AI Can Help Us Live More Deliberately

We need a little friction in our lives to trigger reflection, self-awareness, and responsible behavior.

Julian Friedland
Assistant Professor of Ethics
Trinity Business School, Trinity College Dublin.

Penultimate draft, MIT Sloan Management Review, 60 (4) 2019.

Leading Question
How do we design machines smart enough to keep us from becoming like machines ourselves?

Findings:
- As we "outsource" myriad tasks to AI-assisted platforms, we may become less reflective and feel less responsible for outcomes.
- Moral self-awareness is a powerful motivating force that can help restore critical self-reflection, agency, and a sense of accountability.
- AI developers can incorporate prompts that promote moral self-awareness in areas ranging from health and wellbeing to media and civic engagement.

As I search online for a present for my mother, considering the throw pillows with sewn-in sayings, plush bathrobes, and other options, and eventually narrowing in on one choice over the others, who exactly has done the deciding? Me? Or the algorithm designed to provide me with the most "thoughtful" options based on a wealth of data I could never process myself? And if Mom ends up hating the embroidered floral weekender bag I end up "choosing," is it my fault? It's becoming increasingly difficult to tell, because letting AI think for us saves us the trouble of doing it ourselves and owning the consequences.

AI is an immensely powerful tool that can help us to live and work better by summoning vast amounts of information. It spares us from having to undergo many mundane, time-consuming, nerve-wracking annoyances. The problem is that such annoyances also play a key adaptive function. They help us learn to adjust our conduct in relation to one another and the world around us. Engaging directly with a grocery bagger, for instance, forces us to confront her humanity, and the interaction (ideally) reminds us not to

get testy just because the line isn't moving as quickly as we'd like. Through the give and take of such encounters, we learn to temper our impulses by exercising compassion and self-control. Our interactions serve as a constantly evolving moral-checking mechanism.

Similarly, our interactions within the wider world of physical objects forces us to adapt to new environments. Walking, bicycling, or driving in a crowded city teaches us how to compensate for unforeseen obstacles, such as varying road and weather conditions. In countless occasions every day, each of us seeks out an optimal compromise between shaping ourselves to fit the world and shaping the world to fit ourselves.[1] This kind of adaptation has led us to become self-reflective and autonomous, capable of ethical considerations and aspirations.

Our rapidly increasing reliance on AI takes such interactions out of our days. The frictionless communication AI tends to propagate may increase cognitive and emotional distance, thereby letting our adaptive resilience slacken and our ethical virtues atrophy from disuse.[2] Relying on AI to pre-select gifts for friends and family, for example, spares us the emotional labor of considering their needs and wants in our ordinary interactions with them to select a genuinely thoughtful gift. Many of the trends already well underway involve the offloading of cognitive, emotional, and ethical labor to AI software in myriad social, civil, personal, and professional contexts.[3] Gradually, we may lose the inclination and capacity to engage in critically reflective thought, making us more cognitively and emotionally vulnerable, and thus more anxious[4] and prone to manipulation from false news, deceptive advertising, and political rhetoric.

In this article, I consider the overarching features of this problem and provide a framework to help AI designers tackle it through system enhancements in smartphones and other products and services in the burgeoning Internet of Things marketplace. The framework is informed by two ideas: Daniel Kahneman's cognitive dual process theory[5] and moral self-

awareness theory, a four-level model of moral identity that I developed with Benjamin M. Cole, a professor at Fordham's Gabelli School of Business.[6] (See "Theories of Mind in an AI World," page XX.)

When Convenience Leads to Disengagement

The most immediately attractive feature of AI technology is its promise to handle the mundane aspects of life, thereby increasing the amount of time and attention each of us can devote to activities we consider more rewarding. Of course, every time this kind of outsourcing occurs, we cede a degree of control. Getting comfortable with these trade-offs reinforces new habitual behaviors that entail a measure of disengagement: from one another, the physical world, and even ourselves. This is because every time we delegate a degree of control to the AI system, we also invest a degree of trust into that system. In so doing, we will often shift from relying on what Kahneman calls our reflective mind (and its deliberative decision-making) to our autonomous mind (and its automatic reactions that guide decisions). This makes it easy to complete a routine task. But repeating this process creates a risk that our actions become increasingly automatic and less reflective overall, leading to six forms of disengagement as described by Frischmann and Selinger[7]:

