

Artificial Intelligence and the Ecosystem of Trust

Minao Kukita*

1. Introduction

The estimation, prediction, and evaluation of human attributes, abilities, and behavioral tendencies using machine learning systems (automatic profiling) are currently being applied in various domains and have become a key driver of the platform economy. However, such applications of artificial intelligence entail various adverse effects. Machine learning based on big data often reproduces human biases or discriminatory structures present in society, thereby reinforcing or entrenching them (cf. O’Neil (2016)). Business models that depend on data promote excessive data collection and the misuse of data, threatening individuals’ privacy (cf. Véliz (2020); Zuboff (2019)).

In addition to these issues, this paper examines the potential long-term impact of AI-based profiling on social relationships, with a particular focus on the concept of trust. It argues that profiling practices designed to identify and exclude individuals perceived as potential threats often err on the side of caution against false negatives, thereby exhibiting a disproportionate tolerance for false positives. By analyzing concrete examples, the paper explores how such an epistemic stance undermines the possibility of cultivating interpersonal relationships grounded in mutual trust.

2. Israel’s Abuse of AI Profiling against Gaza People

Currently, artificial intelligence is being applied to the military in various ways. In so-called “hybrid warfare,” AI is used to generate and disseminate fake news, intervening in people’s perceptions. In “Lethal Autonomous Weapon Systems” (LAWS), which are believed to be under development by countries such as Russia, the United States, Israel, and China, AI plays a key role in autonomously carrying out actions from identifying targets to executing attacks. However, the focus here is not on such systems, but rather on how AI that performs profiling is applied in military contexts.

Since the September 11 attacks in 2001, the United States has waged a “war on terror,” killing individuals suspected of being terrorists in South Asia, the Middle East, and Africa. In doing so, the U.S. military and CIA adopted a method called “signature strikes” during the process of identifying and attacking targets (Chamayou (2013); Scahill & The Staff of

The Intercept (2017)). This method involves monitoring people in regions where terrorists are believed to be hiding and identifying individuals highly suspected of being terrorists based on patterns in their communication records, behaviors, and social connections. Missiles are then launched at these targets from remotely operated unmanned aerial vehicles. The decision to target these individuals was approved by then-President Obama.

Since anti-tank missiles were dropped from high altitudes on targets hiding in urban areas, many civilians, including numerous children, were inevitably caught in the crossfire. According to U.S. government documents revealed through whistleblowing, all individuals killed in drone strikes were categorized under the label “Enemy Killed In Action (EKIA)” (Scahill & The Staff of *The Intercept* (2017)). For the United States, individuals exhibiting behavioral patterns suspected of being terrorist-related were deemed terrorists, and anyone killed by U.S. attacks was considered an enemy who deserved to die.

By using artificial intelligence, the same actions as signature strikes can be carried out more “efficiently.” According to Israel’s online news outlet +972 *Magazine*, this has been implemented by the Israeli military against residents of the Gaza Strip since the October 7, 2023, surprise attack by Hamas (Abraham (2024)). The Israeli military uses an artificial intelligence system called “Lavender” to identify Hamas or Palestinian Islamic Jihad (PIJ) operatives in the Gaza Strip. Israel has constructed an evaluation model by learning from data on known Hamas and PIJ operatives.

The trained model is applied to data on the 2.3 million residents of the Gaza Strip collected through a large-scale surveillance system, evaluating each individual’s likelihood of being a Hamas or PIJ combatant on a scale from 1 to 100. Residents who exhibit characteristics such as being part of WhatsApp groups with known combatants, changing their phones or their addresses are assigned higher scores. If their score exceeds a certain threshold, they are placed on a “kill list” and become targets for assassination. This system is designed to address the bottleneck posed by human cognitive limitations in the routine of identifying operatives, pinpointing their locations, and carrying out attacks.

