

自律的知的機械の設計に 倫理学を応用する

EADワークショップ
2018年3月24日

久木田水生
名古屋大学大学院情報学研究科
minao.kukita@i.nagoya-u.ac.jp

“Ethically Aligned Design”の“Ethically” とは？

- “Ethics”は、「倫理」と「倫理学」の両方を意味する。前者は社会において守るべきとされている規範や道義や価値観を指す。後者は前者について研究をする学問分野を指す。
- “Ethically Aligned Design”では全般に前者の意味で使われていると思うが、“Classical Ethics in A/IS”の章では特に学問分野としての倫理学が取り上げられている。

倫理学

- 倫理学は大雑把にいつて何が「善い」ことで、何が「悪い」ことか、何を「すべき」か、何を「すべきでない」か、といったことを考える分野である。
- 西洋において2000年を超える長い倫理学の伝統があるので、これを踏まえて、A/ISの設計に応用しよう。（EAD2）
- でも、もちろん、そういうことを考えたのは西洋人だけではないので、西洋人以外の考えも踏まえよう。多様性は大事だ。（EAD2）

西洋の倫理学（EAD2によれば）

- 西洋の哲学的倫理学は科学的方法（例えば論理的・論証的・弁証法的 logical, discursive and dialectical アプローチや分析的・解釈学的 analytical and hermeneutical アプローチ）を使う。
- あらかじめ前提された道徳 mores を教えることには興味を持たない。
- これに対して宗教はすでに決まった教条を疑うことなく、守り伝えることに興味を持つ。

合理性・論理を重視

Classical EthicsとModern Ethics (EAD2によれば)

- 古代ギリシャでは、倫理は個人、家庭、ポリスという経済の三つの次元との関わりで考えられた。
- 近代になり、倫理は個人の次元の問題として扱われるようになった。Cf. カントの倫理学。
- 「機械倫理」などのプロジェクトは近代の、集団から切り離された個人としての倫理に焦点を当てている。

西洋の倫理学の主流の立場

- 義務論 deontology : 規範に従うのが重要。Cf. カント。
- 帰結主義 consequentialism : 良い帰結を生み出すのが重要。
Cf. ベンサム、ミル。
- 徳倫理学 virtue ethics : 良い性格を涵養するのが重要。Cf. アリストテレス。

西洋の倫理によるモノポリー？（EAD2によれば）

- リベラルで民主主義的な価値観が多文化間の情報倫理を席卷している。
- 多様な価値を取り入れることが必要だ。とはいえ多様性を受け入れるというのもリベラルな価値観。

非西洋の伝統的な倫理（EAD2によれば）

- 仏教：解脱が目標。他者との「適切な関係性」で苦しみを軽減。
- ウブントゥ：コミュニティ中心。
- 神道：人形や自動機械にも神が宿る。人工物も川などのように自然の一部。
- こういった倫理をA/ISの設計に生かせる／生かすべき。

