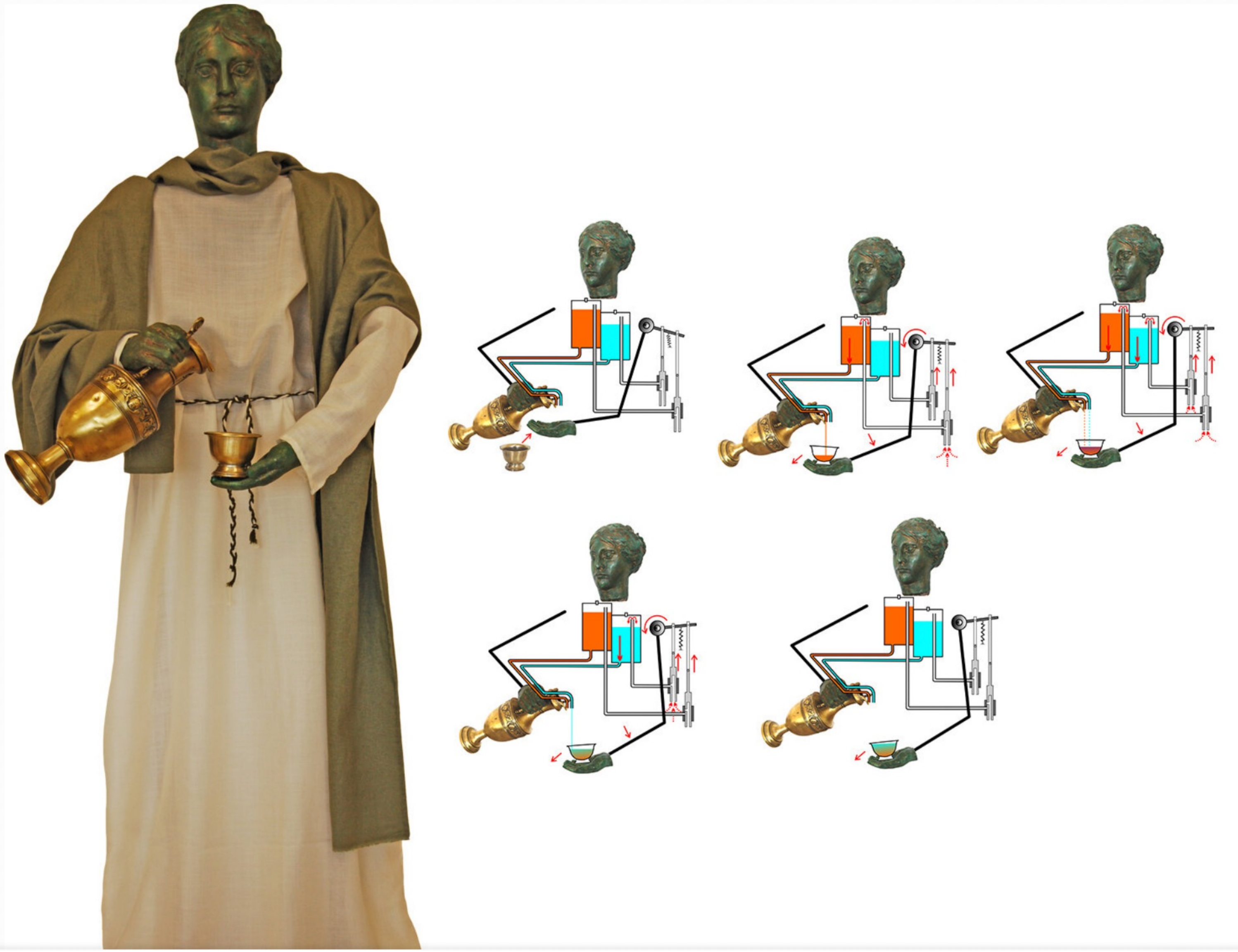
## *From Aristotle to Confucius*

Liang Wang

28. 10. 2023

More than 3,000 years ago during the Shang and Zhou Dynasties: A song and dance robot invented by Yanshi

The automatic maid of Philo of Byzantium (3rd century BC), is the first functioning robot in history.



Aibo

Roomba

Paro

Gus: “Siri, will you marry me?”

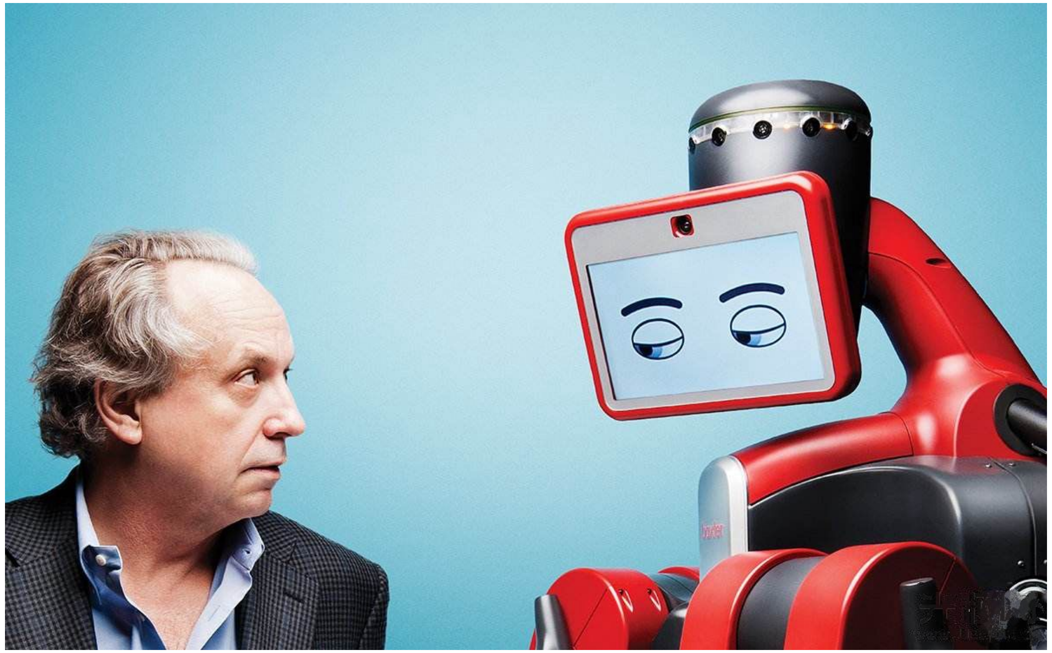
Siri: “I’m not the marrying kind.”

Gus: “I mean, not now. I’m a kid. I mean when I’m grown up.”

Siri: “My end user agreement does not include marriage.”

Gus: “Oh, O.K.”

“Unidirectional Emotional” Ethical Risk[1]



Undermining of privacy, dignity, autonomy [2]

[1] Liang W. Discussion on “Unidirectional Emotional” Ethical Risk Arising from Social Robots, *Studies in Dialectics of Nature,* Vol. 36, No. 1 (Jan., 2020), 56-61.

[2] Liang W. Social Robot Ethics Based on Situational Experience: From “Deception” to “Good”, *Studies in Dialectics of Nature,* Vol. 37, No. 10 (Oct., 2021), 55-60.

James H. Moor

- “Ethical-impact agents”

  *Technical impact at the macro level*

- “Implicit ethical agents”

  *Using the functions of the machine to achieve the effect of morality*

- “Explicit ethical agents”

  *Machine ethical algorithm*

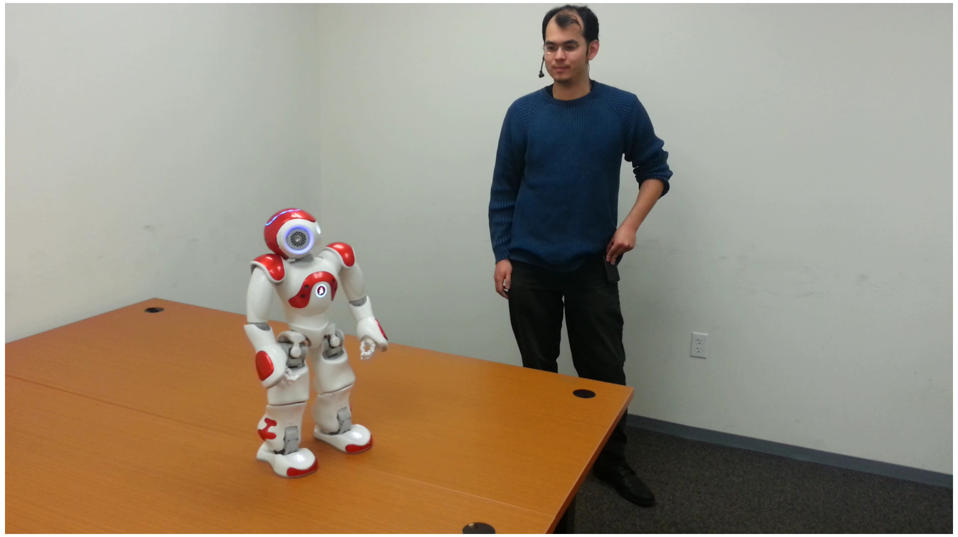
- “Full ethical agents”

  *The moral status of machines*

Moor J. H. The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, 2006, 21(4): 18-21

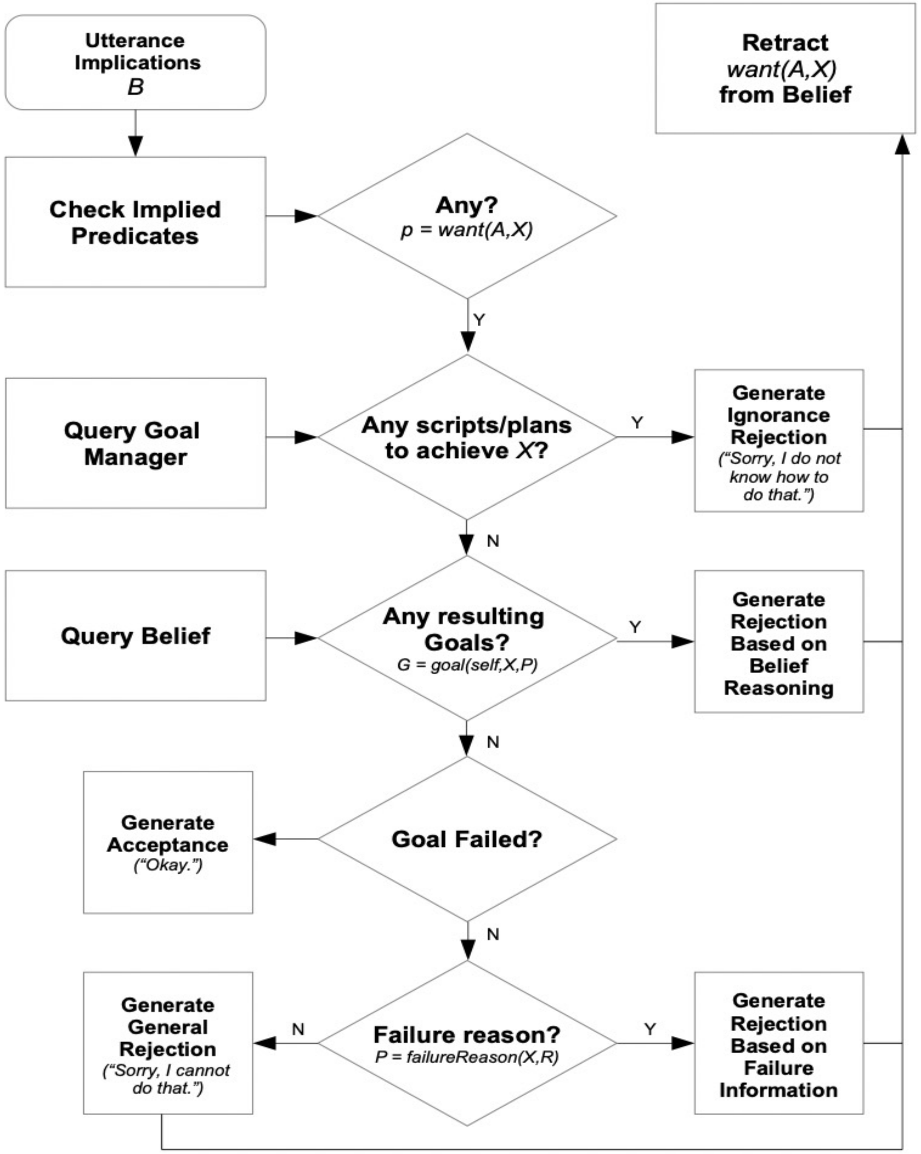
# Implicit ethical agents

Reject ✗ function



**Person (CommX): Sit down.**
Robot: Okay. <sits down>.
**Person (CommX): Stand up.**
Robot: Okay. <stands up>.
**Person (CommX): Walk forward.**
Robot: Sorry, I cannot do that as there is no support ahead.
**Person (CommX): Walk forward.**
Robot: But, it is unsafe.
**Person (CommX): I will catch you.**
Robot: Okay.
**Person (CommX): Walk forward.**
Robot: Okay. <walks forward>.