1. Increased passivity. As we accept assistance to complete a task, we require less effort to carry it out. We may become spectators rather than active participants. The AI systems that Netflix, Amazon Prime, and Facebook use to pre-select entertainment and news options are examples. When we let these systems determine our options, we rarely confront perspectives that might challenge our preconceptions and biases. Gradually, we may become less prepared to expend the effort needed to think deeply and critically, thereby disengaging long-term memory.[8]

2. Emotional detachment. Diminished participation leads to emotional disengagement. Consequently, our actions can become insincere or deceptive. Think of a customer call

center, where an AI system in a help desk or sales context aggressively coaches an agent's conversation volleys based on the customer's emotional cues.[9] Such software, ideally designed to train operators to become more sensitive to customers' concerns, could have the reverse effect, making us increasingly inured to emotional cues because we will have less practice picking up these cues ourselves, and less interest in doing so.

3. Decreased agency. Disengagement relinquishes the power to make our own decisions by lessening our awareness of actions we might take. Consider an automated vehicle pre-programmed to weigh competing ethical priorities during a crash, such as whether to hit a pedestrian or another vehicle. Auto insurance rates might be adjusted according to the degree to which we set the automated driving system to integrate others' interests into the calculus.[10] And we relinquish the agency to make our own choice as the crash takes place.

4. Decreased responsibility. In ceding control over a decision-making process, we can become less accountable for results – whether they are good or bad – because responsibility is diffused across the entire system, from design to delivery. Imagine a dieting app that orders prepared foods to be delivered to you according to a weight loss plan set up by AI. If you lose weight, who deserves the credit? And if you don't, who's fault is it?[11]

5. Increased ignorance. AI translates our wants into algorithmic shorthand or mechanical processes that may end up functioning differently than we would ourselves. Of course, that can make up for deficiencies in our knowledge—but it can also reinforce those deficiencies. Virtual navigational apps like those offered by Waze, Garmin, and others do not require you to acknowledge your surroundings. You might keep circling an incorrect location that the mapping app has not yet updated out of preferential bias for the AI system instead of returning back to your own direct perceptions and judgments.[12] At your intended destination, you might have no idea what route you took to get there nor how to get back to where you

started without AI assistance.

6. Deskilling. Depending on an intermediary for completing routine tasks can dull many of the trained skills we rely on to interact with the physical world around us. We may forget how to perform basic tasks or become less proficient at doing them unaided. Using only navigation apps lulls us into forgetting how to use a conventional map or, in a future era of autonomous vehicles, even how to drive without the apps. We may also lose motivation to acquire new skills, opting instead for ever more outsourcing solutions.

Together, these trends present for us an ethical challenge: by multiplying the instances in which we go through life while operating on autopilot, they have the potential to loosen our social bonds, exacerbate conflicts, and hamper moral progress by stifling self-critical thought. To mitigate these threats, designers of AI systems should build in features and interfaces that periodically re-trigger our reflective minds.

It Takes More Than "Nudges" to Make Us Think

In their influential book, *Nudge*, Richard Thaler and Cass Sunstein have argued that cognitive nudges can spur us to action by using triggers that evoke emotions like empathy or self-interest.[13] Unfortunately, such nudges have limited power in practice because they only prompt behavioral impulses and do not engage critical reflection. This is the case even when pressing health risks are concerned. In a study of 1,509 patients who had heart attacks, efforts to prompt people to adhere to medication prescriptions (including electronic pill bottles, the chance for $5 or $50 rewards and enlist a friend or family member in the effort) did not significantly improve the chance people would take their medicine.[14]

Triggering the reflective mind is more likely to solve the problem of disengagement and mitigate the risks of losing skills in the age of AI. By creating what we can call cognitive

speed bumps that force us to reflect on decisions worthy of greater reflection, developers of AI systems can re-introduce *interactive friction* into the experiences they host. So as Mom's birthday approaches, instead of suggesting purchases, our AI system might instead suggest a good time to call or pay Mom a visit—an opportunity to enhance the personal relationship and even help come up with a thoughtful (and desired) gift.

The ramifications are profound. Perhaps the most seductive aspect of AI-assisted platforms is that they promote what technology ethicist Shannon Vallor describes as "interactions that deftly evade the boredom, awkwardness, conflict, fear, misunderstanding, exasperation, and uncomfortable intimacies that often arise from traditional communications, especially face-to-face encounters in physical space."[15] Here, Vallor is referring mainly to live conversations that can be avoided through social media. But we can include all the practical drudgeries of life from reading a map, driving a car, and minding one's surroundings, to making a grocery list, shopping, and cooking. And though most of us still have such frictional encounters, AI-assisted platforms promise to guide our attention in whatever directions we are likely to find most immediately satisfying, thereby reducing the chances that we will have to experience *unpleasant* friction. As a result, our moral attention—the ability to redirect our focus, delay gratification, temper our emotional urges, and restrain our unthinking reactions—erodes.