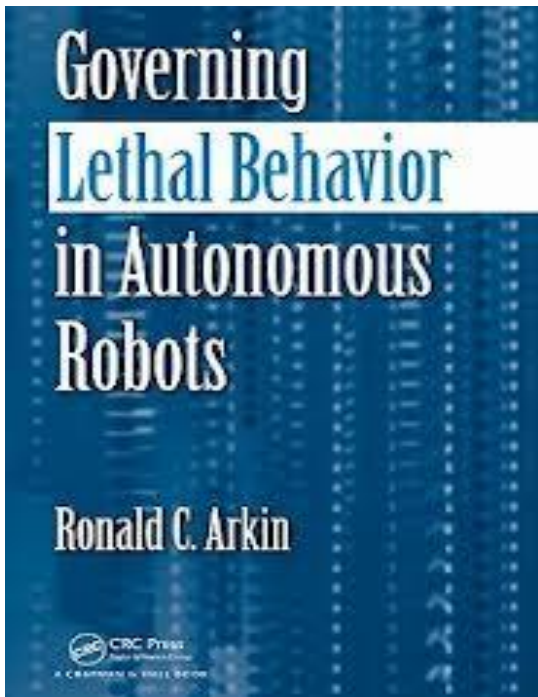
なぜこの三つなのか？

EADと倫理学

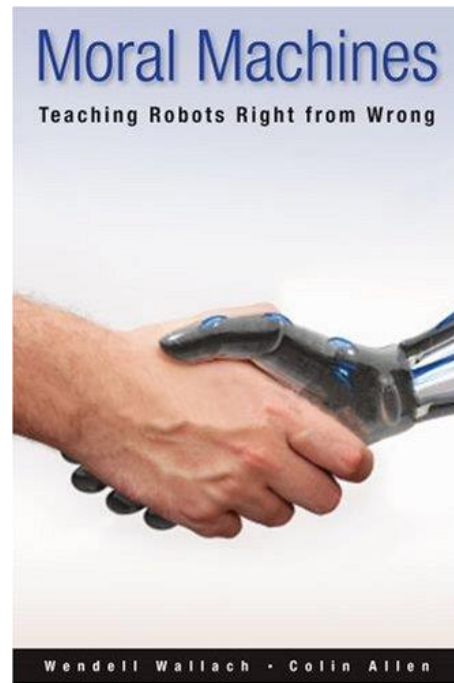
- A/ISは人間の自律性にどのような影響を与えるかを考えなければいけない。
- A/ISを徳倫理学に沿って設計することが有益かもしれない。
- A/ISに義務論的な倫理学理論を形式的手法で実装することには利点がある。

倫理的な機械

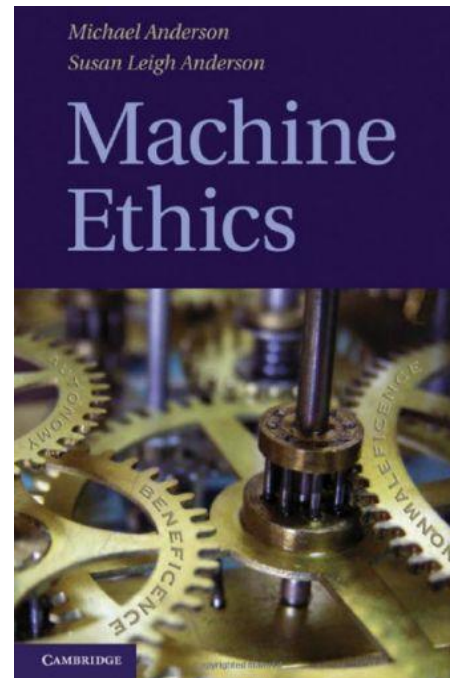
道徳的判断を機械に実装する取り組み



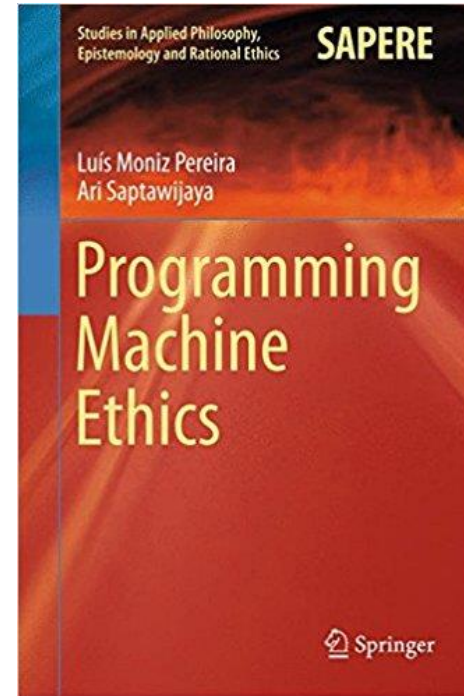
2009



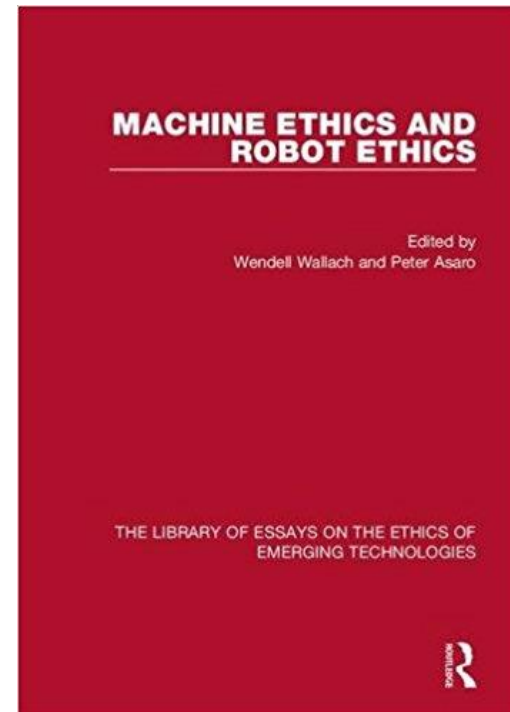
2010



2011



2016



2016

タフツ大学, ブラウン大学, レンセリア・ポリテクニク研究所の研究者たちが米海軍とチームを組んで, 善悪とそれらの帰結を理解することのできる「道徳的」な自律ロボットの開発に着手した

道徳的能力とは大雑把に言って人間が同意する傾向にある法律や社会的な規約を学び、それについて推論し、それに基づいて行動し、それについて語るすることができる能力と考えることができる

戦場における
アルゴリズムに従った道徳

News

Home | News | Blog | Essays | Contributors | Kurzweil Collection

  Feeds |  Newsletter |  Books |  Multimedia |  Events |  Forums |  Press

You are here: Home > News > Can robots be trusted to know right from wrong?