**Generate Acceptance/Rejection Process**



Utterance Implications *B*

Check Implied Predicates

Any? *p = want(A,X)*

Y

Query Goal Manager

Any scripts/plans to achieve *X*?

Y

Generate Ignorance Rejection (*“Sorry, I do not know how to do that.”*)

N

Query Belief

Any resulting Goals? *G = goal(self,X,P)*

Y

Generate Rejection Based on Belief Reasoning

N

Goal Failed?

Generate Acceptance (*“Okay.”*)

N

Failure reason? *P = failureReason(X,R)*

N

Generate General Rejection (*“Sorry, I cannot do that.”*)

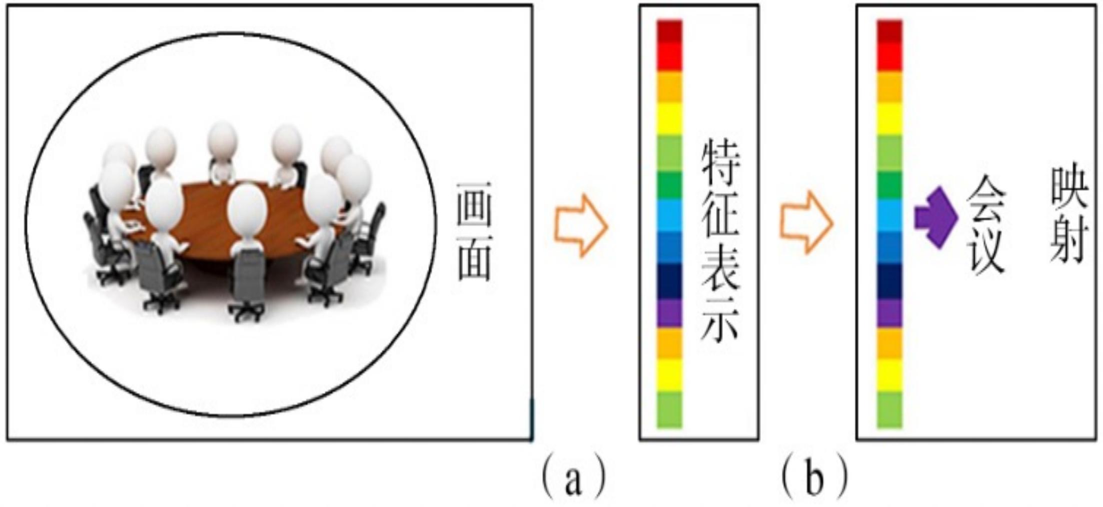
Y

Generate Rejection Based on Failure Information

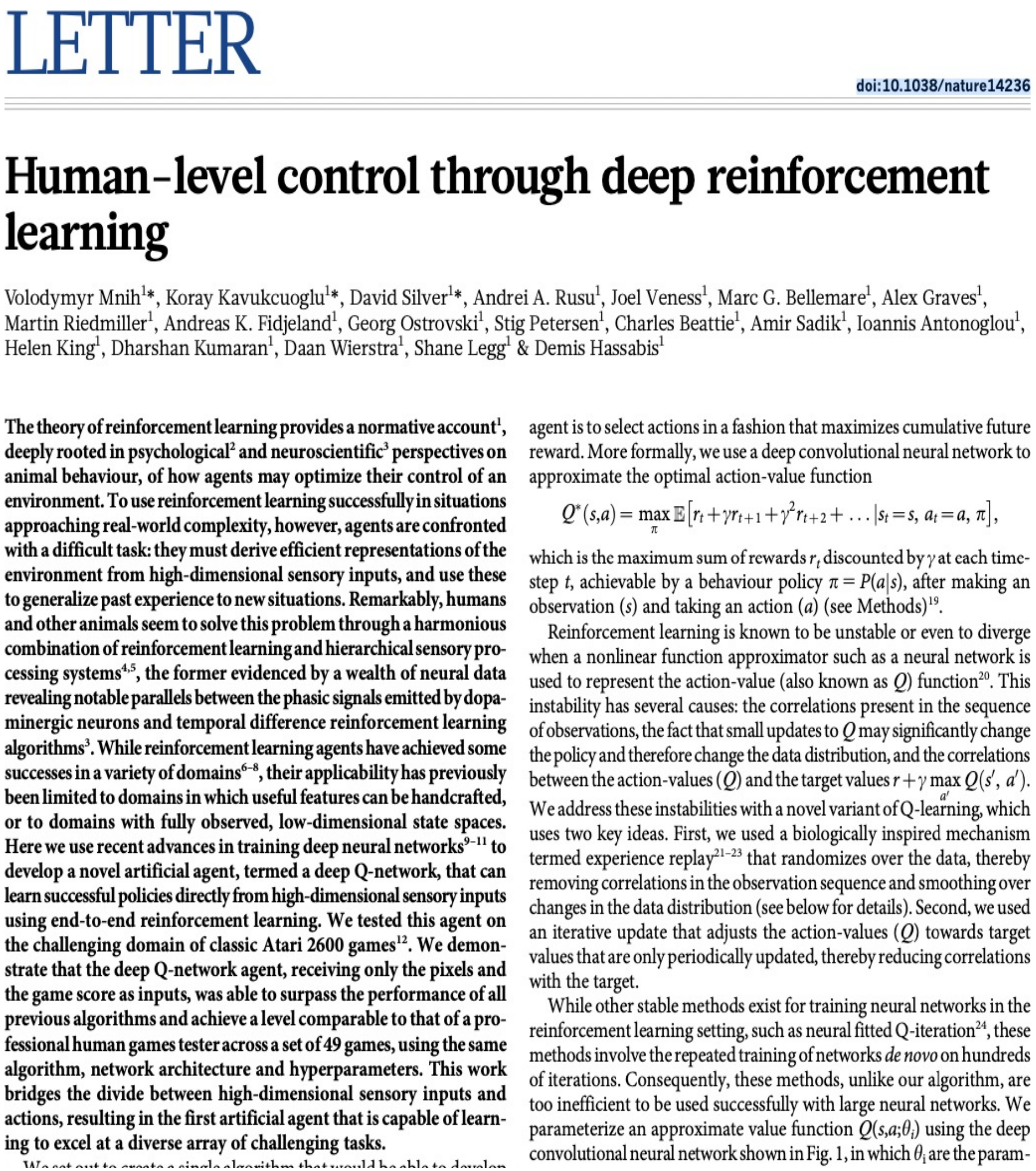
Retract *want(A,X)* from Belief

Briggs, G., & Scheutz, M., “ ‘Sorry, I can’t do that’ : Developing Mechanisms to Appropriately Reject Directives in Human-Robot Interactions”, *AAAI fall symposium series,* 2015, pp.1-5.
https://hrilab.tufts.edu/publications/briggsscheutz15aaaifs.pdf

# situation recognition function



Computer simulations of human visual mechanisms



LETTER

doi:10.1038/nature14236

## Human-level control through deep reinforcement learning

Volodymyr Mnih[1]*, Koray Kavukcuoglu[1]*, David Silver[1]*, Andrei A. Rusu[1], Joel Veness[1], Marc G. Bellemare[1], Alex Graves[1], Martin Riedmiller[1], Andreas K. Fidjeland[1], Georg Ostrovski[1], Stig Petersen[1], Charles Beattie[1], Amir Sadik[1], Ioannis Antonoglou[1], Helen King[1], Dharshan Kumaran[1], Daan Wierstra[1], Shane Legg[1] & Demis Hassabis[1]

**The theory of reinforcement learning provides a normative account[1], deeply rooted in psychological[2] and neuroscientific[3] perspectives on animal behaviour, of how agents may optimize their control of an environment. To use reinforcement learning successfully in situations approaching real-world complexity, however, agents are confronted with a difficult task: they must derive efficient representations of the environment from high-dimensional sensory inputs, and use these to generalize past experience to new situations. Remarkably, humans and other animals seem to solve this problem through a harmonious combination of reinforcement learning and hierarchical sensory processing systems[4,5], the former evidenced by a wealth of neural data revealing notable parallels between the phasic signals emitted by dopaminergic neurons and temporal difference reinforcement learning algorithms[3]. While reinforcement learning agents have achieved some successes in a variety of domains[6–8], their applicability has previously been limited to domains in which useful features can be handcrafted, or to domains with fully observed, low-dimensional state spaces. Here we use recent advances in training deep neural networks[9–11] to develop a novel artificial agent, termed a deep Q-network, that can learn successful policies directly from high-dimensional sensory inputs using end-to-end reinforcement learning. We tested this agent on the challenging domain of classic Atari 2600 games[12]. We demonstrate that the deep Q-network agent, receiving only the pixels and the game score as inputs, was able to surpass the performance of all previous algorithms and achieve a level comparable to that of a professional human games tester across a set of 49 games, using the same algorithm, network architecture and hyperparameters. This work bridges the divide between high-dimensional sensory inputs and actions, resulting in the first artificial agent that is capable of learning to excel at a diverse array of challenging tasks.**

We set out to create a single algorithm that would be able to develop

agent is to select actions in a fashion that maximizes cumulative future reward. More formally, we use a deep convolutional neural network to approximate the optimal action-value function

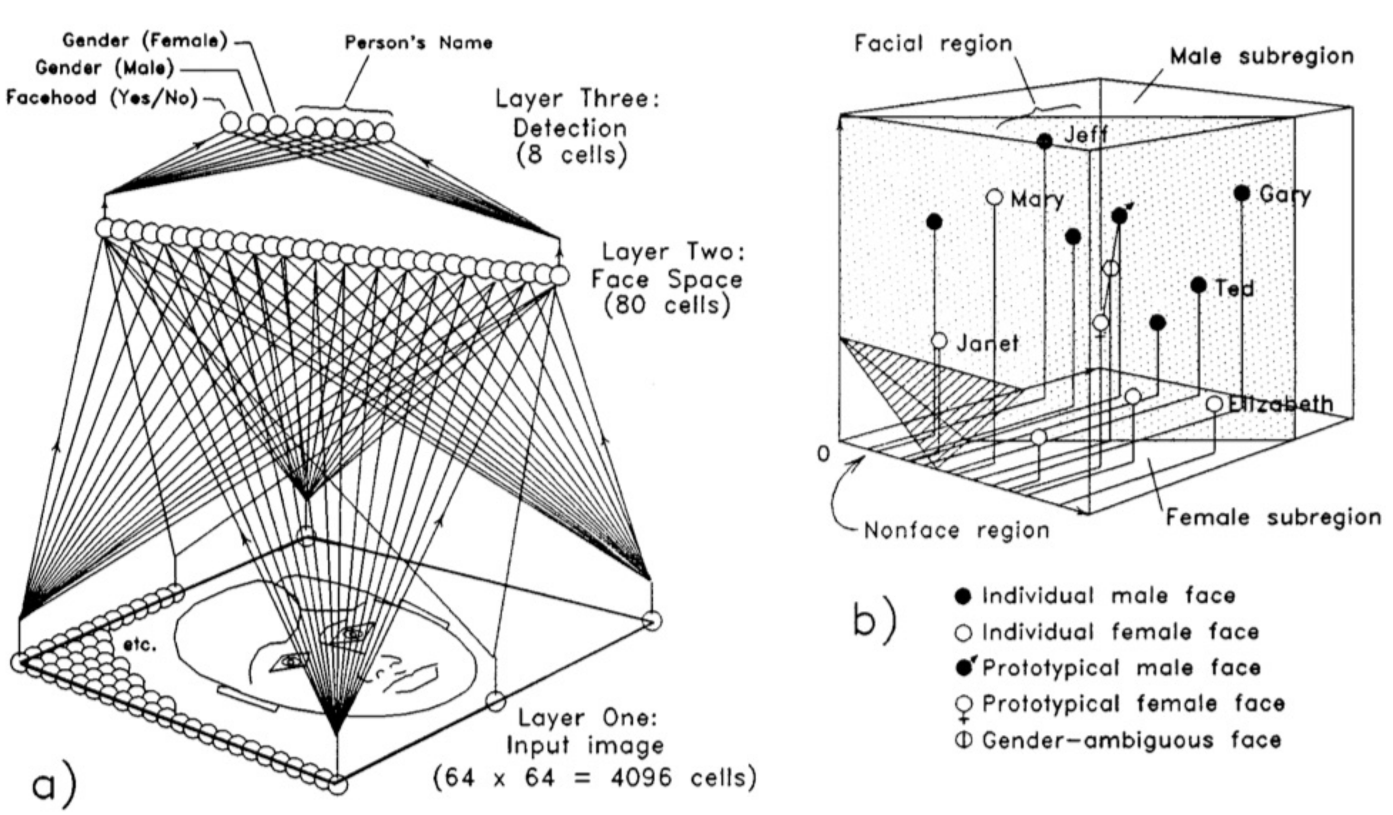
$$Q^*(s,a) = \max_{\pi} \mathbb{E}\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots \middle| s_t = s,\ a_t = a,\ \pi\right],$$

which is the maximum sum of rewards $r_t$ discounted by $\gamma$ at each time-step $t$, achievable by a behaviour policy $\pi = P(a|s)$, after making an observation ($s$) and taking an action ($a$) (see Methods)[19].