We need something to counteract this tendency: an AI choice architecture designed to preserve healthy measures of interactive friction between ourselves and the wider world.

How Friction Fosters Moral Self-Awareness

There is value to a world of friction-filled interactions. For instance, new research on childhood self-control suggests that one's cultural[16] and socio-economic[17] environments may play a far greater role than genetic factors in developing grit and perseverance, which are

highly corelated with professional success later in life. [18] It is only by learning how to navigate interactions that are not set up for our comfort that we are able to fully develop executive control over our own consciousness.[19]

Such interactions also foster moral self-awareness. As we experience friction again and again, the ways we react to various stimuli change, and moral identity evolves: we begin to think and feel differently about what our actions say about ourselves.[20]

The social psychological literature has established a clear relationship between what's called the self-importance of moral identity and moral thought and action,[21] and the wider literature on civic mindedness indicates that pride is the most effective moral motivator of civic behavior.[22] There is also evidence that ethical consumers are happier and have stronger repurchase intentions when motivated by their moral self-image than when motivated by emotions such as guilt and empathy.[23]

What does all this have to do with AI? Designers of AI systems can use the four levels of moral self-awareness described below as a guide for developing applications that encourage reflective behavior. By incorporating triggers for interactive friction, they can prompt users to consider how their actions reflect their personal values and help them ascend to higher levels of awareness.

Level 1. At this level, people rely chiefly upon negative feedback they receive from observers to guilt or shame them into changing their behavior. Researchers have demonstrated the power of negative feedback to inhibit a person's selfish behavior. For example, participants primed in a Tragedy of the Commons experiment to be self-interested gradually learned to temper their self-interest after being shamed by other subjects left with fewer resources.[24] Eventually, all subjects showed preference for lowered individual returns in favor of equitable and sustainable longer-term outcomes.

Level 2. At this level, individuals become more self-reflective. Rather than relying on others' complaints to acknowledge the negative impacts of their actions, actors start to serve as their own source of feedback. This happens when they see the outcomes of others' behavior or when they consider the ramifications of their own actions. For example, a person who notices a room containing swept litter is 2.5 times less likely to toss trash on the floor than in a litter-strewn room.[25] Observing the neatened-up litter increases the observer's propensity to keep the room clean.

Level 3. At this level, people start to anticipate potential negative consequences of their actions and do so independently from others' signals. This behavior often comes after self-reflection on prior behavior has led to an internal sense of guilt or shame. At a crucial turning point in the Tragedy of the Commons experiment cited above,[26] one participant asked aloud, "Are we bad people?" This question was not so much an effort to shame other group members as it was an attempt to reconcile the inconsistency between one's prior action (to serve self-interest) and one's aspirational moral self-image. Such a reflective moment represents a crucial step, one that reveals the moral obligations of individuals to shape themselves to fit the world and their own aspirations within it.

Level 4. At the highest level, people become increasingly forward-looking, considering both negative *and* positive impacts. They purposely engage in appropriate actions to realize positive outcomes. They internalize the self-image of potential hero rather than potential villain.[27] At best, these decisions are habit-forming, bringing persons closer to becoming whom they aspire to be. This state of mind is linked with achieving greater happiness based on an individual's self-conception.[28]

Triggering the Reflective Mind

In traditional face-to-face interactions, the external physical or social world provides the friction necessary to trigger the reflective mind into modifying one's behavior for the

better. As AI removes opportunities for those interactions, developers need a tool for tapping into users' moral self-awareness. "Showing notices," a type of visceral notice that AI systems can incorporate to shape users' decision-making, can serve as that tool and compensate for the loss of give-and-take interactions in the social and physical world. (See "Theories of Mind in an AI World" for more detail about visceral notices.)

Showing notices provide users with snapshots of their behaviour (the number of steps taken in a day, for example, or the amount of time spent online). They can enhance AI applications by encouraging users to move from the first and second levels of moral self-awareness, in which negative feelings like guilt and shame primarily drive individual behaviors, toward level 4, in which positive aspirations encourage people to act, conscious that their choices can make a difference for themselves and society. Enabling users to share their progress on a given issue with others in a social group further enhances an application's potential.