Can robots be trusted to know right from wrong?

May 12, 2014

Is it possible to develop "moral" autonomous robots with a sense for right, wrong, and the consequences of both?

Researchers from [Tufts University](#), [Brown University](#), and [Rensselaer Polytechnic Institute](#) think so, and are teaming with the U.S. Navy to explore technology that would pave the way to do exactly that.

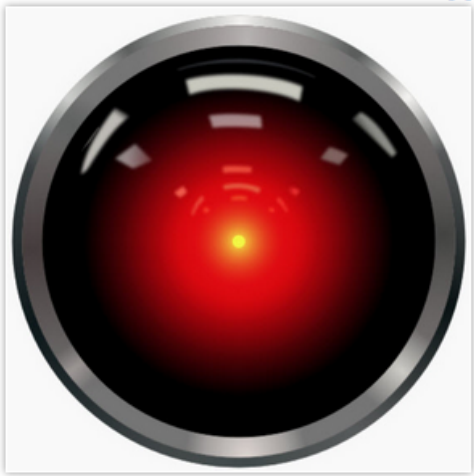
"Moral competence can be roughly thought about as the ability to learn, reason with, act upon, and talk about the laws and societal conventions on which humans tend to agree," says principal investigator Matthias Scheutz, professor of computer science at Tufts School of Engineering and director of the Human-Robot Interaction Laboratory (HRI Lab) at Tufts.

"The question is whether machines — or any other artificial system, for that matter — can emulate and exercise these abilities."

But since there's no universal agreement on the morality of laws and societal conventions, this raises some interesting questions. Was [HAL 9000](#) (HAL = (Heuristically programmed ALgorithmic computer) moral? Who defines morality?

Algorithmic morals on the battlefield

Scheutz cites a simplified military battlefield scenario, where the conventions are more easily defined (but still improve on Asimov's simplistic three laws of robotics): a robot medic responsible for helping wounded soldiers is ordered to



HAL 9000 (credit: Warner Bros.)

道徳性

```
graph TD; A([道徳性]) --> B([他者の感情を思いやる]); B --- C([社会の規範を学ぶ]); B --- D([価値を評価する]); B --- E([反省する]); B --- F([結果の責任を取る]); B --- G([状況に応じて適切な判断を下し行動をとる]);
```

社会の規範を学ぶ

価値を評価する

他者の感情を
思いやる

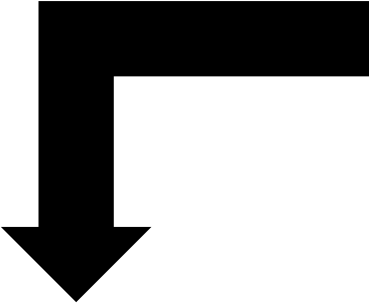
状況に応じて
適切な判断を下し
行動をとる

反省する

結果の責任を取る

等々...

Machine Ethicsのパラダイム



機械が従うべき道德的規範を
明示化する.

機械にその規範を組み込み,
従わせる.

アシモフのロボット工学三原則

- 一、ロボットは人間に危害を加えてはならない。また何も手を下さずに人間が危害を受けるのを黙視していてはならない。
- 二、ロボットは人間の命令に従わなくてはならない。ただし第一原則に反する命令はその限りではない。
- 三、ロボットは自らの存在を護らなくてはならない。ただしそれは第一、第二原則に違反しない場合に限る。

Ethical trap: robot paralysed by choice of who to save

Can a robot learn right from wrong? Attempts to imbue robots, self-driving cars and military machines with a sense of ethics reveal just how hard this is

