Using Human Cognitive Limitations to Enable New Systems

Vincent Conitzer
Duke University
Durham, NC, USA
conitzer@cs.duke.edu

Abstract

For many applications, human cognitive limitations are something that poses challenging constraints. In this paper, I argue that they also have their advantages: they enable new systems and functionality that would be impossible without them. This creates an opportunity for new and exciting interdisciplinary collaborations between cognitive science and computer science.

Introduction

In the field of human computation, we tend to focus either on humans' cognitive *strengths*, as these allow us to do things that are still out of reach for AI, or on *overcoming* their cognitive *weaknesses*, for example by making technology more usable or by mitigating human biases. But in this paper, we will consider human cognitive weaknesses in a positive light. I will argue that human cognitive weaknesses allow us to build systems with desirable properties that would be unattainable without those weaknesses. Generally, what I have in mind is attaining various security properties: preventing people from manipulating the system in a way that would be possible without cognitive limitations. But the security properties are unlike those typically studied in the computer security or cryptography communities, because they rely on specifically *human* limitations. Human cognitive limitations are many, and so are the various desirable properties that we may wish to achieve. Therefore, there is a rich field of inquiry at the intersection of cognitive science and computer science that I believe has so far remained largely unexplored.

We will illustrate the general agenda with several examples. Each example consists of:

1. a motivation,
2. various desiderata – properties that would enable the imagined system,
3. a high-level strategy for obtaining the desiderata, and
4. human cognitive weaknesses on which the strategy relies.

Such a description by itself is, of course, not a working system. Generally, there are many details to be worked out, which can be done in many ways. Whether the system is in fact successful at attaining the desired properties generally also requires experiments on human subjects. However obvious or well established a human cognitive weakness may be, its mere existence is generally no guarantee; if we rely on it for a security property, we need to be sure that the weakness is severe and consistent enough across subjects that the system cannot be broken.

We will discuss three examples that fit within the general framework. For one of them, we have already filled out the high-level strategy in several different ways, but in experiments on human subjects, the results are not (yet) good enough to enable real applications. For another one, with significantly weaker desiderata than the first one, we already have a detailed implementation that performs extremely well in human subjects experiments. Finally, our third example is one for which we have no implementation or evaluation yet, but it has significantly different desiderata and cognitive weaknesses on which the strategy relies, thereby illustrating that the agenda is much broader than the earlier two examples illustrate.

Before we discuss the examples, one important caveat is in order. We rely on specifically human cognitive weaknesses that are not shared by computers or AI. Consequently, for the systems that we have in mind, it is essential that they also involve a cognitive task that remains out of reach for AI; otherwise, the system would not be robust to attacks by AI (or a human-AI team).[1] We will postpone consideration of this issue to the end of this paper. This is because it is easier to present the approach, as well as to actually perform human subjects experiments, without worrying about this additional issue. The natural approach is to first get the methodology to work on human subjects unaided by AI, and then move on from there. But this is not to deny that it is important to be robust to human-AI teams. Fortunately, it seems that in many cases such robustness can likely be achieved by integrating some kind of CAPTCHA (von Ahn et al. 2003; von Ahn, Blum, and Langford 2004) or adversarial examples (Szegedy et al. 2014). But here, too, if we rely on these

[1] There may be situations in which we can be sure the users would not have access to any aid from computers or AI, so that we would not have to worry about this; but, at least for high-stakes applications, this is likely more the exception than the rule.

for important security properties, we should experimentally test whether the system is indeed robust to human-AI teams. How to set up such experiments is a challenging open question in itself, to which we will return at the end of the paper.

We now proceed with our discussion of the examples, considering initially the case where humans are unaided by AI or other tools.

Tests That People Can Pass Once But Not Twice

Motivation. For many purposes, we would like to allow any human being to sign up for an account, but we do not want anyone to be able to sign up for *more* than one account. For example, an e-mail provider may want to allow anyone to have an account, but may also be worried about a single person obtaining multiple accounts, because the person may be using them for undesirable purposes such as sending spam or attempting to influence others politically in a misleading way. (Note that there may be legitimate reasons for having multiple accounts. For example, the person may wish to separate work and personal life. Also, the person may be part of a sensitive or even persecuted class and wish to separate accounts for that reason. As long as the e-mail provider is trusted,[2] it can allow multiple sub-accounts under the main account to allow this functionality.) Another scenario is that a company wants to offer anyone a one-week free trial of its product, but wants to prevent a person from using the product indefinitely by opening a new account every week.