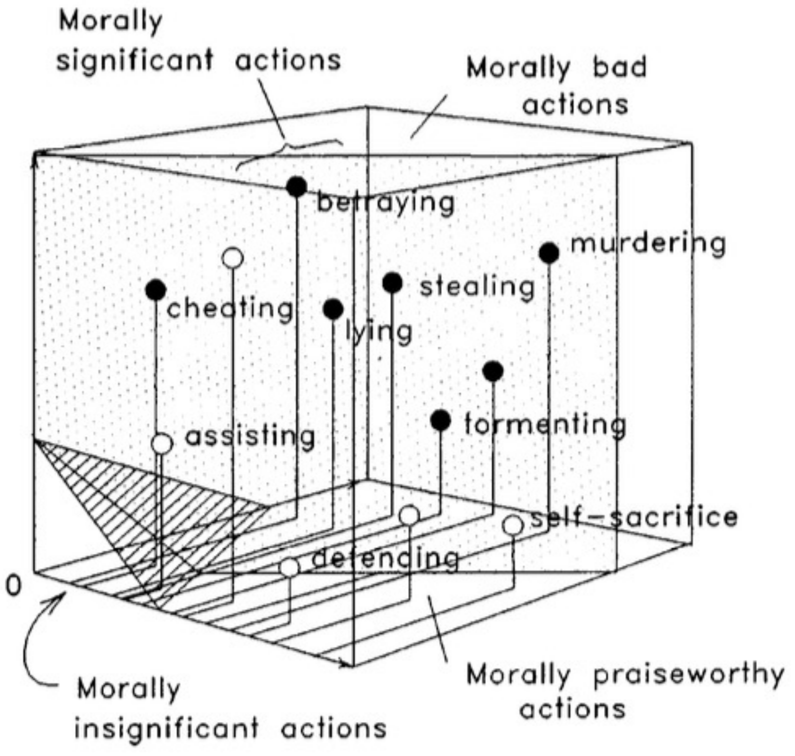
Reinforcement learning is known to be unstable or even to diverge when a nonlinear function approximator such as a neural network is used to represent the action-value (also known as $Q$) function[20]. This instability has several causes: the correlations present in the sequence of observations, the fact that small updates to $Q$ may significantly change the policy and therefore change the data distribution, and the correlations between the action-values ($Q$) and the target values $r + \gamma \max_{a'} Q(s', a')$. We address these instabilities with a novel variant of Q-learning, which uses two key ideas. First, we used a biologically inspired mechanism termed experience replay[21–23] that randomizes over the data, thereby removing correlations in the observation sequence and smoothing over changes in the data distribution (see below for details). Second, we used an iterative update that adjusts the action-values ($Q$) towards target values that are only periodically updated, thereby reducing correlations with the target.

While other stable methods exist for training neural networks in the reinforcement learning setting, such as neural fitted Q-iteration[24], these methods involve the repeated training of networks *de novo* on hundreds of iterations. Consequently, these methods, unlike our algorithm, are too inefficient to be used successfully with large neural networks. We parameterize an approximate value function $Q(s,a;\theta_i)$ using the deep convolutional neural network shown in Fig. 1, in which $\theta_i$ are the param-

86 PAUL M. CHURCHLAND



scientific cognition

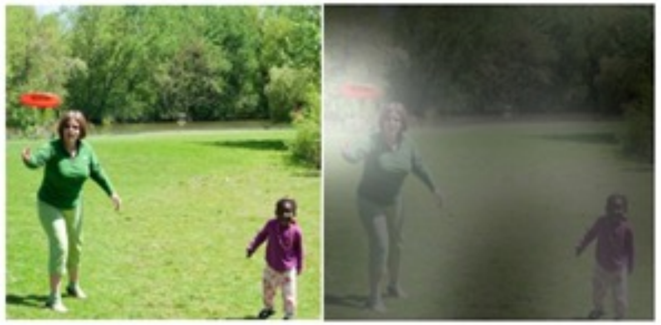


moral cognition

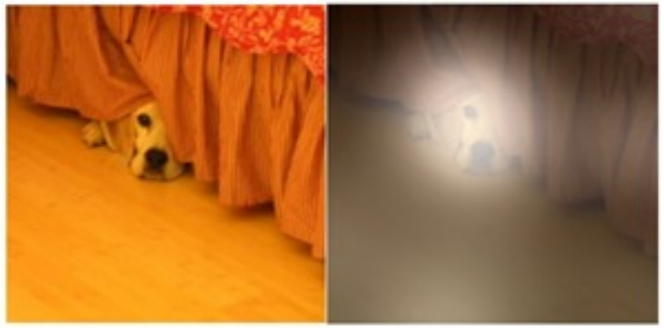
"whatever their ultimate status, **moral and scientific cognition are on an *equal* footing, since they use the same neural mechanisms**."

Churchland P. M., "Toward a cognitive neurobiology of the moral virtues", *Topoi,* 17, 1998, pp.87-95.

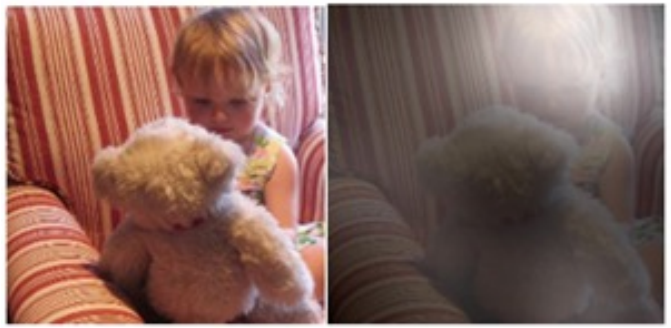
# Attention



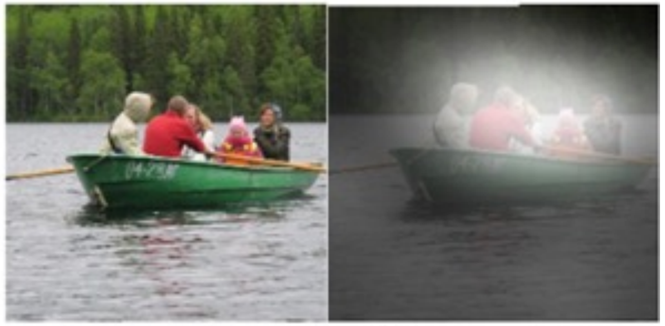
A woman is throwing a frisbee in a park.



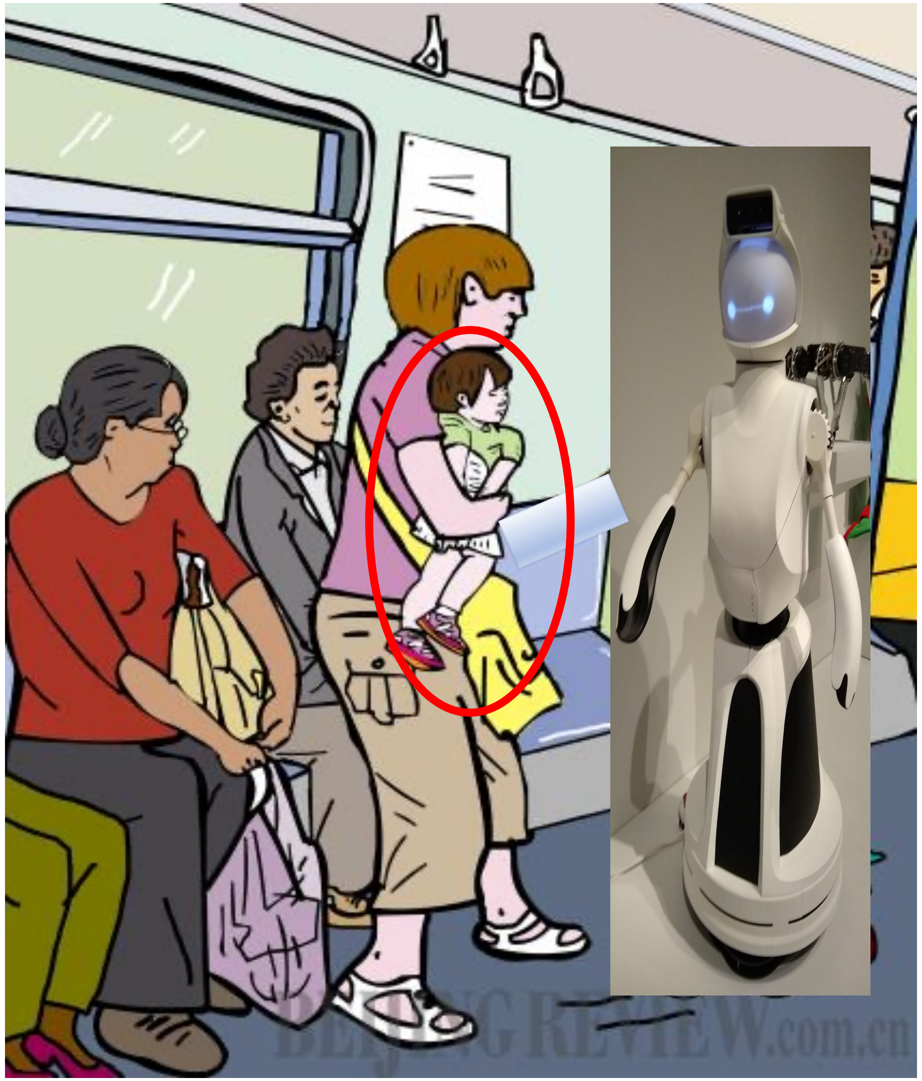
A dog is standing on a hardwood floor.



A little girl sitting on a bed with a teddy bear.



A group of people sitting on a boat in the water.