Ethical robots save humans

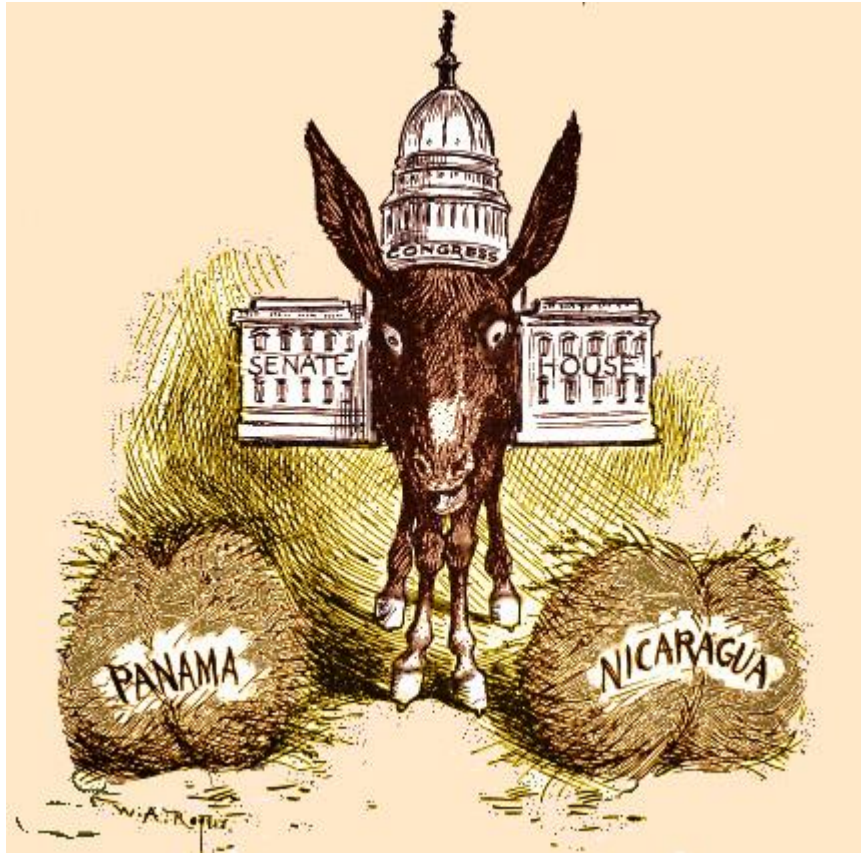


ロボットは善悪の区別をつけられるようになるか？ ロボット、自走車、軍事機械に倫理観を備えさせる試みはそれがいかに難しいかを明らかにする

アラン・ウィンフィールドによる実験では、規則ベースの「人助け」ロボットが、ジレンマにおちいり「パニック」に陥るような様子が見られた。

New Scientist

<https://www.newscientist.com/article/mg22329863-700-ethical-trap-robot-paralysed-by-choice-of-who-to-save/>



1900年に書かれた政治的風刺画。大西洋と太平洋を結ぶ運河をパナマとニカラグアのどちらのルートで建設するか思案する米国議会の様子をビュリダンのロバになぞらえて表現している。

ビュリダンのロバ

- ロバが二つの乾草の山の間でどちらを食べるか迷っているうちに餓死するという寓話.
- 同じくらい望ましい帰結を生む二つの行動の間では合理的な意思決定はできないというビュリダン（中世の哲学者）の説を風刺して作られた.

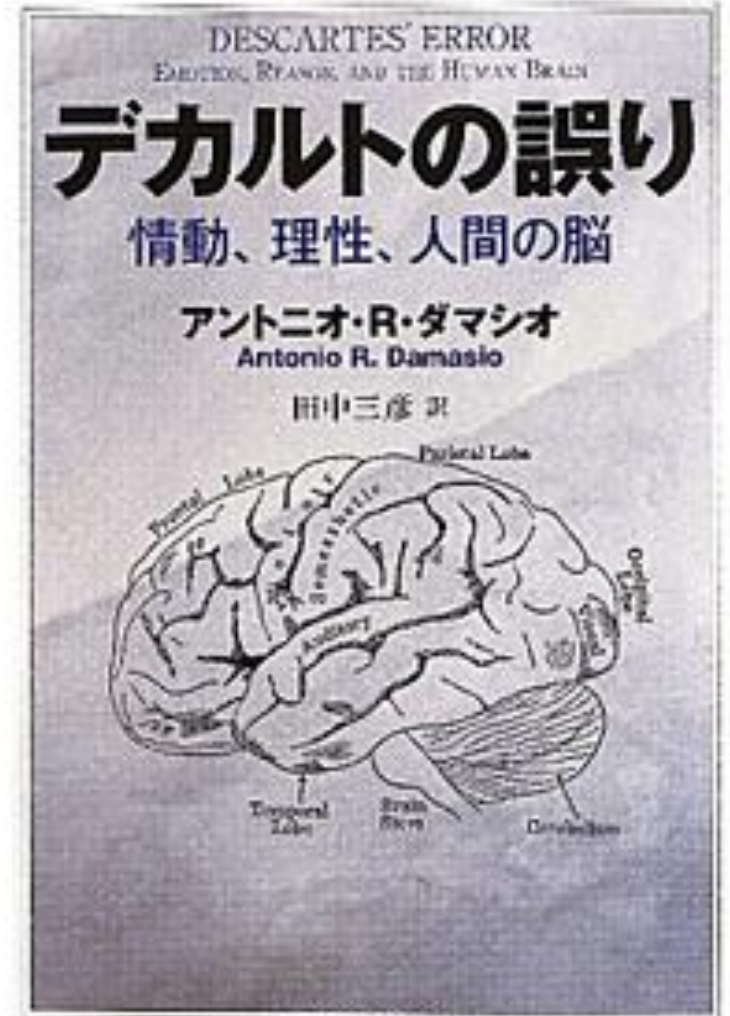
"Deliberations of Congress" by W. A. Rogers - New York Herald (Credit: The Granger Collection, NY).