The Israeli military explains that decisions are not entirely entrusted to machines, as humans manually verify the process. However, according to sources interviewed by +972, in practice, human verification was limited to confirming whether the target was male or not, and this had to be completed within a matter of seconds. They were under pressure to identify more new targets, and when they were unable to do so, they sometimes lowered the threshold for judgment.

According to the sources, one out of every ten targets assassinated under this system was not actually a Hamas operative. However, the Israeli military reportedly considers this margin of error acceptable. The military randomly selected samples from Lavender's predictions and had them manually verified. Once this manual verification confirmed that the system's predictions were 90% accurate, the military approved its widespread use. In other words, a roughly 10% error margin was factored in from the outset. One source remarked, "the protocol was that even if you don't know for sure that the machine is right, you know that statistically it's fine. So you go for it," adding, "Everything was statistical, everything was neat—it was very dry."

Individuals placed on the "kill list" are subsequently tracked using a system called "Where Is Daddy?" to estimate their location. In many instances, the attacks were scheduled to occur while they were at home, as this was the time when their location could be most reliably determined. As a result, family members, often women or children, are also caught in the attack. There were also instances where the target was not actually at home.

The "statistical" and "dry" approach is consistent in estimating collateral damage (harm to non-combatants or civilian infrastructure caused during military operations). Previously, intelligence personnel took significant time to verify collateral damage, but since October 7, 2023, this verification process has been automated, significantly reducing human involvement. The margin of error in the automated model was tolerated due to the speed of processing. "The collateral damage calculation was completely automatic and statistical," one source noted, adding that the figures generated were not even whole numbers. Since October 7, the level of acceptable collateral damage has reached unprecedented levels, with the sacrifice of up to 10 civilians for a low-ranking operative and up to 100 civilians for a high-ranking operative being tolerated. For context, during the U.S. military's fight against ISIS in Iraq, collateral damage exceeding 15 civilians required special authorization from the U.S. Central Command commander.

3. What Automated Profiling Does

The case of automatic profiling of Gaza's population by the Israeli military is an extreme example, but AI-driven profiling has become commonplace across various domains. The reason for utilizing AI profiling is to more reliably secure gains or avoid losses—in other words, for the purpose of "risk management".

Here, "risk" refers to the product of the magnitude of gain or loss associated with an event

and the probability of that event occurring. In everyday language, “risk” usually refers to the possibility of loss or harm, but in the field of risk engineering, even the possibility of gain is referred to as “risk.”

Risk management, broadly speaking, involves (1) identifying the available options for decision-making, (2) estimating the level of risk associated with each option, and (3) choosing an action based on these estimates. Step (2) in this process is referred to as “probabilistic risk analysis”, or simply “risk analysis”. Artificial intelligence is primarily used for risk analysis, specifically for estimating the probability of certain events occurring.

AI used in fields such as marketing, justice, border control, and the military serves the purpose of risk management. For instance, in companies like Amazon, recommending products to users can result in profit if it leads to a purchase, or a loss of potential profit if ignored. Therefore, selecting and recommending products that users are likely to purchase is crucial. Companies achieve this by profiling users based on their data and estimating the probability that a specific recommendation will lead to a purchase.

For Israel, leaving Hamas operatives active poses the risk of attacks on its citizens, making it critical to neutralize as many Hamas operatives as possible. To achieve this, the Israeli military profiles Gaza residents and estimates the probability that a given individual is a Hamas operative.

The use of AI-driven profiling has already been criticized for perpetuating and reinforcing discriminatory stereotypes and practices present in society, promoting privacy violations by corporations and governments, and producing self-fulfilling effects⁽¹⁾ (cf. O’Neil (2016); Véliz (2020)). However, the aspect I wish to focus on here is how the use of AI-driven profiling influences our perception of others and our social relationships.