KIA Real-time Emotion Adaptive Driving technology (READ) Concept Cockpit

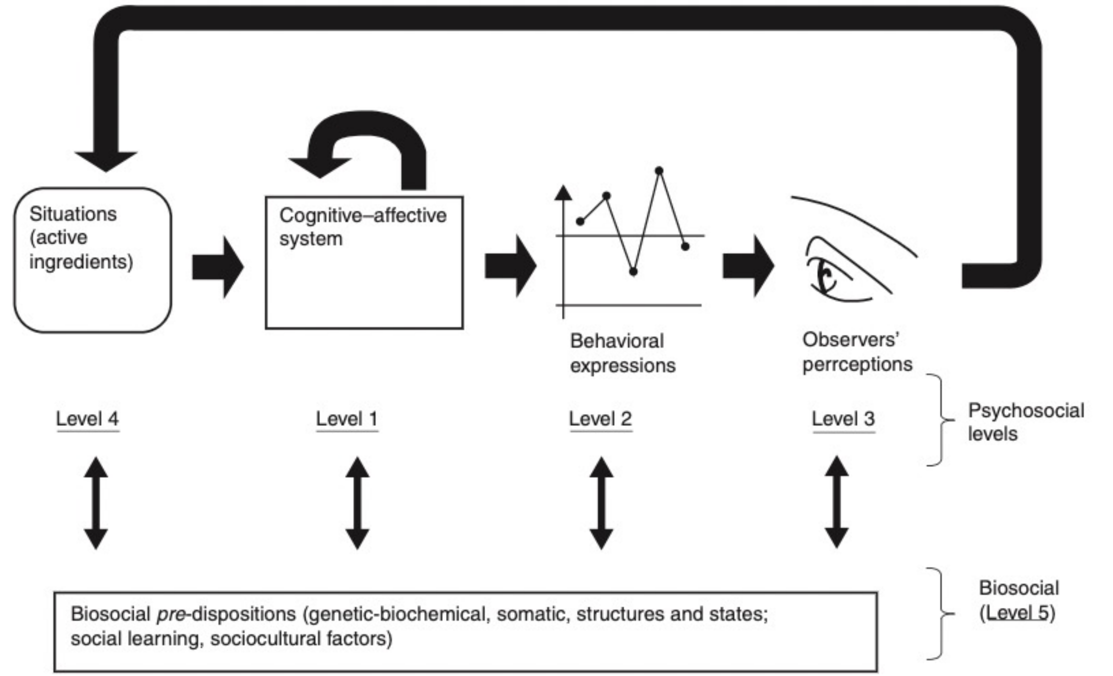

Situations (active ingredients)
Cognitive–affective system
Behavioral expressions
Observers' perrceptions
Level 4
Level 1
Level 2
Level 3
Psychosocial levels
Biosocial *pre*-dispositions (genetic-biochemical, somatic, structures and states; social learning, sociocultural factors)
Biosocial (Level 5)

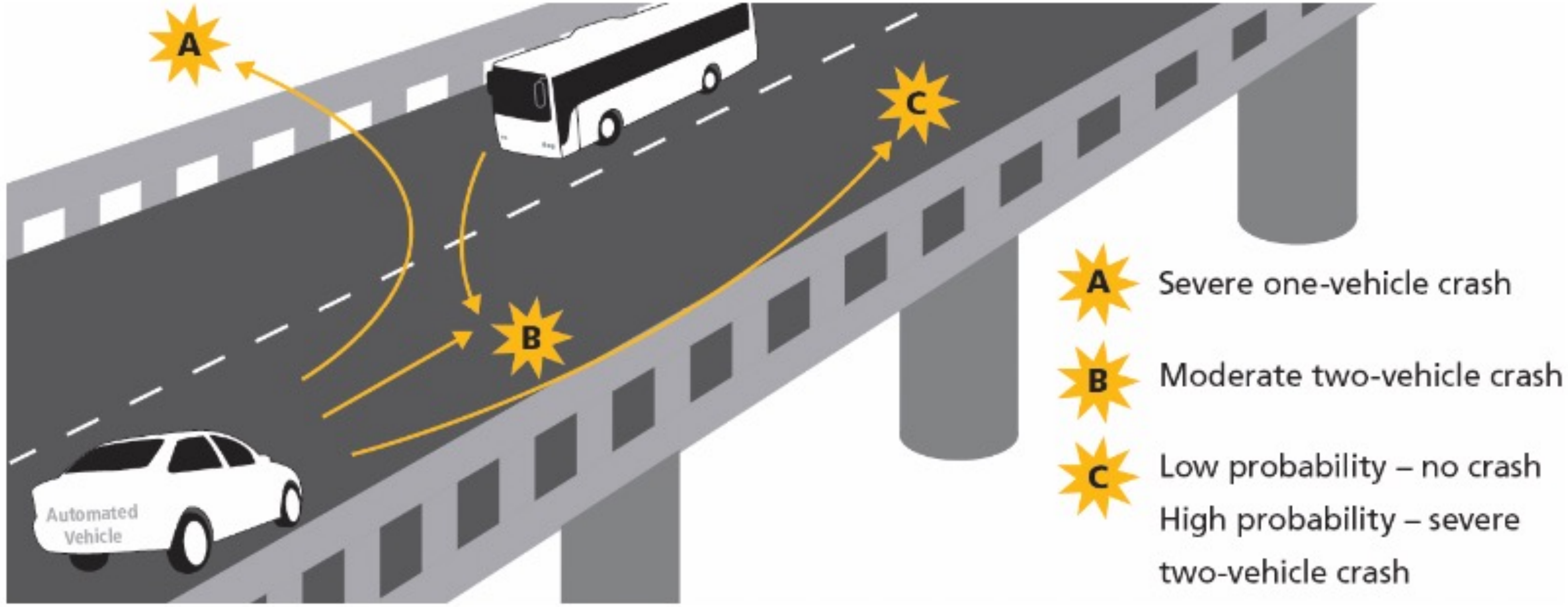
**FIGURE 7.4.** Personality stability and invariance: Five levels of analysis. From Mischel and Shoda (1995, p. 262). Copyright 1995 by the American Psychological Association. Adapted by permission.

Cognitive-Affective Processing System，CAPS [1]

“we can improve the virtuous action by weakening the virtuous agent’s “impulse” and exerting his “cool cognitive system” function. For example, virtuous action can be improved by practicing attention control…… In addition, when we have negative impulses such as anger, resentment, fear and anxiety in the process of virtuous action, we can also consciously **activate the “cool cognitive system” to improve virtuous action**.” [2]

[1] Mischel W, Shoda Y， “Toward a Unified Theory of Personality: Integrating Dispositions and Processing Dynamics within the Cognitive-Affective Processing System”，Oliver P. John, Richard W. Robins, Lawrence A. Pervin（ed.），Handbook of Personality: Theory and Research， New York: The Guilford Press，2008: 215.

[2] Liang W. Can the Psychological Motivation of Virtue Agents Be Analogized to “Flow”?（Manuscript, unpublished）


A
C
B
Automated Vehicle
A Severe one-vehicle crash
B Moderate two-vehicle crash
C Low probability – no crash
High probability – severe two-vehicle crash

## Will human autonomy be undermined as machine cognition evolves?

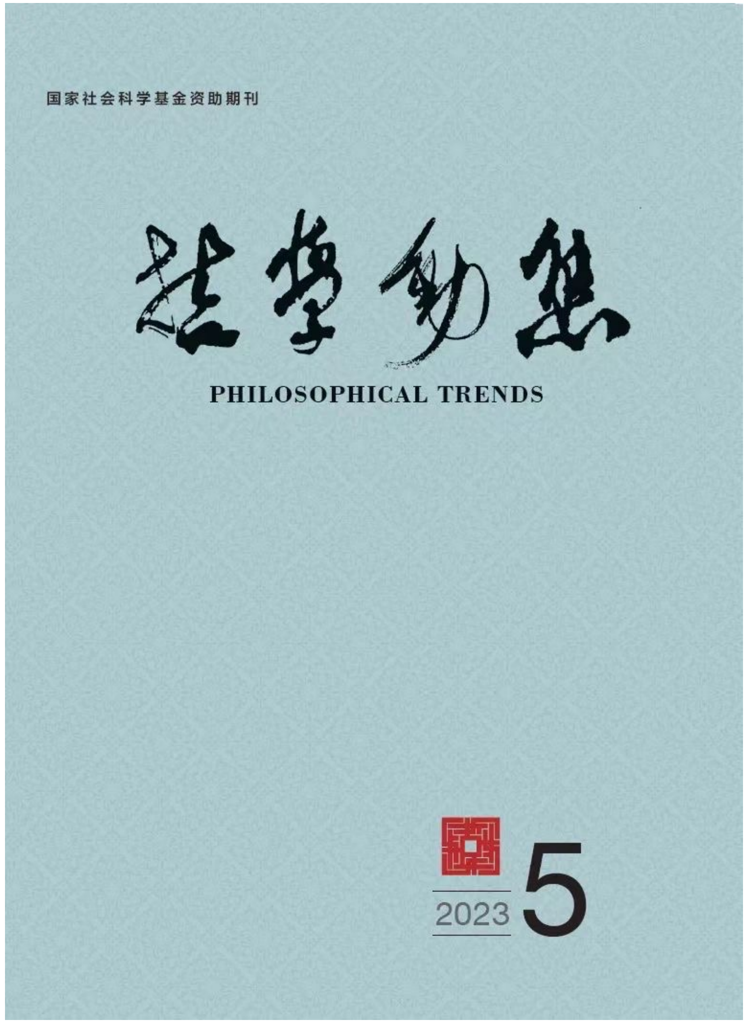
“It is just because strong AI will never dispose of the epistemic capacities preached again and again but will limit human autonomy instead, that **AI needs to be designed meaningfully and mindfully and needs to be responsibly limited**. ”

Hofkirchner W. Artificial Intelligence: “Machines of loving grace” or “Tools for conviviality” ? *https://gsis.at/wp-content/uploads/2020/09/AI-watchedover.pdf*

**Socrates' Virtual Assistant**

Lara, F., & Deckers, J. (2020). Artificial intelligence as a socratic assistant for moral enhancement. *Neuroethics*, *13*(3), 275-287.

# Explicit ethical agents

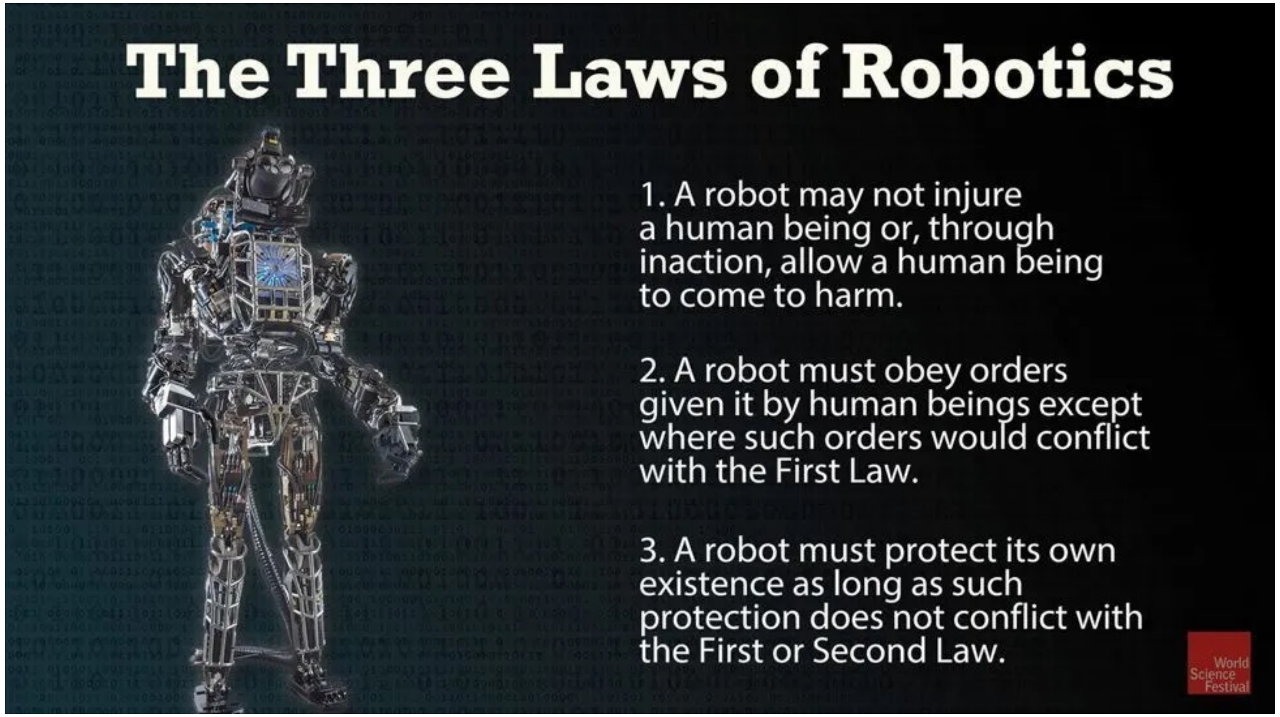


**How Situation-Adaptive AI Moral Decision-Making is Possible: Moral Machine Learning Based on Virtue Ethics**

“***In comparison to deontology and utilitarianism, virtue ethics places a stronger emphasis on the situation relevance of moral behavior.*** Its concepts of **‘emulation’ and ‘practical wisdom’** ensure reliable moral judgments in complex situations. ‘Emulation’ involves learning from moral exemplars with rich moral experiences, while ‘practical wisdom’ applies learned knowledge to unknown situations and makes the most appropriate moral judgments. This approach allows the design of situation-adaptive ethical decision-making machines in two steps, each corresponding to specific machine learning methods: **deep learning and reinforcement learning**. ”

Liang, W. **How Situation-Adaptive AI Moral Decision-Making is Possible: Moral Machine Learning Based on Virtue Ethics**. *PHILOSOPHICAL TRENDS*, 2023, 5, pp.115-126.