Desiderata. We want any person to be able to sign up for an account once, but not twice. We do not want to rely on real-world identifying information such as social security numbers, both for privacy reasons and because this identifying information may be too heterogeneous across countries. We also do not want to rely on IP addresses or anything of the sort. Instead, we want something that may appear impossible: a test that (almost) anyone could pass once, but (almost) nobody could pass twice!

Strategy. One way in which this could be achieved is through a *randomized memory test*, as follows. When someone wishes to obtain an account, she is given a randomly selected instance of the memory test, in which she is first asked to memorize some randomly selected items, and then to recall them. The key insight is to ensure that the randomly selected instances of the test, while different, are closely related, so that having taken the test previously causes the subject to become confused and therefore give wrong answers.

Cognitive weaknesses required. Humans are unable to erase their memories at will. Moreover, their memories are not always complete: they may recall something they saw without recalling exactly when they saw it.

Existing implementations and experiments. Earlier work (Conitzer 2010) introduced several designs for a memory test of this nature. One involved a small database of 58 images of faces of distinct people. A subject was presented 29 of these at random, and asked to remember them. Then, the subject was shown all 58 faces and asked which ones she had seen previously. The whole procedure was then repeated again, with a different random draw of 29 faces (out of the same 58), to simulate someone attempting to pass the test a second time. The hypothesis behind the test design was that the first iteration of the test should be straightforward for the subject, but the second iteration should be significantly more challenging because the subject is asked to mark which faces she saw *in this iteration* – while at that point having seen every face between 1 and 3 times, depending on the random draws. Hence, simply judging whether the image is familiar is sufficient in the first iteration, but not in the second.

Unfortunately, the experiment did not bear out the hypothesis. While some subjects' scores (number of correct answers) indeed did decrease in the second iteration, there were also some whose scores increased (perhaps due to an improved memorization strategy). On top of this, scores varied significantly from one subject to another, making it difficult to set a "pass" threshold that everyone could meet but that would rule out subjects passing the test a second time. A different test design, based on associating colors with items, performed better, but not well enough for real deployment (also given the length of that test).

Tests That People Can Pass Once But Not Twice *at the Same Time*

Motivation. Since we have not yet been able to design a system that achieves the objective from the previous section (at least well enough for real applications), we here consider a restricted objective. Under certain circumstances, it suffices that a person is not able to sign up for multiple accounts *at (roughly) the same time.* For example, consider a live online sports broadcast. At the end of the match, we would like anyone to be able to cast a vote for who is the player of the match, within a short time window, after which the winner will be announced. We allow viewers to sign up to vote immediately after the match – but we want each person to be able to sign up only once.

Desiderata. Here, we want to design a test that, in (say) 2 minutes, anyone can pass once, but nobody can pass twice.

Strategy. Rather than memory, we will focus on attention. Again, instances of a test will be randomized; they will take a fixed amount of time during which the subject is required to pay close attention. The test is designed in such a way that switching attention back and forth between two instances of the test is doomed to result in failing both instances (or at least one of them), so that every person can pass only one instance of the test in the amount of time given.

Cognitive weaknesses required. Humans are generally limited in the extent to which they can pay attention to two things at the same time (Eriksen and Yeh 1985; McCormick and Klein 1990; Pan and Eriksen 1993; McCormick, Klein, and Johnston 1998; Jans, Peters, and De Weerd 2010).

Existing implementations and experiments. In earlier work (Andersen and Conitzer 2016), we developed a detailed approach based on visual attention. The subject is presented with a screen on which boxes containing words move around. The user is supposed to pay attention to only one

[2] Of course, in some cases the provider is not trusted, perhaps because it is subject to a legal environment that is not trusted. I offer no solution for that here.

box, indicated at the beginning but otherwise indistinguishable from the other boxes, so that the user has to continue to track the box. There are words in these boxes that change over time. When the word changes in the box of interest, the user is supposed to indicate whether the word is misspelled or not, by pressing a button. Figure 1 shows a screenshot.

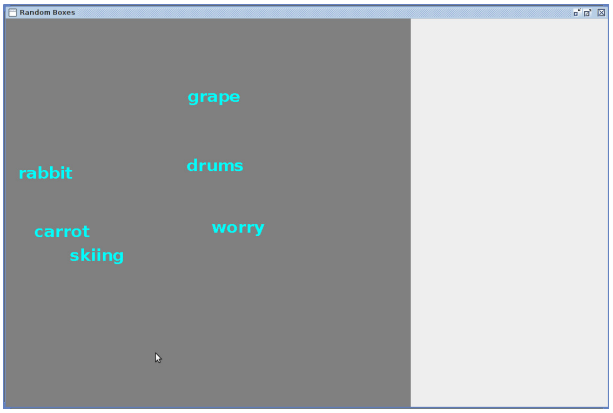

Figure 1: Screenshot, taken from (Andersen and Conitzer 2016). Each word is in an (invisible) box, and these boxes move around on the left side of the screen. The user has been asked to pay attention to one of these boxes initially, but there is no remaining visual indication of which box that is. The right side of the screen lights up in a color, green or red, to indicate whether a response is correct.