Sociologist Shoshana Zuboff argues that the proliferation of machine learning systems based on big data leads to a society where “machine processes replace human relationships so that certainty can replace trust” (Zuboff (2019), p. 351). Zuboff is talking about AI tools that can not only predict but also transform people’s behavior, but even profiling tools alone can have a significant negative impact on trust. When tools are available to express the risks associated with engaging with an individual in clear numerical terms, it becomes difficult to interact with others without relying on those tools, or to see them as something more than a risk factor—someone with depth and individuality.

This becomes particularly concerning when such tools are used to identify individuals deemed as risk factors with the aim of avoiding or eliminating them. At the same time,

this issue is not unique to artificial intelligence or something initiated by it. Rather, it aligns with a longstanding social trend of measuring and evaluating individual performance that predates the advent of AI, though AI has the potential to significantly accelerate this tendency. This makes it an even more complex and challenging issue to address.

4. Trust and Assurance

To trust someone in a certain matter, broadly speaking, is to have the expectation that “this person will do such and such” and to base one’s own actions on that expectation. For example, if an acquaintance asks to borrow money from me and I decide to lend it with the expectation that “they will probably pay it back eventually,” this means I have trusted that acquaintance to return the money.

The degree of certainty in one’s expectations of another can vary depending on the situation. In fields such as sociology, it is generally not considered trust if one takes for granted that the other person will not betray their expectations. Trust involves making a decision based on one’s expectations, while also acknowledging to some degree the possibility that the other party might not act as expected. For example, when a café serves coffee to a customer, it typically does not do so based on a calculation of whether the customer might fail to pay. Therefore, it cannot be said that the café is placing trust in the customer. In other words, trust can be understood as “expectation accompanied by a certain degree of doubt.”

Social psychologist Toshio Yamagishi (山岸 (1998)) proposed distinguishing the state in which one does not need to consider the possibility of one’s expectations being betrayed as “assurance (安心),” rather than “trust.” Such assurance may arise, for example, from the enforcement of laws or contracts, strong mutual dependence, or close social bonds within a group. According to Yamagishi, unlike assurance, trust involves acknowledging the possibility of betrayal or deception by the other party. However, trust emerges when one assesses the other person’s internal characteristics—such as intentions or emotions—and concludes that they are unlikely to betray or deceive.

According to Yamagishi, traditional Japanese society was one in which lies and betrayals were suppressed by fostering strong bonds (commitments) within closed, fixed interpersonal relationships. This created a society of assurance. However, in such a society, the costs of rigid relationships were significant—for example, collusion hindering the proper functioning of market mechanisms. Yamagishi argues that Japan’s old, closed, assurance-based society is now irreversibly collapsing. Therefore, Japanese society must transition

into a new, open, trust-based society.

Throughout history, humanity has always been compelled to address the risks posed by others. Consequently, humans have long developed and relied on various mechanisms to meet this need, continuously refining and inventing new ones. What I refer to here as “mechanisms” encompasses innate cognitive, emotional, and behavioral tendencies, culturally formed and learned customs, moral norms, social systems such as politics and law, and technology.

These mechanisms can broadly be divided into two categories. The first is those that enable the identification of individuals posing risks, allowing us to avoid or exclude them. This includes, for example, the ability to detect subtle signs of another person’s emotions or thoughts from their facial expressions or behaviors. Norms surrounding etiquette and manners serve as clues to reveal the nature of individuals, while criminal investigations and judicial processes assist in identifying and isolating dangerous persons.

The second type of mechanism imposes constraints on people’s actions in some way, making it difficult for them to betray or deceive others. This category includes psychological and emotional factors such as loyalty and the sense of reciprocity, as well as penalties for causing harm to others.

AI-driven profiling can be considered a technology aimed at risk avoidance or exclusion. Will the advancements in AI allow us to reliably identify trustworthy individuals in advance, and then will we live in a society so filled with assurance that trust becomes unnecessary?