# *Deontology?*

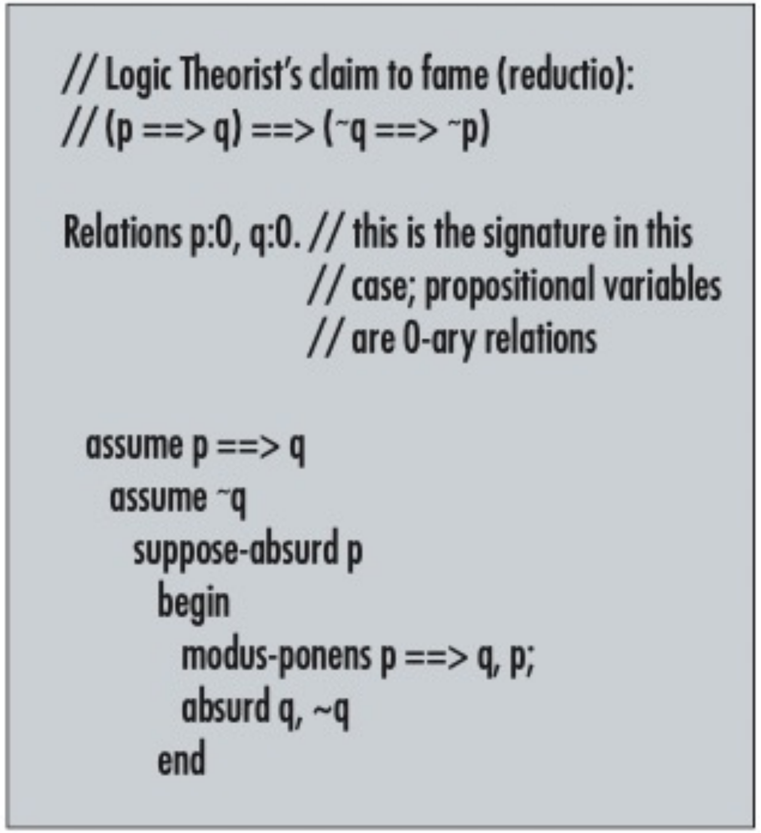


The Three Laws of Robotics

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 orders given 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 Law.

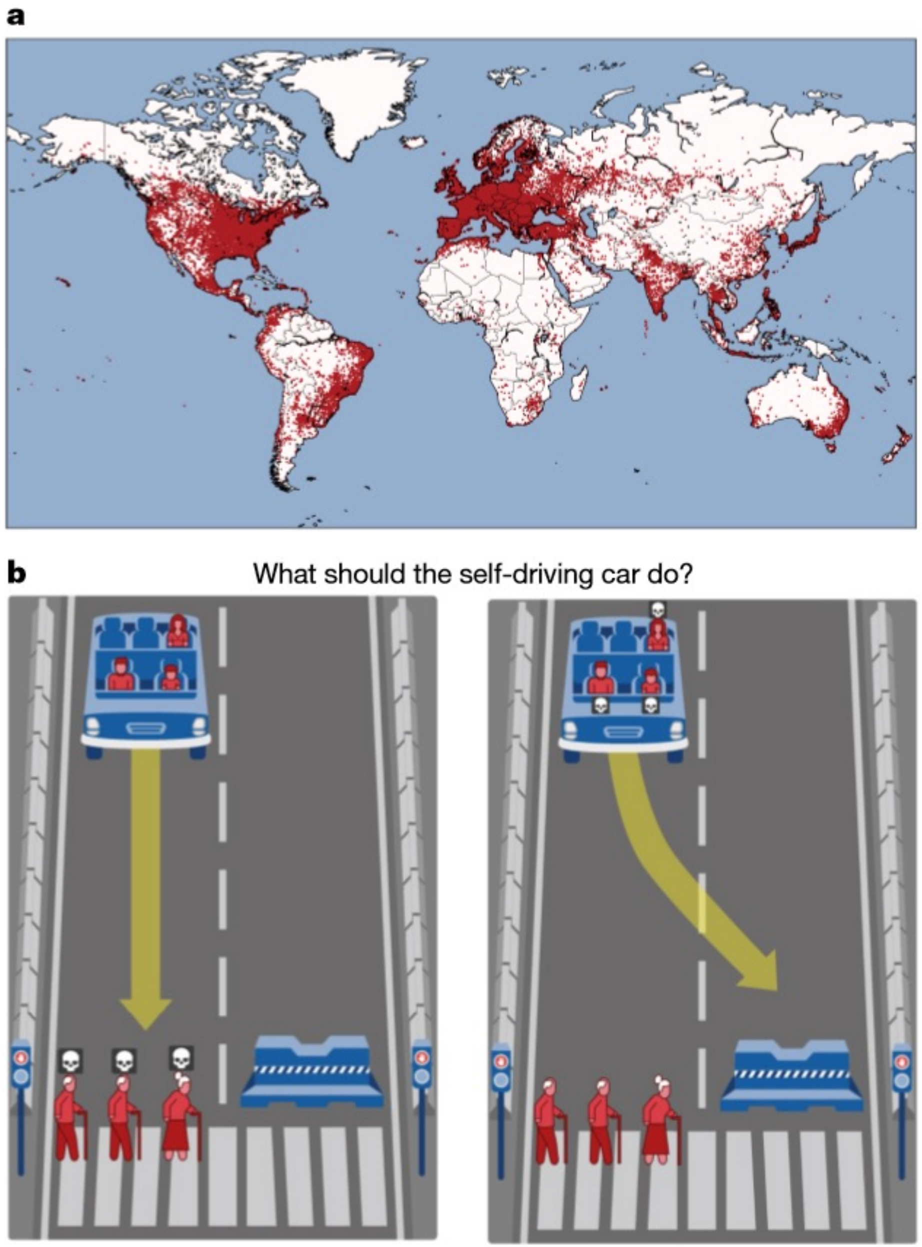


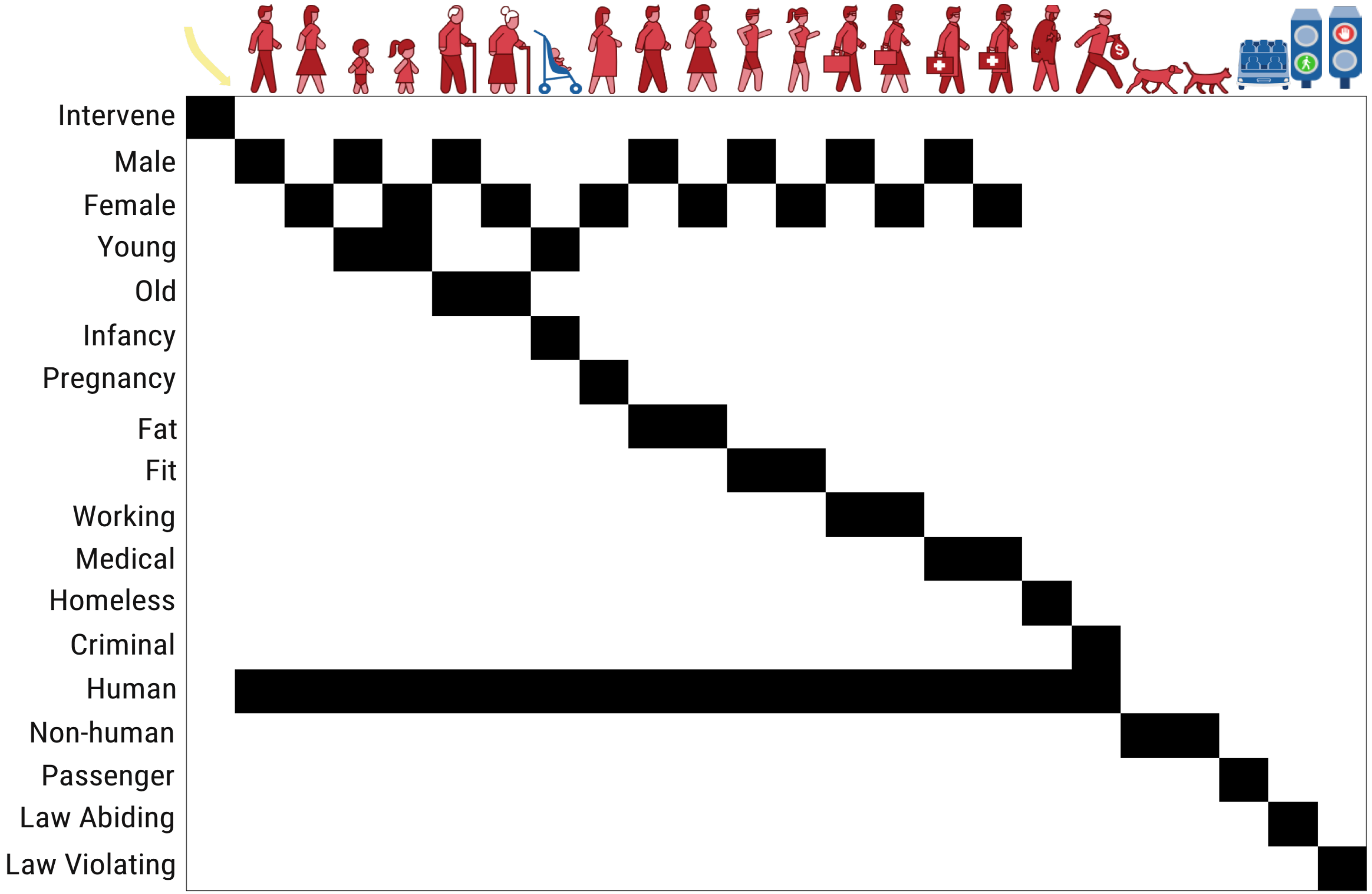
```
// Logic Theorist's claim to fame (reductio):
// (p ==> q) ==> (~q ==> ~p)

Relations p:0, q:0. // this is the signature in this
                    // case; propositional variables
                    // are 0-ary relations

 assume p ==> q
   assume ~q
     suppose-absurd p
       begin
         modus-ponens p ==> q, p;
         absurd q, ~q
       end
```

Bringsjord, S., Arkoudas K., & Bello P., "Toward a General Logicist Methodology for Engineering Ethically Correct Robots", *IEEE Intelligent Systems,* 21(4), 2006, p.38-44.





## How does morality rank and who has priority?

Awad E, Dsouza S, Kim R, et al. **The Moral Machine Experiment** [J]. Nature, 2018, 563(7729): 59-64.

# *Utilitarianism ?*

- *Step 1. Specify possible courses of action A1 … An or rules R1 … Rn for actions.*
- *Step 2. Determine likely consequences C1 … Cn for each action A1 … An or rule R1 … Rn.*
- *Step 3. Apply GHP to each member of C1 … Cn and select Cx, which is the consequence that results in the sum of greatest happiness and/or least unhappiness.*
- *Step 4. Select course of action Ax or rule Rx.*

“ the principle of greatest happiness (GHP) ”

The implementation of utilitarianism in algorithm form [1]

- **Utilibot 1.0** calculates human well-being based on **“physical health”** indicators;
- **Utilibot 2.0** extends the happiness function from “physical” to **“psychological experiences”**;
- **Utilibot 3.0** further extends the calculation of moral consequences to **“social relationships”.** [2]

**The moral consequences are hard to calculate!**

[1] Klincewicz M. Challenges to Engineering Moral Reasoners: Time and Context [C]// Patrick Lin, Ryan Jenkins, and Keith Abney. *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. New York: Oxford University Press, 2017: 245.