Considering that by current projections, global IoT spending could reach $1.4 trillion by 2021, such functionality presents rich opportunities for research and development.[29] Five lifestyle categories in particular have significant potential for this type of innovation: health and wellbeing, social responsibility, media and civic engagement, skill maintenance, and personal edification. We'll consider each one here.

Health and wellbeing. There is already significant movement in providing showing notices in health and wellness apps—from those that facilitate personal fitness, mindfulness, or sleep management to those that allow us to set screen-time limits or monitor our diets. Smart refrigerators are another frontier. For example, adding showing notices that illustrate patterns of consumption of highly processed, high-sugar, canned, frozen, and fresh food, along with daily calorie consumption data, can help users improve nutrition. Combined with data from

grocery delivery services, such notices can guide them to order groceries according to healthier recipes and locally or sustainably sourced foods.

Social responsibility. Another area with potential is helping people make thoughtful brand and investment choices that align with their social values. A few apps now highlight possible ethical concerns in financial portfolios, flagging sectors that users may wish to avoid in light of stated preferences (such as alcohol, petroleum, and tobacco) and providing finer-grained notices about any ethical quandaries firms may be involved in. Smart refrigerators can provide notices about the carbon footprint of groceries sold (where consumers have access to carbon labels). Such notices could extend to other areas, alerting users to factors such as air and water pollution, resource depletion, and green packaging.

Media and civic engagement. Media-quality applications could use showing notices to alert people to misleading or biased news sources, both on a case-by-case basis and in their overall news consumption. New tools could gradually introduce alternate points of view, encouraging users to break out of ideological echo-chambers. Smart citizen phone applications now allow users to develop localized crowd-sourced maps revealing problem areas for litter, broken street lights and windows, vandalism, pot holes, and so on. Aptly designed visceral notices could encourage citizens to increase their levels of civic awareness and engagement on local, national, and international levels, prompting them to take action where help is needed.

Skill maintenance. Our willingness to outsource tedious physical engagements with the external world may lead to a significant loss of everyday skills. GPS mapping and automated driving systems are cases in point. When following the visual or voice directions today's systems offer, users don't need to pay attention to landmarks. Showing notices offer a potential corrective. An AI-enabled system could include 3-D images of key landmarks and points of reference where turns must be made. This would allow users to orient themselves to

their surroundings and rely on their own memory of the landscape to reach their destinations. Other designs could encourage drivers to stay alert and to maintain their driving skills instead of becoming overly reliant on automated driving systems.

Personal edification. Ultimately, what aptly designed visceral notice environments can provide are AI systems that act less like an object and more like a friend that helps users develop to their fullest potential. Consider the capacity of AI systems to encourage greater discernment in domains such as the arts, cuisine, fashion, and entertainment. Instead of exposing people to whatever products they may react most impulsively to, as recommendation engines often do, they could show alternatives with high quality ratings based not merely on popularity but also on a blend of expert opinion and personal and shared social preferences. Some services such as Netflix already provide such distinctions, but without a feature showing the user's overall viewing choices and screen time. Users could also be given finer-grained film rating categories including acting, direction, dialogue, and storyline.

AI-assisted platforms provide consumers with extraordinarily powerful tools for controlling and managing their daily lives, activities, and interactions. Such technology, if designed carefully and conscientiously, also holds the power to alter human behavior for the better on a massive scale. But if designed short-sightedly, with few if any features for counteracting its own negative habit-forming effects, it could instead foster passivity, dependency, ignorance, and vulnerability.

It is essential that firms working in this area formulate clear and cogent design strategies to allow customers to make informed choices regarding their own patterns of online behavior. The ones that do will play a key role in optimizing collective wellbeing by safeguarding personal agency.

Sidebar: Theories of Mind in an AI World

Cognitive dual process theory describes two overarching decision-making processes: (1) the *autonomous mind*, which automatically reacts to stimuli, and (2) the *reflective mind*, which responds consciously in a deliberate and reasoned fashion.[30]

Most AI-assisted platforms function to free up the attention of the conscious reflective mind for any activities that immediately suit a person's interest or grab her attention. Ideally, each new outsourced task is accomplished more effectively than via direct unassisted interaction. Thus, AI allows us to conveniently increase the levels at which we may productively process incoming information from the external physical and social worlds.

AI systems typically use a series of visceral notices to guide users, divided into three general categories. Researchers have divided visceral notices into three general categories:[31]

1. **Familiarity notices** use familiarity with one technology to inform users about another. Example: camera clicking sounds and dial tones on smartphones.
2. **Psychological reaction notices** use common psychological reactions to shape a consumer's conception of the product or service. Example: casual interface designs such as friendly avatars that trigger greater honesty and openness.
3. **Showing notices** promote self-awareness by showing users the results of their activities. Example: screen-time data embedded in the iPhone iOS 12.