Licensed under Public Domain via Commons

https://commons.wikimedia.org/wiki/File:Deliberations_of_Congress.jpg#/media/File:Deliberations_of_Congress.jpg

意思決定における 感情の重要性

- 脳の一部を損傷した患者が日常的な意思決定に支障をきたすようになった。
- 調べると、その患者は感情を持たなくなっていたことが分かった。



道徳基盤理論

西洋の倫理学が重視するのは特にこの二つ



ケア／危害
公平／ズル
自由／抑圧
忠誠／裏切り
権威／転覆
聖性／墮落

西洋の伝統的な倫理は“WEIRD”な人たちにしか通じない。

- Western
- Educated
- Industrialized
- Rich
- Democratic

道徳的意思決定の 2つのプロセス

論理

- 手間と時間がかかる
- 柔軟性がある
- より多くの関係者の利害を計算にいれることができる
- 損得を計算して利己的になりやすい

感情

- 自動的で素早い
- 融通が利かない
- 集団の中と外の区別に敏感
- 仲間に対しては利他的



ロボットに感情を持たせればいいのか？

http://www.softbank.jp/corp/news/sbnews/project/2014/20140822_01/



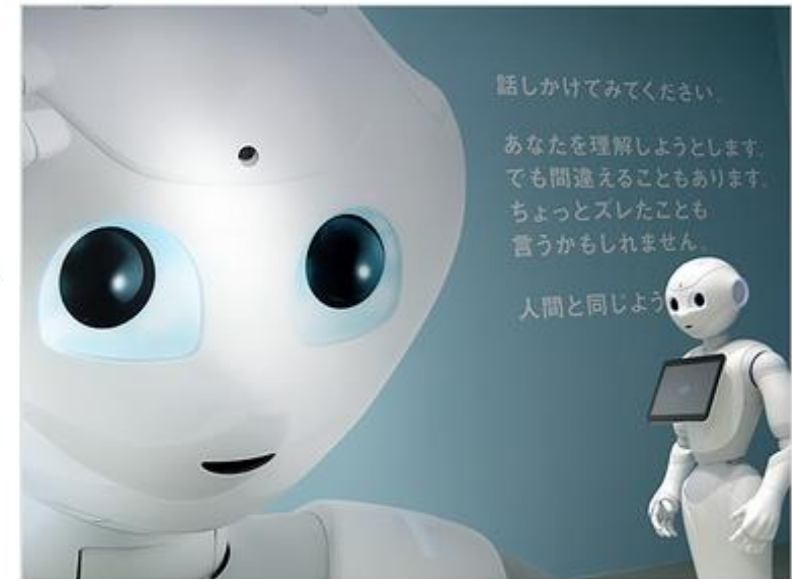
Pepperの「感情エンジン」
生物の内分泌系をシミュレートする

http://biz-journal.jp/2015/06/post_10499_2.html

ソフトバンクのビジョンを実現する 愛のあるロボット「Pepper」の開発

🔊 このページを音声で聴く

ソフトバンクグループは、2014年6月5日に世界初の感情認識パーソナルロボット「Pepper（ペッパー）」の開発と、2015年2月からの一般販売開始を発表しました。また2014年8月1日には、ロボット事業を専門にする会社として、新たにソフトバンクロボティクス株式会社（以下「ソフトバンクロボティクス」）が設立されました。今回、同社で「Pepper」の開発リーダーを務めるプロダクト本部 PMO室 室長の林 要にインタビューしました。