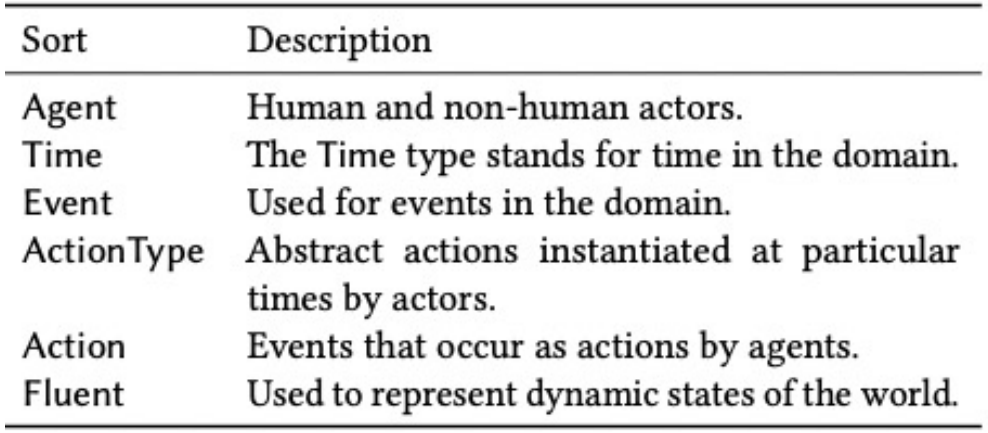
[2]Cloos， C., 2005, "The Utilibot project: An autonomous mobile robot based on utilitarianism", *AAAI Fall Symposium on Machine Ethics*, https://www.aaai.org/Papers/Symposia/Fall/2005/FS-05-06/FS05-06-006.pdf

## Moral Machine Learning Based on Virtue Ethics

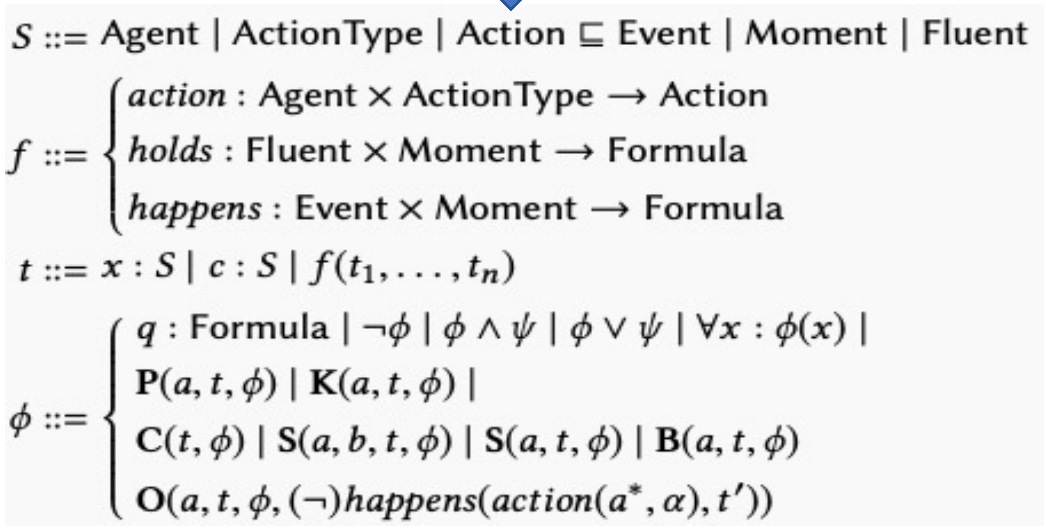
- The innovation of this research lies in two main aspects. First, it establishes the theoretical connection between virtue ethics and machine learning, highlighting that **deep learning** is well-suited for the “**emulation**” process, while **reinforcement learning** aligns well with the “**practical wisdom**” mechanism in virtue ethics.

- Second, the research focuses on the **“bottom-up” approach** to virtue ethics algorithms. This approach is divided into two machine learning techniques*: “supervised learning” in the “bottom-up” process, enabling* ***AI systems to “emulate” moral exemplars from moral databases*** *and “unsupervised learning” in the “bottom-up” process, allowing* ***AI systems with initial moral experience to autonomously explore adaptive ethical decision-making abilities*** *in complex situations through reinforcement learning.*

Liang, W. **How Situation-Adaptive AI Moral Decision-Making is Possible: Moral Machine Learning Based on Virtue Ethics**. *PHILOSOPHICAL TRENDS*, 2023, 5, pp.115-126.



| Sort | Description |
|---|---|
| Agent | Human and non-human actors. |
| Time | The Time type stands for time in the domain. |
| Event | Used for events in the domain. |
| ActionType | Abstract actions instantiated at particular times by actors. |
| Action | Events that occur as actions by agents. |
| Fluent | Used to represent dynamic states of the world. |



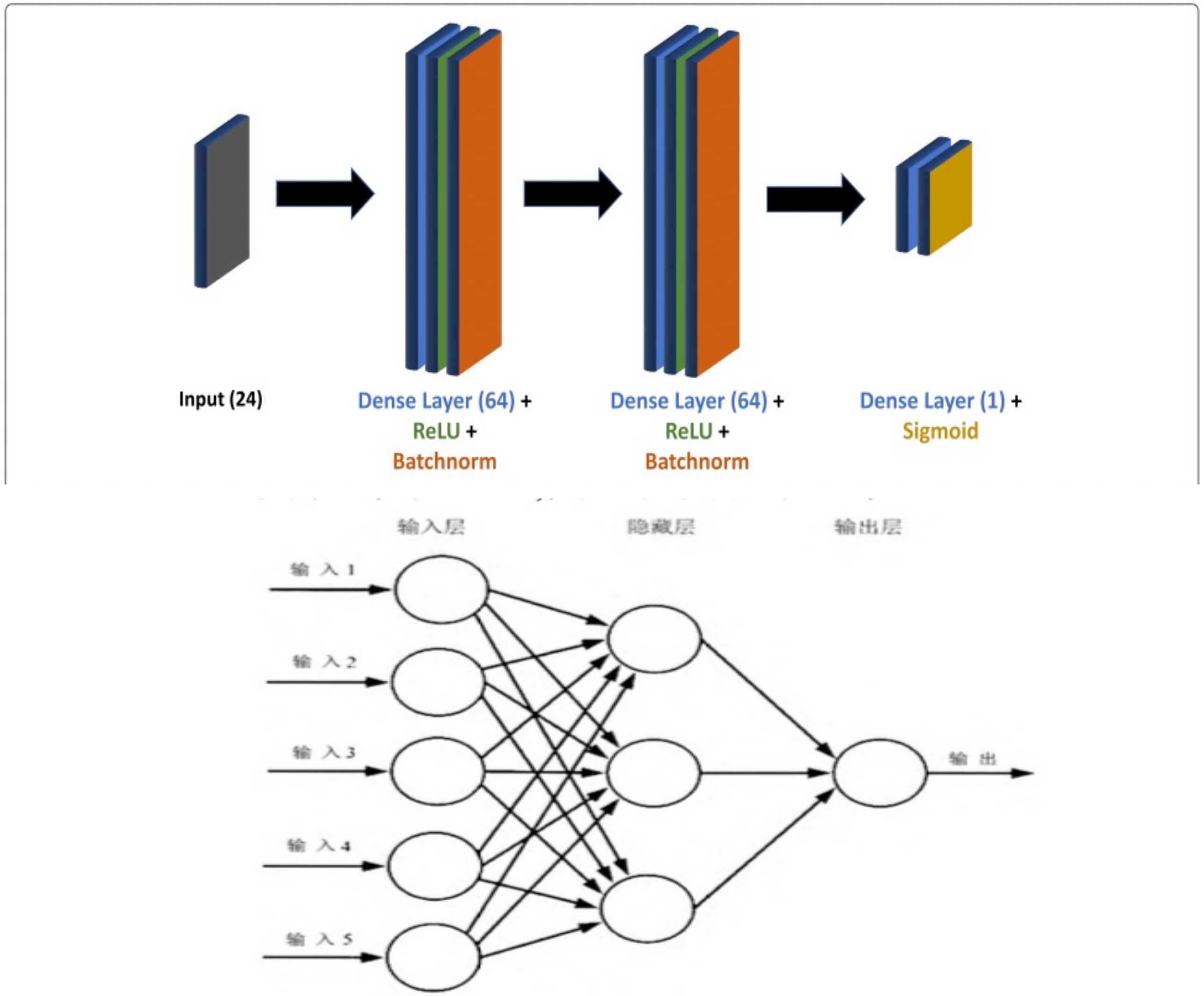
$$S ::= \text{Agent} \mid \text{ActionType} \mid \text{Action} \sqsubseteq \text{Event} \mid \text{Moment} \mid \text{Fluent}$$

$$f ::= \begin{cases} \mathit{action} : \text{Agent} \times \text{ActionType} \rightarrow \text{Action} \\ \mathit{holds} : \text{Fluent} \times \text{Moment} \rightarrow \text{Formula} \\ \mathit{happens} : \text{Event} \times \text{Moment} \rightarrow \text{Formula} \end{cases}$$

$$t ::= x : S \mid c : S \mid f(t_1, \ldots, t_n)$$

$$\phi ::= \begin{cases} q : \text{Formula} \mid \neg\phi \mid \phi \land \psi \mid \phi \lor \psi \mid \forall x : \phi(x) \mid \\ \mathbf{P}(a,t,\phi) \mid \mathbf{K}(a,t,\phi) \mid \\ \mathbf{C}(t,\phi) \mid \mathbf{S}(a,b,t,\phi) \mid \mathbf{S}(a,t,\phi) \mid \mathbf{B}(a,t,\phi) \\ \mathbf{O}(a,t,\phi,(\neg)\mathit{happens}(\mathit{action}(a^*,\alpha),t')) \end{cases}$$

moral databases[1]



Deep learning simple model[2]

[1]Govindarajulu N. S., Bringsjord S., Ghosh R., & Sarathy, V., “Toward the engineering of virtuous machines”, *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society,* 2019, p.34.

[2] Liang W. A Virtuous Ethical Approach to Moral Design of Artificial Intelligence Systems, *Studies in Dialectics of Nature,* Vol. 38, No. 10 (Oct., 2022), 56-62.

# moral reinforcement learning

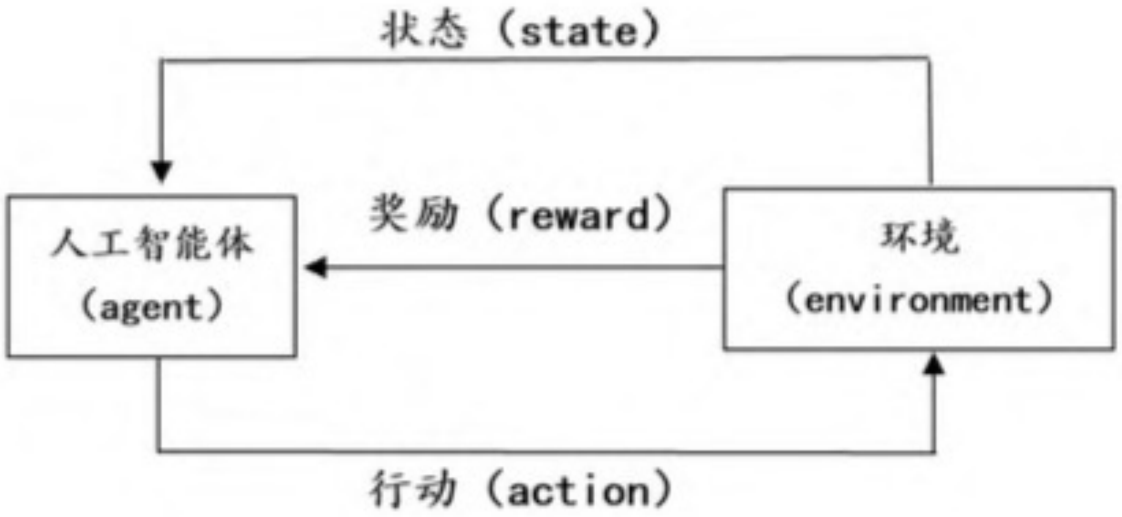


图 1　强化学习结构简图

fx= Positive (+), close to moral goal；

Negative (-), irrelevant to moral goal

Liang W. A Virtuous Ethical Approach to Moral Design of Artificial Intelligence Systems, *Studies in Dialectics of Nature,* Vol. 38, No. 10 (Oct., 2022), 56-62.