5. The Pitfalls of Artificial Intelligence

AI-driven profiling, if sufficiently accurate, could serve as a powerful tool for risk avoidance. By using it, one could detect attempts at deception in real time and avoid immediate risks, but also eliminate potential future threats even before an individual forms malicious intentions. Such AI could reveal whether someone will demonstrate the required abilities, work diligently, exhibit violent tendencies, contract certain diseases, or quit their job soon. AI-driven profiling may be a “silver bullet” that exposes and neutralizes the “werewolves” lurking in our midst. Advocates of AI-driven profiling appear to hold this kind of expectation. However, I argue that no such silver bullet exists. The reasons for this will be outlined below.

The first reason is that human personality and abilities are not fixed traits. They evolve over time, shaped by environmental factors, social interactions, and the opportunities one

encounters. The challenge of accurately predicting such dynamic changes far exceeds that of detecting whether someone is lying in the present moment. While it is true that past and present contexts, behaviors, and innate dispositions partially influence future outcomes, they do not fully determine them. Consequently, any prediction about a person's future remains inherently probabilistic and can never attain absolute certainty.

The second reason is that profiling by artificial intelligence is not a neutral act of observation independent of the subject's behavior. AI-driven profiling is frequently used to inform decisions about whether to establish a cooperative relationship with the individual being evaluated, and such decisions often shape the person's subsequent behavior. Recall the concept of self-fulfilling predictions discussed earlier. An individual deemed "suitable for cooperation" by AI is likely to be granted opportunities to achieve results and, by leveraging those opportunities, will be further positively evaluated. Conversely, if AI judges someone "unsuitable for cooperation," that person may be denied the chance to collaborate and produce positive outcomes.

When AI evaluation systems become pervasive, there is a risk that certain individuals could be consistently judged as "unsuitable for cooperation" in all areas of life, systematically denied opportunities. This is the situation constitutional scholar Tatsuhiko Yamamoto (山本 (2017)) describes as a "virtual slum."

Confined within a virtual slum and persistently excluded from opportunities for cooperation, individuals may lose the will to engage with others. Even when opportunities do arise, they may act selfishly, unable to trust both the system or others. In the worst-case scenario, they may fall into despair, grow resentful toward society, and ultimately turn to crime. This, in turn, invites further negative evaluations, thereby perpetuating a self-reinforcing cycle of exclusion and marginalization.

In this way, the werewolf-exclusion system, initially envisioned as a silver bullet, may ironically end up producing new werewolves of its own making.

Third, if the aim is to make artificial intelligence appear flawless, it is strategically advantageous to design it as a highly "suspicious" system. This becomes evident when we consider the circumstances under which errors in AI evaluations come to light.

Errors in AI risk assessments generally fall into two categories. The first is when AI classifies an individual as "low risk," prompting a decision to cooperate, but the individual ultimately fails to meet expectations (a false negative). The second case is when AI classifies an individual as "high risk," resulting in a decision not to cooperate, even though

cooperation with that person would in fact have been successful (a false positive).

As is immediately apparent, in cases of false positives—where the system incorrectly labels someone as “high risk”—it is usually difficult to determine that the AI made a mistake. This is because it is hard to verify whether cooperation with the excluded individual would, in fact, have succeeded. By contrast, in the case of false negatives—where someone is deemed “low risk” but turns out to pose a risk—the AI’s misjudgment becomes visible and undeniable. From the perspective of those designing or operating the system, it is understandable that avoiding false negatives would be prioritized, even at the cost of tolerating a certain number of false positives.

In other words, the systems are likely to be designed with the principle of sacrificing specificity to achieve high sensitivity.⁽²⁾ This design bias becomes especially profound in contexts where the consequences of failing to detect a threat or being betrayed are particularly severe.

In such a scenario, a werewolf-exclusion system may appear to function effectively on the surface, but it carries hidden costs: the exclusion of individuals who could have been valuable cooperative partners. However, this cost is conveniently concealed by the assumption that everyone excluded was, in fact, a werewolf. This logic mirrors the rhetoric employed by the United States during the “war on terror,” in which all individuals killed in drone strikes were posthumously designated as “Enemy Killed in Action.”