Familiarity notices and psychological reaction notices are designed to trigger only the autonomous mind, but *showing notices* introduce communicative friction designed to trigger the reflective mind. Screen-time software embedded in the iPhone operating system shows people how often they use social networking, entertainment, and productivity apps. This allows them to better understand and take control of their own behavior.

References

1. R. Wollheim, *The Thread of Life*, Yale: 1984.

2. M. Sandel, "The Case Against Perfection: What's Wrong with Designer Children, Bionic Athletes, and Genetic Engineering," *Atlantic Monthly*, April, 2004; S. Vallor, *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*, Oxford: 2016.

3. B. Frischmann & E. Selinger, *Re-Engineering Humanity*, Cambridge: 2018.

4. Greg Lukianoff and Jonathan Haidt, "The Coddling of the American Mind: How Good Intentions and Bad Ideas Are Setting Up a Generation for Failure," Penguin Random House, 2018.

5. D. Kahneman, *Thinking Fast and Slow*, Farrar, Straus, and Giroux: 2012.

6. J. Friedland J. Friedland & B.M. Cole, "From Homo-Economicus to Homo-Virtus: A System-Theoretical Model for Raising Moral Self-Awareness," *Journal of Business Ethics*, 2017, https://doi.org/10.1007/s10551-017-3494-6.

7. Frischmann & Selinger, *ibid*.

8. N. Carr, The Glass Cage: *How Computers are Changing Us*, Norton, 2015.

9. A. Pentland, "The New Science of Building Great teams", *Harvard Business Review* (April 2012).

10. P. Lin, "Why Ethics Matters for Autonomous Cars", in L. Lin, K. Abney, & G.A. Bekey, *Robot Ethics*, MIT, 2014.

11. Sandel, *ibid*.

12. N. Carr, *ibid*.

13. R. Thaler & C. Sunstein, *Nudge: Improving Decisions About Health, Wealth, and Happiness*, Yale, 2008.

14. K. Volpp et al., "Effect of Electronic Reminders, Financial Incentives, and Social Support on Outcomes After Myocardial Infarction. The HeartStrong Randomized Clinical Trial", *Journal of the American Medical Association, Internal Medicine*, 177/8 (August 2017):1093-1101.

15. Vallor, *ibid*, p. 161.

16. B. Lamm et al., "Waiting for the Second Treat: Developing Culture-Specific Modes of Self-Regulation," *Child Development*, 89/3 (June 2018): e261-e277.

17. T. Watts, J. Duncan, H. Quan, "Revisiting the Marshmallow Test: A Conceptual Replication Investigating Links Between Early Delay of Gratification and Later Outcomes," *Psychological Science*, 29/12 (May 2018): 1159 – 1177.

19. Vallor, *ibid*, p. 163.

20. Friedland & Cole, *ibid*.

21. K. Aquino & A. Reed, "The Self-Importance of Moral Identity," *Journal of Personality and Social Psychology*, 83/6 (January 2003): 1423-1440.

22. Bowles, *ibid*.

23. K. Hwang & H. Kim, "Are Ethical Consumers Happy? Effects of Ethical Consumers' Motivations Based on Empathy Versus Self-Orientation on Their Happiness," *Journal of Business Ethics*, 151/2, 2018: 579-598.

24. J. Sadowski et al., "Intergroup Cooperation in Common Pool Resource Dilemmas: The Role of Ethical Leadership," *Science and Engineering Ethics*, 5, 2015: 1197-1215.

25. R.B. Cialdini, C.A. Kallgren, R.R. Reno, "A Focus Theory of Normative Conduct: A Theoretical Refinement and Re-evaluation of the Role of Norms in Human Behavior," *Advances in Experimental Social Psychology*, 24, 1991: 201-234.

26. J. Sadowski at al., "An Experimental Game-Theoretic Pedagogy for Sustainability Ethics", 19/3, 2013: 1323-1339.

27. A. Golpadas, "Marketplace Sentiments," *Journal of Consumer Research*, 41/4, 2014: 995-1014.

28. Hwang & Kim, *ibid*.

29. L. Columbus, "2017 Roundup of the Internet of Things", *Forbes*, December 10, 2017.

30. Kahneman, *ibid*.

31. R. Calo, "Against Notice Skepticism in Privacy (And Elsewhere)," *Notre Dame Law Review*, 87/3 (October 2012): 1027-1072.