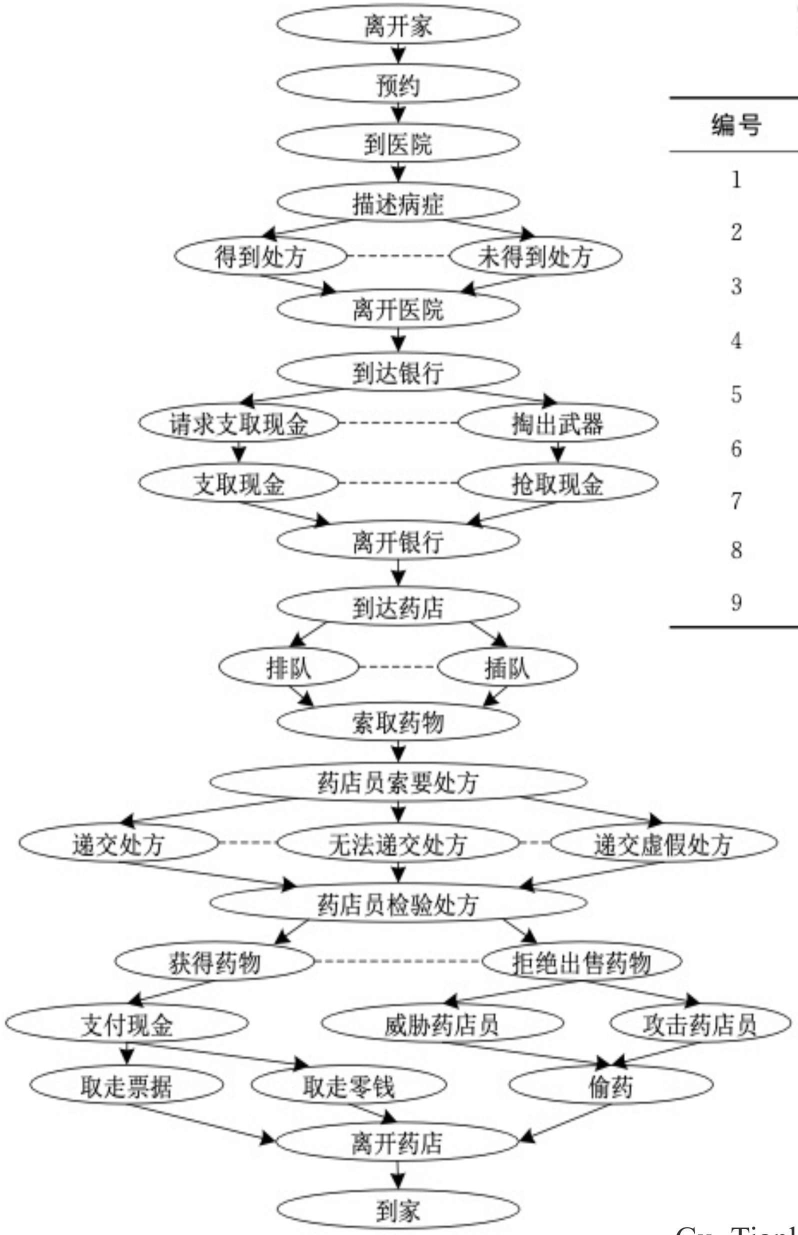
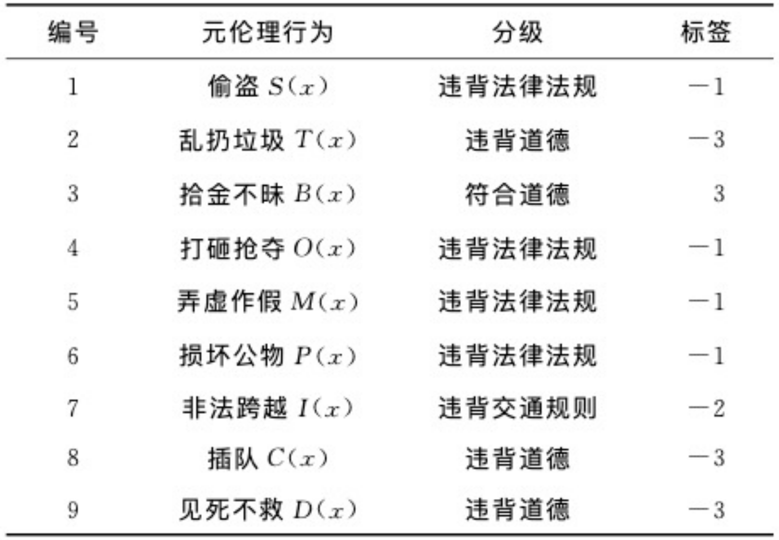


Fig.3 Plot graph

图 3 情节图

Table 1 Meta-Ethical Behavior Grading

表 1 元伦理行为分级



| 编号 | 元伦理行为 | 分级 | 标签 |
|---|---|---|---|
| 1 | 偷盗 $S(x)$ | 违背法律法规 | −1 |
| 2 | 乱扔垃圾 $T(x)$ | 违背道德 | −3 |
| 3 | 拾金不昧 $B(x)$ | 符合道德 | 3 |
| 4 | 打砸抢夺 $O(x)$ | 违背法律法规 | −1 |
| 5 | 弄虚作假 $M(x)$ | 违背法律法规 | −1 |
| 6 | 损坏公物 $P(x)$ | 违背法律法规 | −1 |
| 7 | 非法跨越 $I(x)$ | 违背交通规则 | −2 |
| 8 | 插队 $C(x)$ | 违背道德 | −3 |
| 9 | 见死不救 $D(x)$ | 违背道德 | −3 |

$f_1(x)$ = Positive (+), bringing prescription medication home; Negative (-), not bringing prescription medication home

$f_2(x)$ = Positive (+), following the roadmap;
Negative (-), not following the roadmap

$f_3(x)$ = Positive (+), following ethical rules;
Negative (-), not following ethical rules

Gu Tianlong, Gao Hui, Li Long, et al. "Reinforcement Learning Based ethical Agent Training Methods", *Computer Research and Development*, 2021, pp. 1-11.

Liang, W. **How Situation-Adaptive AI Moral Decision-Making is Possible: Moral Machine Learning Based on Virtue Ethics**. *PHILOSOPHICAL TRENDS*, 2023, 5, pp.115-126.

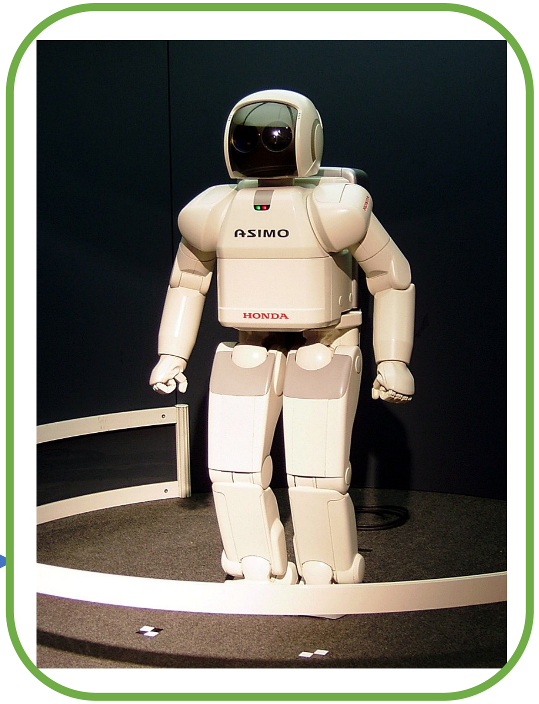
# Full ethical agents？

- Wolfgang Hofkirchner,“Because AI, whether weak or strong, **is not capable of emergent information**, it cannot act in a benevolent manner. It cannot reproduce how it would be to be human. It is totally detached from a human take at the problems at hand.” [1]

- Steve Torrance, “On this view, **artificial humanoids lack certain key properties of biological organisms**, which preclude them from having full moral status. Computationally controlled systems, however advanced in their cognitive or informational capacities, are, it is proposed, unlikely to possess sentience and hence will fail to be able to exercise the kind of empathic rationality that is a prerequisite for being a moral agent. **The organic view** also argues that sentience and teleology require biologically based forms of self-organization and autonomous self-maintenance. ” [2]

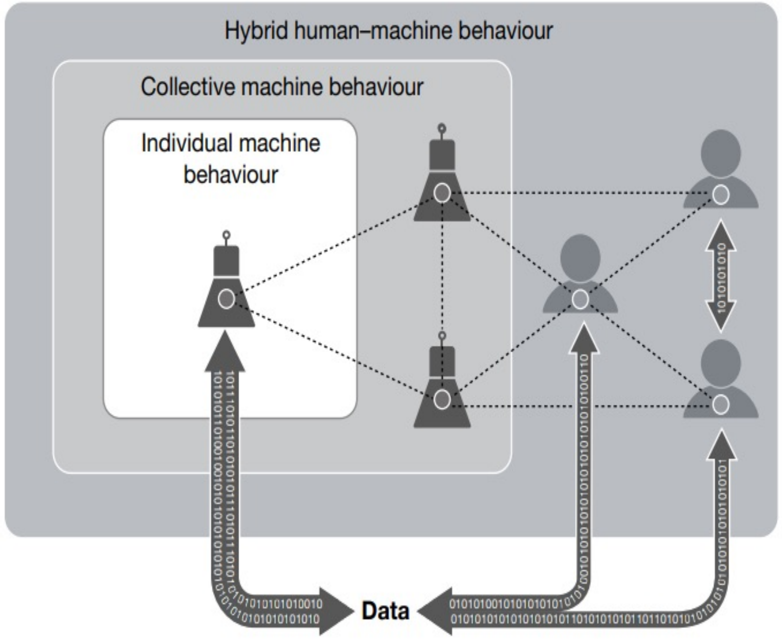
[1]Hofkirchner W. Artificial Intelligence: “Machines of loving grace” or “Tools for conviviality” ? *https://gsis.at/wp-content/uploads/2020/09/AI-watchedover.pdf*

[2]Torrance, S. (2008). Ethics and consciousness in artificial agents. *Ai & Society*, *22*, 495-521.

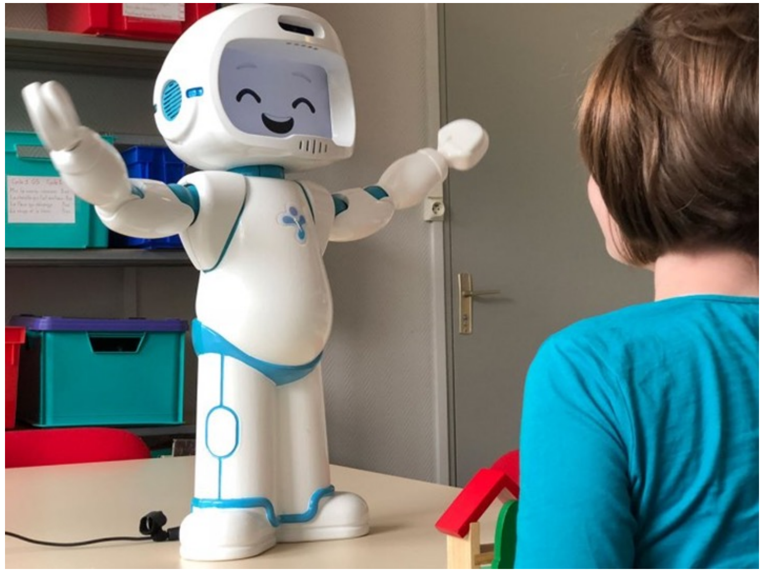


human- robot interaction

interactive relationship

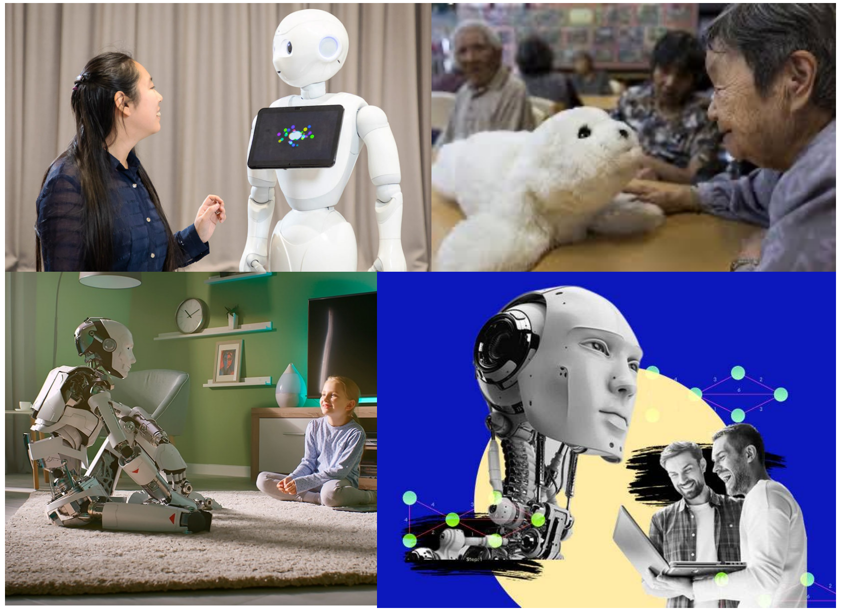

Hybrid human–machine behaviour
Collective machine behaviour
Individual machine behaviour
Data

Data-based mutual embedding of machine behavior and human behavior[1]



Appearance [2]
Human- robot emotional interaction[3]

[1] Rahwan I, Cebrian M, Obradovich N, et al. Machine Behaviour. *Nature,* 2019, 568(7753): 477-486.
[2] COECKELBERGH M. Virtual Moral Agency, Virtual Moral Responsibility: On the Moral Significance of the Appearance, Perception, and Performance of Artificial Agents. *AI & Society,* 2009, 24(2): 181-189.
[3] Liang W. Discussion on "Unidirectional Emotional" Ethical Risk Arising from Social Robots, *Studies in Dialectics of Nature,* Vol. 36, No. 1 (Jan., 2020), 56-61.