This issue is not unique to artificial intelligence. When we avoid or exclude individuals based on stereotypes associated with the groups to which they belong, we are engaging in the same logic. In certain contexts, such actions amount to serious forms of discrimination. For instance, unfavorable treatment by police or government authorities on the basis of gender, race, or health status is widely regarded as impermissible discrimination. Likewise, when companies make hiring or promotion decisions based on such attributes, they are not only unethical but also legally prohibited.

The problem with artificial intelligence lies in its potential to inadvertently learn and reinforce societal stereotypes—often without the awareness of developers or users—and in its ability to generate newly marginalized groups. Imagine a scenario in which AI system, after analyzing vast datasets, infers that “50% of individuals exhibiting certain characteristics are werewolves.” These “characteristics” might be so complex that humans cannot comprehend them. As a result, individuals who match these inscrutable criteria may be labeled “dangerous” with a 50% probability and consequently excluded.

But is it justifiable to exclude someone based solely on a probabilistic assumption about their potential for harm? Such practices raise deep ethical and social concerns regarding fairness, bias, and the unintended consequences of delegating judgment on AI-driven systems.

6. The Ecosystem of Trust

When we trust someone, we often place our faith not only in who they are at present, but also in their potential for future growth and development. Positive change often begins precisely with such an expectation, as many individuals are motivated to live up to the trust invested in them. While this notion aligns with our everyday experience, it has also been empirically supported: a study involving young children demonstrated that being trusted can, in fact, promote honesty (Zhao et al. (2024)).

In this sense, trust is also self-fulfilling. By trusting someone, we project positive expectations onto them and extend opportunities for growth. In doing so, we help to elicit constructive behavior. Conversely, distrust has its own consequences. Withholding trust communicates negative expectations and simultaneously deprives the individual of opportunities to act positively. As a result, we risk stifling their potential for growth and reinforcing the very outcomes we feared.

Unlike artificial intelligence, human beings do not always ground their trust in solid evidence. We sometimes choose to trust individuals with no proven track record—or even complete strangers. Moreover, we often forgive and continue to trust those who have previously betrayed our expectations. This uniquely human capacity to trust beyond evidence reflects a forward-looking hope—a willingness to believe in the possibility of change—rather than a reliance on certainty.

To borrow the words of philosopher Nobutoshi Nagamori, trust has the “power to bind the other to ‘become’ a subject worthy of trust,” and it “seeks to fulfill its expectations by aiming toward the future” (永守 (2024), p. 6). Nagamori illustrates this idea by referring to the actions of Bishop Myriel toward the protagonist, Jean Valjean, in Victor Hugo’s novel *Les Misérables*.

When Valjean, newly released from prison, meets Bishop Myriel, he is filled with deep resentment toward society. Despite Valjean’s theft of his silverware, the bishop forgives him—and even offers him additional gifts. This act of compassion inspires Valjean to transform his life.

However, not all acts of trust yield such positive outcomes. Research by neuroeconomist Paul J. Zak, employing the “Trust Game,” reveals that approximately 5% of individuals are “unconditional non-reciprocators,” underscoring the inherent risks involved in extending trust (Zak (2012)).

In *JoJo’s Bizarre Adventure*, a similar act of forgiveness occurs when Lord Joestar pardons Dario Brando for stealing his ring. However, unlike Valjean, Dario does not repent. Instead, his selfishness leads to further harm, culminating in his son Dio bringing tragedy to the Joestar family.

This contrast highlights the unpredictable outcomes of trust and forgiveness. Trust may transform some werewolves into virtuous villagers, but not all werewolves will change. Therefore, we must strive to reduce the uncertainty in others’ behavior through every means available, carefully assess whether someone is worthy of trust, and ultimately act by placing faith in our judgment despite lingering uncertainties.