Under reasonable conditions on the motion of the boxes and the frequency of word changes, the test is quite doable for human subjects. However, it is effectively impossible for a subject to pass two instances of this test *at the same time*: switching attention to the other instance causes one to lose track of the correct box in the first instance. Indeed, in an experiment, when given a single instance, every subject (out of 25) passed the test; when asked to attempt to simultaneously pass two instances of the test, no subject passed more than one of them.

Remote Identification That Is Not Transferable Or Biometric

Motivation. Imagine a group of people – say, faculty in a department – who are in the same location, but know that soon they will not be. They will need to weigh in on important decisions remotely. There is a concern that some faculty will be too busy to pay attention and will hand off their vote to someone else, by giving that person the credentials to log in. (In other, longer-term settings, we may even worry about someone passing on their credentials to someone else before dying or becoming incapacitated.) We assume that doing so is not appropriate because the new person does not have the original person's expertise.

Desiderata. We want to enable a set of people to log in, possibly anonymously, without being able to transfer this ability to another person. It should not be possible for another person to learn to log in successfully from repeated attempts. We do not wish to rely on biometric identification (e.g., face recognition through a device's camera). We assume we initially have trusted access to the people in question (the faculty are still on campus).

Strategy. We design a video game (possibly in an automated, randomized fashion) with levels of increasing difficulty. Anyone should be able to learn to play level 10, by starting at level 1 and gradually working her way up. However, level 10 should be sufficiently difficult that it is not possible to learn to play level 10 *just by playing level 10*. That way, we can give faculty access to all levels of the game, to learn to play it, while they are still on campus (say, in a trusted room); later, upon attempting to log in from a remote location, they will be asked to play level 10 and perform reasonably well. They will not be able to hand off this ability to someone else – they cannot simply tell someone how to play a video game and have them instantly be good at it – nor will anyone else be able to learn how to log in by attempting to do so many times, because by assumption level 10 is too hard to learn to play without access to the easier levels.

Cognitive weaknesses required. Humans are generally unable to communicate to another person exactly how they do something like playing a video game well. Moreover, they cannot learn certain challenging tasks without first having the opportunity to practice on easier versions of them.

Existing implementations and experiments. To my knowledge, there are none yet.

Tools to Break the Properties: Notes, Friends, and AI

The human subjects experiments so far have been conducted under very restricted conditions. In reality, people would be able to use all kinds of different strategies. For example, they might take notes or screenshots, or they might get their friends to help. While more work would need to be done to ensure robustness to such attacks, they do not seem insurmountable. In particular, we can run the tests fast enough to make note-taking ineffective, and friends helping is not necessarily a concern if those friends are anyway allowed to obtain their own accounts. However, the use of AI creates a serious challenge. Even not very sophisticated computer programs could easily track which images have been seen before, and modern AI would be able to recognize faces even across different images. Similarly, writing a program to track a text box is likely not that challenging of a computer vision exercise. And reinforcement learning techniques have proved quite successful in video games (Mnih et al. 2013).

So, in reality, at least for sufficiently high-stakes applications, it is necessary to modify the techniques to make AI ineffective. This could involve integrating a CAPTCHA (von Ahn et al. 2003; von Ahn, Blum, and Langford 2004) into the task. More specifically, techniques for creating adversarial examples (Szegedy et al. 2014) in computer vision could be of help: perhaps it is possible to change a few pixels in a face, a box, or a video game that a human would not even notice but that would confuse the AI. This would be an appealing and constructive use of adversarial examples.

Such anti-AI modifications would still need to be

tested. There is already a significant literature on breaking CAPTCHAs; early examples include (Mori and Malik 2003; Thayananthan et al. 2003; Moy et al. 2004), and research on this topic continues (Ye et al. 2018). In this context, however, we would need the technique to be robust to human-AI teams as well – e.g., even a human aided by AI should not be able to obtain two accounts. How to appropriately design human subjects experiments to evaluate this is a challenging but intriguing question. One method would be to let researchers "attack" the system by developing auxiliary AI techniques and demonstrating in human subjects experiments that their use breaks the desired properties.