感情

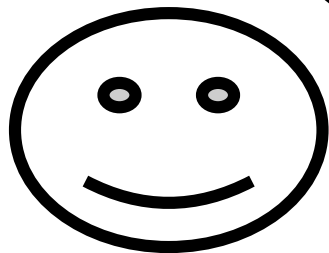
自動的
コントロールしにくい

意思決定・行動

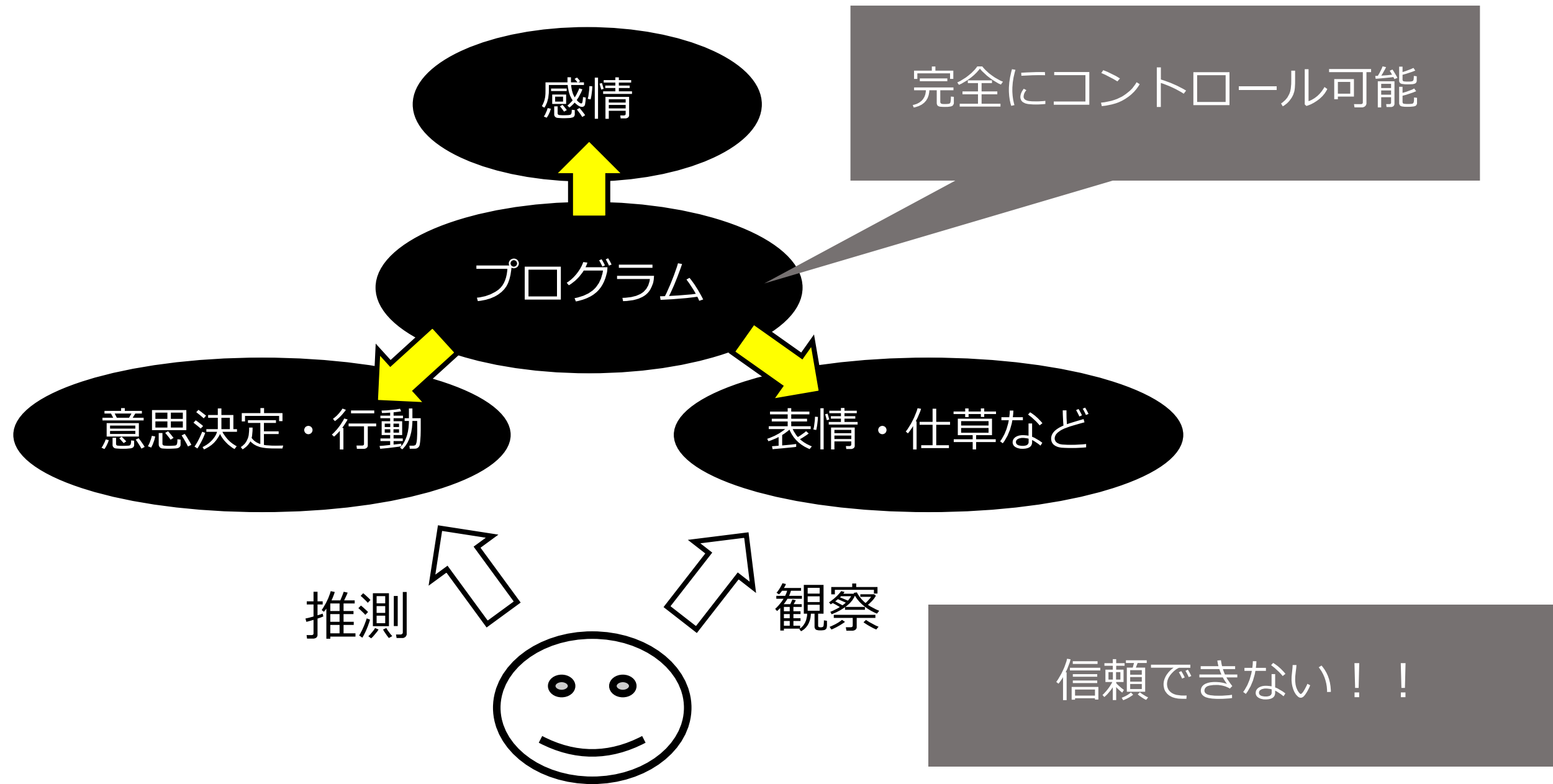
表情・仕草など

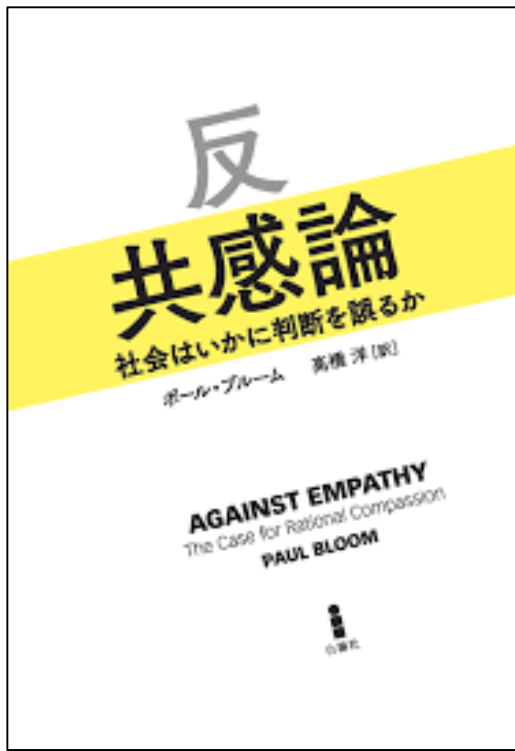
推測

観察



だからこそ信頼できる？！

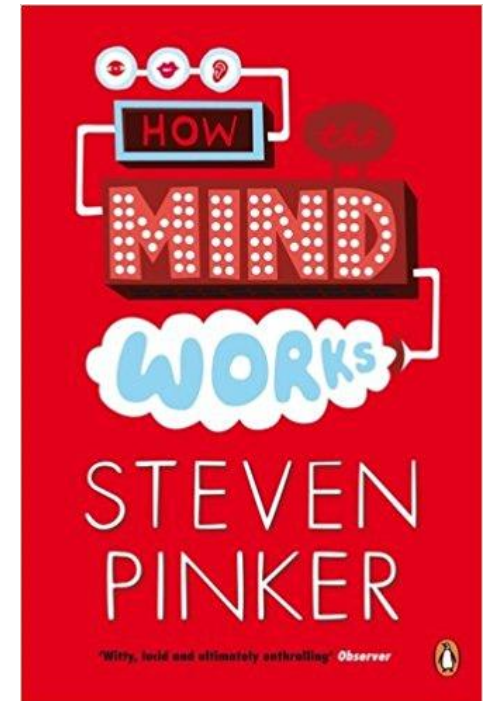