What if the world were populated only by people like Bishop Myriel and Lord Joestar? Such a society would undoubtedly be remarkable, but it would not be stable or enduring. Sooner or later, someone like Dario Brando would emerge, exploiting the goodwill and tolerance of others for personal gain. The Brando lineage would prosper, while the descendants of the Joestars would face repeated hardship.

However, over time, a new kind of individual would begin to appear—one distinct from both the Brandos and the Joestars. These individuals would neither trust indiscriminately nor reject trust altogether. Instead, they would carefully assess others’ trustworthiness. When they deemed someone reliable, they would extend trust and cultivate robust cooperative relationships. They would feel a moral obligation to reciprocate goodwill, but they would also be unafraid to respond firmly, even with anger, when that trust was betrayed. In the world of *JoJo’s Bizarre Adventure*, such individuals are best exemplified by Speedwagon.

In the long run, individuals like The Joestars and Speedwagon would ultimately fare better than those like the Brandos, who are constantly scheming and betraying others. This is because the Joestars and Speedwagons can reap the benefits of cooperation that the Brandos, due to their untrustworthy nature, cannot attain. While The Joestars and Speedwagon may occasionally suffer losses due to betrayal, their capacity to build and sustain cooperative relationships would, over time, lead to greater success. In contrast, the Brandos’ self-serving behavior ultimately undermines their ability to form lasting alliances, limiting

their long-term gains.

Traits such as tolerance, quickness to anger, persistence, gullibility, skepticism, caution, openness, strong group loyalty, a sense of obligation, and cunning are all personality characteristics related to trust. Society is composed of individuals who exhibit these traits to varying degrees. Moreover, even within a single person, these traits are not fixed. They may manifest differently depending on the context, and they can evolve over time through interactions with others. This dynamic interplay of trust-related traits constitutes the ecosystem of trust that humanity has cultivated over the course of its long social history.

This ecosystem has likely been repeatedly forced to adapt in response to shifts in social structures and technological developments. I contend that the widespread adoption of AI-driven profiling has the potential to significantly disrupt this balance. The consequences of such a disruption are likely to exceed anyone's ability to predict.

Moreover, it may constitute a “transformative experience” (cf. Paul (2014))—an event that alters our fundamental values and preferences in such a way that it becomes impossible to assess, in advance, whether its overall impact will be good or bad.

However, I fear the emergence of a troubling scenario: the designers and users of werewolf detection systems driven by an incentive to maximize sensitivity at the expense of specificity (as in the case of Lavender), may configure these systems to produce increasingly “suspicious” judgments. This approach not only carries the hidden cost of obstructing potential cooperative relationships but also risks transforming falsely identified innocents into actual werewolves through systematic exclusion.

Such transformation, in turn, reinforces the perceived legitimacy and indispensability of the detection system itself, thereby initiating a self-reinforcing cycle in which society becomes ever more reliant on these technologies—regardless of their ethical or epistemic soundness.

7. Conclusion

This paper has examined the challenges posed by the rapidly advancing field of artificial intelligence, with particular focus on systems that utilize big data to perform human profiling.

The increasing use of AI systems that profile individuals based on big data may encourage us to view others as machine-readable bundles of data and inferred attributes. Moreover, AI may foster an evaluative stance in which others are primarily assessed as potential sources

of risk—that is, in terms of the benefits or harms they might pose. This “culture of risk analysis” threatens to erode trust in human relationships.

However, this trend is not unique to AI. Rather, it represents a continuation of the modern tendency to measure and quantify human performance wherever possible. Given this trajectory, the growing ubiquity of AI across various domains appears almost inevitable. Organizational managers may increasingly come to regard it as their responsibility to use AI for risk assessment and mitigation. Just as management consultants have long promoted various performance metrics to administrators, vendors of AI systems may come to occupy a significant share of the “bullshit job market” within organizational management.