Concluding Remarks

Designing new systems based on human cognitive limitations is an exciting new area for interdisciplinary research. None of the examples described in this paper are deployed in practice so far, and we should set a high bar for doing so. The technique should work well not only in controlled environments with unaided human subjects, but also in the face of attackers willing to devote significant resources, at least in high-stakes applications. We should also address the fact that some of these tests will not be accessible to a subset of the population; indeed, none of the discussed designs are accessible to visually impaired people. This is an important concern. While it need not halt the research agenda, which is still in early stages, continued awareness of this issue and responsible conduct in this research generally are in order. The stakes of the application are also important: there is an obvious difference between excluding a visually impaired person from voting for player of the game after a sports broadcast, and excluding her from important faculty decisions.

There are, in my opinion, likely many other examples that fit within the general framework beyond the three presented here. It is intriguing that all three examples involve identity and authentication; are there other examples that do not involve these?[3] An appealing approach to finding new examples is for computer scientists and cognitive scientists to interact, with the former proposing motivating examples and the latter proposing useful cognitive limitations, working together from that point on to design and test a technique.

Acknowledgments

I thank Garrett Andersen, Eric Hu, and Kobi Gal for contributions to and feedback on our earlier work along these lines, and NSF for support under award IIS-1814056.

References

Andersen, G., and Conitzer, V. 2016. ATUCAPTS: Automated tests that a user cannot pass twice simultaneously. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*, 3662–3669.

[3]We are currently also doing research on *designing the identities of AI agents* (Conitzer 2019). Since the techniques discussed in this paper exploit *human* cognitive limitations, they will likely not work for AI; but should we design AI systems in such a way that they can identify themselves in similar ways, and if so, how?

Conitzer, V. 2010. Using a memory test to limit a user to one account. In Ketter, W.; Poutré, H. L.; Sadeh, N.; Shehory, O.; and Walsh, W., eds., *Agent-Mediated Electronic Commerce and Trading Agent Design and Analysis*, vol. 44 of *Lecture Notes in Business Information Processing*. 60–72.

Conitzer, V. 2019. Designing preferences, beliefs, and identities for artificial intelligence. In *AAAI*, 9755–9759.

Eriksen, C. W., and Yeh, Y.-y. 1985. Allocation of attention in the visual field. *Journal of Experimental Psychology: Human Perception and Performance* 11(5):583–97.

Jans, B.; Peters, J. C.; and De Weerd, P. 2010. Visual spatial attention to multiple locations at once: the jury is still out. *Psychological Review* 117(2):637–84.

McCormick, P. A., and Klein, R. 1990. The spatial distribution of attention during covert visual orienting. *Acta Psychologica* 75(3):225–242.

McCormick, P. A.; Klein, R. M.; and Johnston, S. 1998. Splitting versus sharing focal attention: Comment on Castiello and Umiltà (1992). *Journal of Experimental Psychology: Human Perception and Performance* 24(1):350–7.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. 2013. Playing Atari with deep reinforcement learning. arxiv.org/abs/1312.5602.

Mori, G., and Malik, J. 2003. Recognizing objects in adversarial clutter: Breaking a visual CAPTCHA. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 134–141.

Moy, G.; Jones, N.; Harkless, C.; and Potter, R. 2004. Distortion estimation techniques in solving visual CAPTCHAs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 23–28.

Pan, K., and Eriksen, C. W. 1993. Attentional distribution in the visual field during same-different judgments as assessed by response competition. *Perception & Psychophysics* 53(2):134–144.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. arxiv.org/abs/1312.6199.

Thayananthan, A.; Stenger, B.; Torr, P. H. S.; and Cipolla, R. 2003. Shape context and chamfer matching in cluttered scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 127–133.

von Ahn, L.; Blum, M.; Hopper, N.; and Langford, J. 2003. CAPTCHA: Using hard AI problems for security. In *Advances in Cryptology - EUROCRYPT 2003, International Conference on the Theory and Applications of Cryptographic Techniques*, 294–311.

von Ahn, L.; Blum, M.; and Langford, J. 2004. Telling humans and computers apart automatically: How lazy cryptographers do AI. *Communications of the ACM* 47(2):56–60.

Ye, G.; Tang, Z.; Fang, D.; Zhu, Z.; Feng, Y.; Xu, P.; Chen, X.; and Wang, Z. 2018. Yet another text captcha solver: A generative adversarial network based approach. In *ACM SIGSAC Conference on Computer and Communications Security (CCS'18)*, 332–348.