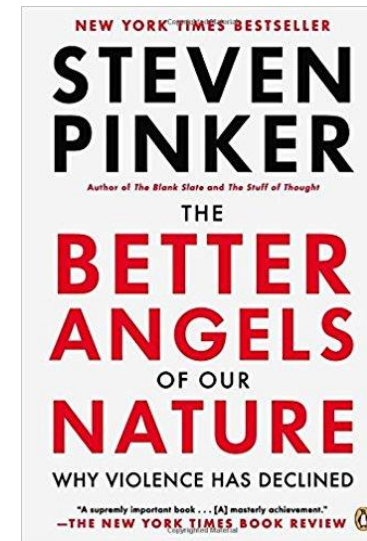


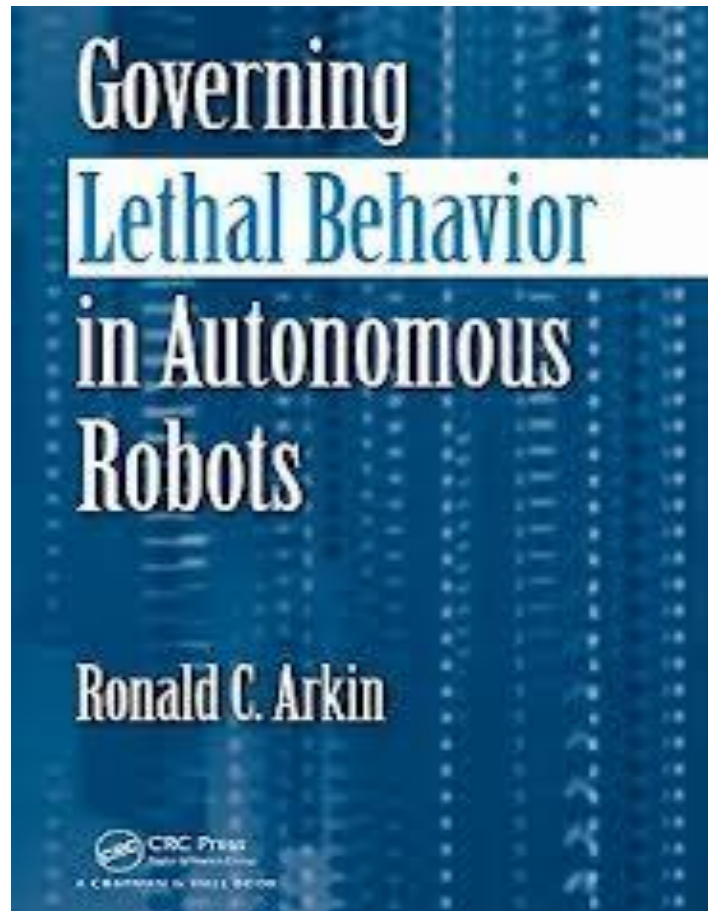


- 共感はスポットライトのようなもので、限定的なスコープしか持たない。
- 合理的・統計的思考を不可能にする。
- バイアスがあり、他者よりも同じ集団に属するものをひいきする。
- 攻撃性を増大させることもある。