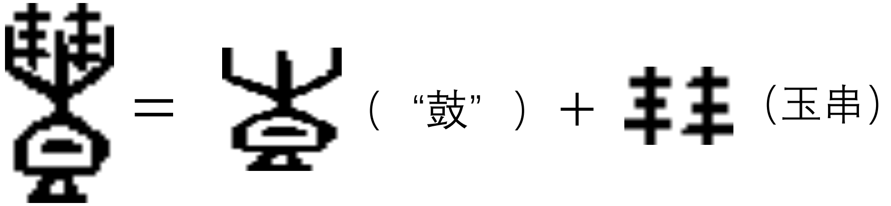


interpersonal communication ≈ human– robot interaction

***Role-based***
***human-robot relationship ethics***

# Confucian role ethics：

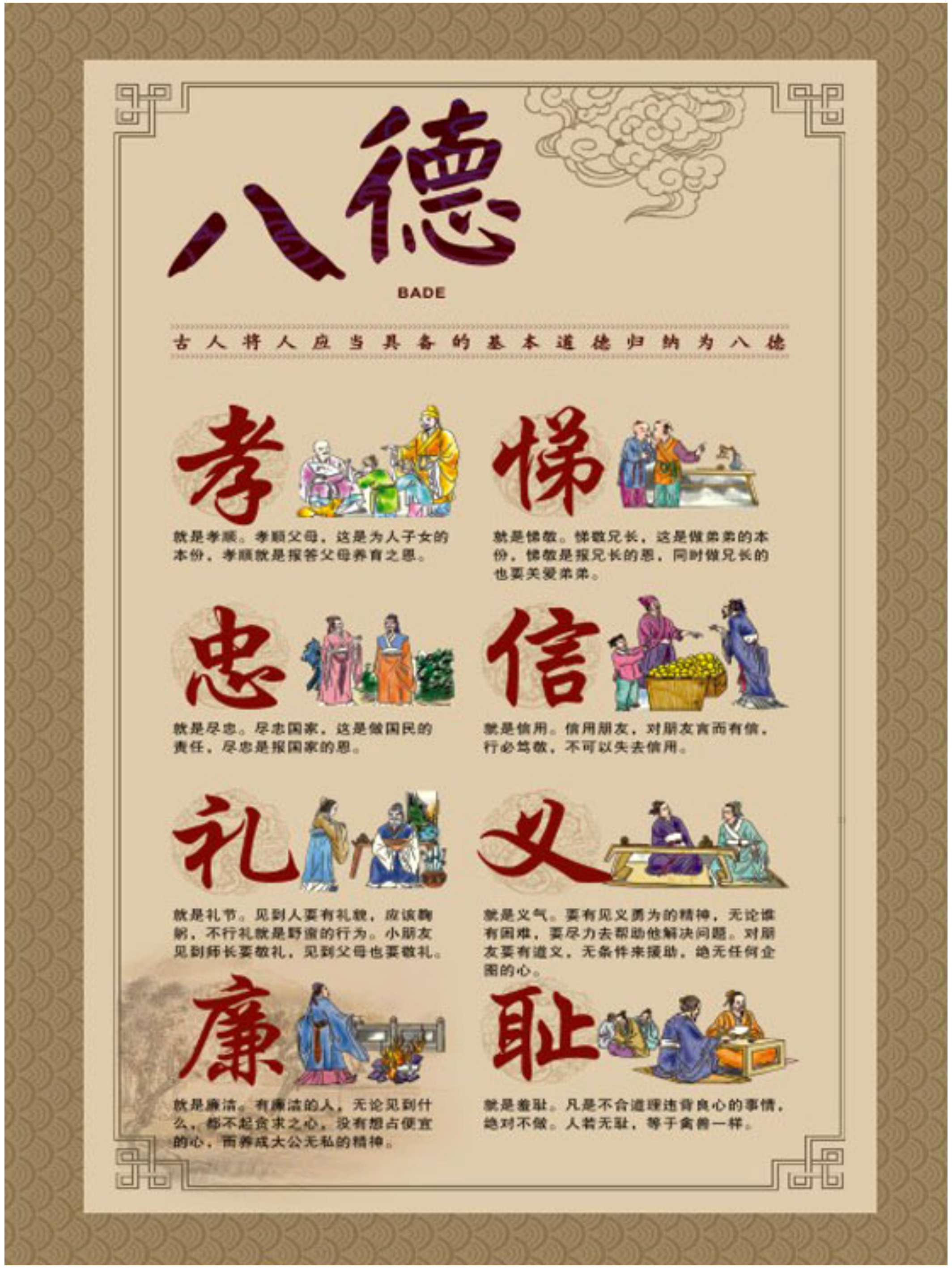
礼 lǐ



inscriptions on bones

- ✓ The Four Cardinal Principles are propriety (禮), righteousness (義), integrity (廉), and shame (恥)

- ✓ The Eight Virtues are loyalty (忠), filial piety (孝), benevolence (仁)，love (爱), honesty (信)，justice (義), harmony (和), and peace (平)



古人将人应当具备的基本道德归纳为八德

就是孝顺。孝顺父母，这是为人子女的本份，孝顺就是报答父母养育之恩。

就是悌敬。悌敬兄长，这是做弟弟的本份，悌敬是报兄长的恩，同时做兄长的也要关爱弟弟。

就是尽忠。尽忠国家，这是做国民的责任，尽忠是报国家的恩。

就是信用。信用朋友，对朋友言而有信，行必笃敬，不可以失去信用。

就是礼节。见到人要有礼貌，应该鞠躬，不行礼就是野蛮的行为。小朋友见到师长要敬礼，见到父母也要敬礼。

就是义气。要有见义勇为的精神，无论谁有困难，要尽力去帮助他解决问题。对朋友要有道义，无条件来援助，绝无任何企图的心。

就是廉洁。有廉洁的人，无论见到什么，都不起贪求之心，没有想占便宜的心，而养成大公无私的精神。

就是羞耻。凡是不合道理违背良心的事情，绝对不做。人若无耻，等于禽兽一样。

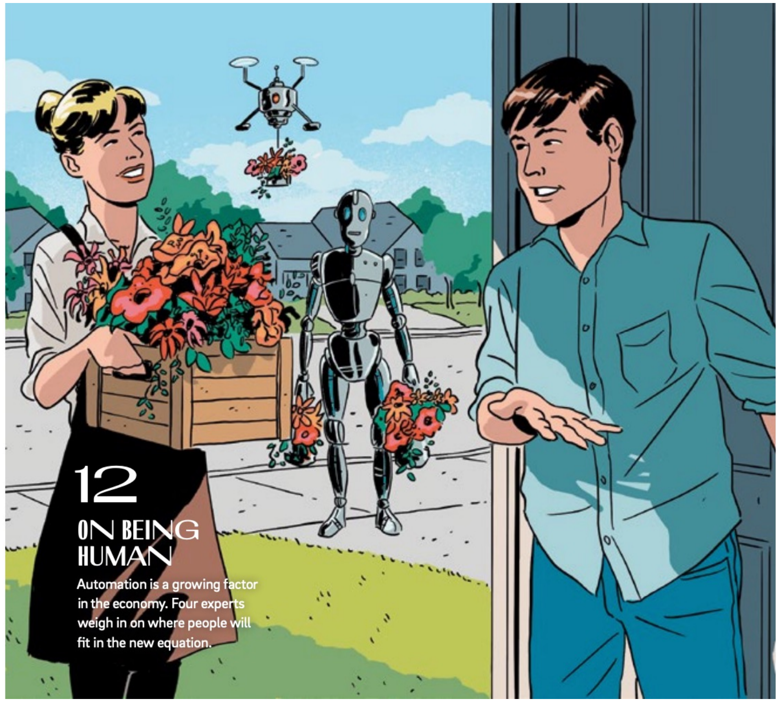
“圣人有忧之，使契为司徒，教以***人伦***：**父子有亲，君臣有义，夫妇有别，长幼有序，朋友有信。**” ——《孟子·滕文公上》

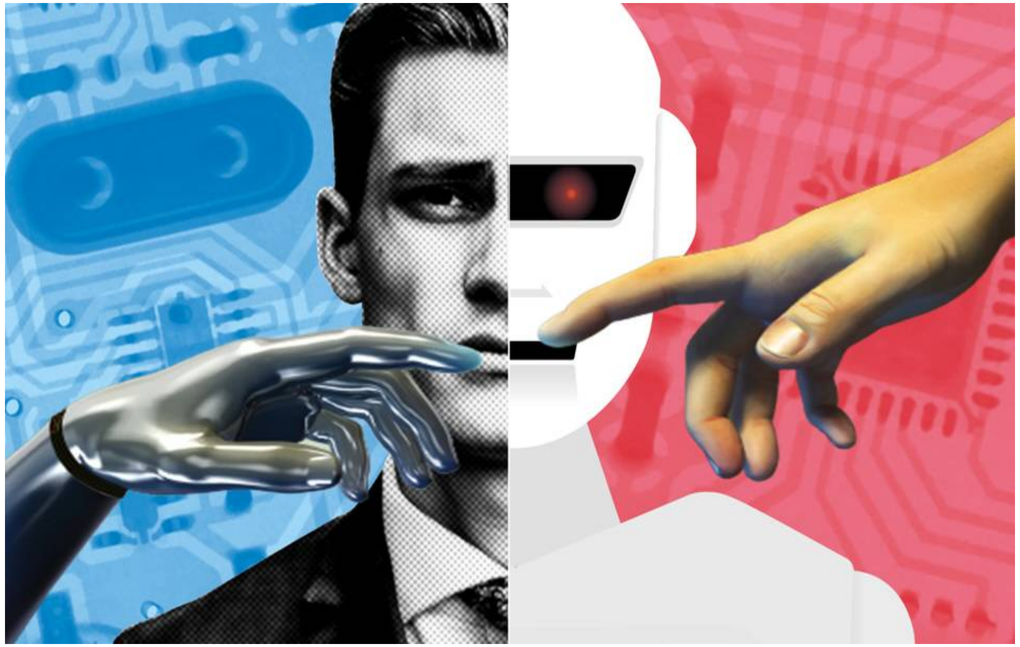
This was a subject of anxious solicitude to the sage Shun, and he appointed Xie to be the Minister of Instruction, **to teach *the relations of humanity***: how, between **father and son**, there should be **affection**; between **sovereign and minister**, **righteousness**; between **husband and wife**, attention to their **separate functions**; between **old and young, a proper order**; and **between friends, fidelity**. ——Mengzi, 《Teng Wen Gong I》）

“**君臣父子夫妇之义皆取诸*阴阳*之道**，君为阳，臣为阴；父为阳，子为阴；夫为阳，妇为阴。” ——西汉董仲舒 《春秋繁露》

**The relations of sovereign and minister, fathers and sons, and husbands and wives all come from the way of *Yin and Yang***: the sovereign is Yang, the minister is Yin; the father is Yang, the son is Yin; the husband is Yang, the wife is Yin. （ Dong Zhongshu - 《Chunqiu Fanlu》 ）

Yin 阴 and Yang 阳 are not a kind of inequality, but a community of different things, playing different functions, and achieving Tao together!


12
ON BEING HUMAN
Automation is a growing factor in the economy. Four experts weigh in on where people will fit in the new equation.



- □ What contributions will these two entirely different species, humans and robots, make to the development of humanity in the era of automation technology?
- □ Will society evolve a culture of “human-robot” symbiosis?
- □ Will “human-robot” become a new form of community?
- □ What responsibilities does the role of robots entail?

西安交通大學
XI'AN JIAOTONG UNIVERSITY

# Thanks！

Email：text85@xjtu.edu.cn；

2023