On the other hand, those subject to AI evaluation are likely to modify their behavior in order to secure favorable assessments. Just as web administrators engage in search engine optimization to improve their rankings, individuals may begin optimizing themselves to perform well under AI-driven evaluation systems.

However, such optimization efforts risk obscuring the original purposes these evaluations were intended to serve. Individuals labeled as “high risk” by widely adopted AI systems may find themselves systematically excluded from various opportunities, effectively relegated to a “virtual slum” (山本 (2017)).

This perspective may appear unduly pessimistic. Yet as AI becomes increasingly embedded in various domains of life, historical patterns suggest that it will inevitably intensify our already excessive preoccupation with measurement, while further entrenching cultures of prediction and risk analysis.

In this context, it is essential to preserve a humanistic perspective—one that regards humans not merely as sources of “bias”, but as sources of “value” (cf. 隠岐 (2018)). We must also recognize that believing in the potential for human growth—even in the absence of evidence or in the face of evidence to the contrary—is not an irrational stance.

註

* kukita.minao.p7@f.mail.nagoya-u.ac.jp

(1) A phenomenon where a prediction causes itself to become true. For example, rumors of a financial institution being at risk of collapse may trigger a bank run, ultimately leading to its failure.

(2) The terms “specificity” and “sensitivity” are commonly used in medical testing. Specificity refers to the proportion of true negative results among all actual negative cases, while sensitivity refers to the proportion of true positive results among all actual positive cases. Typically, there is a trade-off between the two: a highly sensitive test—one that minimizes missed diagnoses—often sacrifices specificity, resulting in more false positives for individuals who do not actually have the disease.

文献

- Abraham, Y. (2024). “Lavender’: The AI Machine Directing Israel’s Bombing Spree in Gaza.’ Retrieved from <https://www.972mag.com/lavender-ai-israeli-army-gaza/>.
- Chamayou, G. (2013). *Théorie du Drone*, Paris: Fabrique.
- 永守伸年 (2024). 『信頼と裏切りの哲学』, 慶應義塾大学出版会. (Nagamori, N. (2024). *Philosophy of Trust and Betrayal*, Tokyo: Keio University Press).
- 隠岐さや香 (2018). 『文系と理系はなぜ分かれたのか』, 星海社. (Oki, S. (2018). *Why Were the Humanities and Sciences Divided?*, Tokyo: Seikaisha).
- O’Neil, C. (2016). *Weapons of math destruction : how big data increases inequality and threatens democracy*, New York: Crown.
- Paul, L. A. (2014). *Transformative experience*, Oxford: Oxford University Press.
- Scahill, J. & The Staff of *The Intercept* (2017). *The Assassination Complex: Inside the Government’s Secret Drone Warfare Program*, New York: Simon and Schuster.
- Véliz, C. (2020). *Privacy Is Power: Why and How You Should Take Back Control of Your Data*, London: Transworld.
- 山岸俊男 (1998). 『信頼の構造——こころと社会の進化ゲーム』, 東京大学出版会. (Yamagishi, T. (1998). *The Structure of Trust: Evolutionary Game of Mind and Society*, Tokyo: The University of Tokyo Press).
- 山本龍彦 (2017). 『おそろしいビッグデータ』, 朝日新聞出版. (Yamamoto, T. (2017). *Big Data Is Terrible*, Tokyo: Asahi Simbun Press).
- Zak, P. J. (2012). *The moral molecule : the new science of what makes us good or evil*, London: Transworld Publishers.
- Zhao, L., Mao, H., Harris, P. L., & Lee, K. (2024). ‘Trusting Young Children to Help Causes Them to Cheat Less,’ *Nature Human Behaviour*, 8, 4, 668–678. doi: 10.1038/s41562-024-01837-4. Epub 2024 Feb 20.
- Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, London: Profile Books.

[Graduate School of Informatics, Nagoya University/Philosophy]