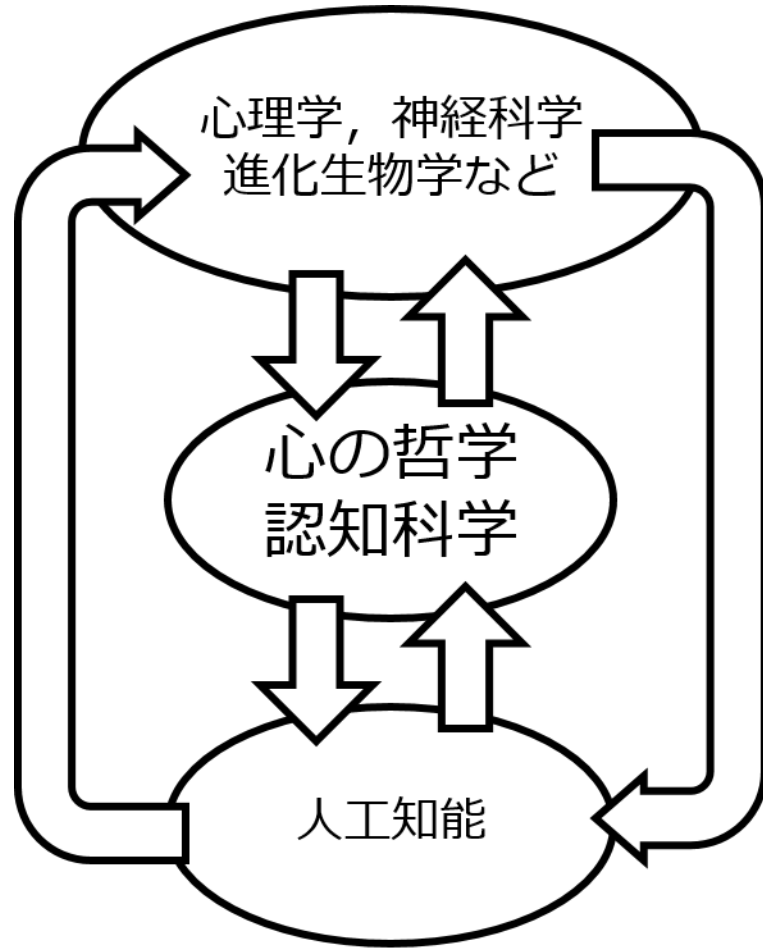
感情は有害？

- 感情は生存と繁殖の機会を増大させるように働く。
- それは幸福や賢明さや道徳的価値を促進するようには機能しない。

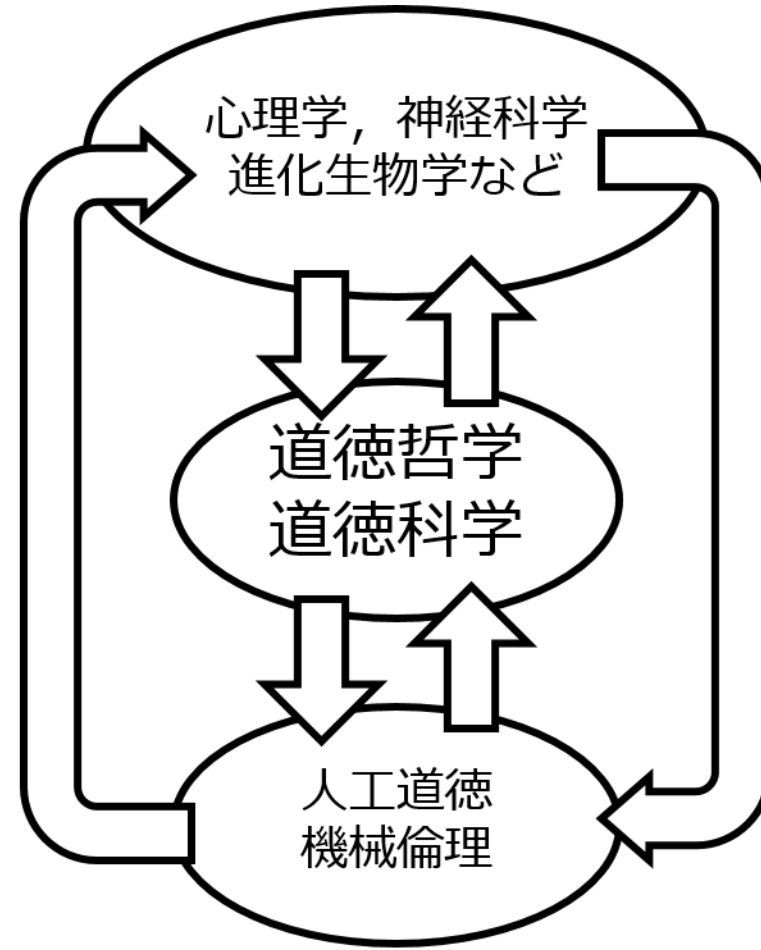




ロボット兵士はそれらは怒りや憎しみや恐怖のような感情にかられることがないから、戦場をより倫理的にする。



20世紀



21世紀