On Surprise-Hacking

Teppo Felin
University of Oxford
teppo.felin@sbs.ox.ac.uk

Mia Felin
University College London

Joachim I. Krueger
Brown University

Jan Koenderink
KU Leuven and Utrecht University

Forthcoming in
Perception

> Disbelief is not an option. The results are not made up, nor are they statistical flukes.
>
> Daniel Kahneman (2011)
> on perceptual priming research

Many perception experiments—like stage magic—engage in what might be called "surprise-hacking." Like magicians, investigators first divert the attention of experimental subjects with some kind of task, priming stimulus, or competing cue. They then go on to point out how humans are blind to things that should be obvious. This type of surprise-hacking has allowed cognitive scientists to claim that there is an epidemic of human perceptual blindness, irrationality, and delusion.

The famous "gorilla" experiment provides an excellent example of magic-like surprise hacking (Simons and Chabris, 1999). Participants were asked to view a video clip of individuals passing a basketball and to count the number of passes. As participants were attending to this task, a person in a gorilla suit walked across the scene. Participants were then asked whether they saw anything unusual. Only half of them reported that they had noticed the gorilla. The authors infer the scientific value of this finding from its surprise value. Many academic commentators have endorsed this interpretation (e.g., Kahneman, 2011). Similar research claims to show evidence for many other forms of blindness and illusion (e.g., O'Regan et al., 1999; Rensink et al., 1997; Simons and Levin, 1997; cf. Kuhn et al., 2008). Importantly, the leveraging of surprise is central to the preferred interpretation in many paradigms of perceptual and social priming. Together, these lines of research have given rise to a meta-theory claiming that the human mind can best be understood—or understood only—in terms of its delusions and deficits (Bargh and Chartrand, 1999; Bargh, 2008; Chater, 2018; Kahneman, 2011). We suggest that this meta-theory misrepresents the fundamental properties of mind.

We define surprise-hacking as the planning, design, and staging of experiments by which researchers seek to elicit counterintuitive, and typically, negative effects. Surprise-hacking is evident in the types of experiments that scientists craft and the conclusions they draw from the data. For example, if experimenters—*ex ante*—assume that humans are blind, they can then stage perceptual experiments by diverting attention, thus confirming their assumption and interpreting the data accordingly. Mimicking magic, much psychological research places an epistemic premium on surprise, and particularly, errors and irrationality (Krueger and Funder, 2004; Lamont and Wiseman, 2005; Sharpe, 1988).

The goals of stage magic are well-aligned with crafting psychological experiments to provide evidence of human blindness. The key tool of magic is misdirection, defined "as the intentional deflection of attention for the purpose of disguise" (Sharpe, 1988: 47). Like a Swiss army knife, magic offers cognitive scientists with an arsenal of useful tools, experimental treatments and interventions, to mislead respondents and to conjure the desired effect. In fact, some cognitive scientists have assembled a "taxonomy of misdirection" (Kuhn et al., 2014; cf. Macknick et al., 2008; Rensink and Kuhn, 2015). Different forms of misdirection—whether subtle and passive or overt and active—include distraction with irrelevant stimuli or tasks, subtle behavioral cues (e.g., pointing), delay and memory, varied types of priming and (incidental) manipulation of situations and ambient environments, misattribution and automaticity, verbal and non-verbal suggestion, and so forth.

The problem with surprise-hacking arises from the amount of time that goes into the preparation and staging of the experimental effect or illusion. This creates a radical mismatch as research participants have only a few moments to detect the purpose of a study. The asymmetry between the careful staging of the experiment, and the limited time available for task performance by subjects, stacks the odds in favor of the experimenter. The typical visual or experimental scene is teeming with any number of potential things that a subject might attend to, be aware of, or point toward. It is this

visual abundance that provides scientists with the ingredients from which to create distractions and diversions to conjure blindness.

However, there is an important difference between magic and staged perception studies. James 'The Amazing' Randi put it this way: "Magicians are the most honest people in the world; they tell you they're gonna fool you, and then they do it." At a magic show, one expects to be deceived and misdirected. This can't be said of scientific experiments that hack for surprise. Participants tend to trust the experimenters' instructions, which makes them easy prey. Given the implicit trust (or perhaps gullibility; cf. Forgas and Baumeister, 2019), experimental participants can readily be deceived. Participants may also wish to be "good subjects" for the experimenter, to ensure they don't ruin a sought-after effect (Orne, 1962; cf. Klein et al., 2012).

An essential question is this: should participants take the instructions they receive seriously, trusting the experimenter, and engage in the task intended to divert their attention? In order to *not* be blind (as blindness is conceived of by the experimenter), they shouldn't. Ironically, in the gorilla study, the data for those who did *not* properly engage in the diversion task (e.g., if they lost count of the passes) were *excluded* from the analysis (see Simon and Chabris, 1999: 1068). The exclusion of these data provides a window into the experimenters' mindset and what they were seeking. Some of the excused participants must have figured out that they were deliberately being distracted, and therefore lost count or didn't get the "right" answer—but did see the gorilla. Yet, missing the gorilla is repeatedly called an "error," "illusion," and "mistake"—a key flaw of the human mind (Chabris and Simons, 2010).

Our central concern is that surprise-hacked findings may *depend* on the clever and subtle staging and crafting to get the surface-level, magic-like "ta-da!" effect. Daniel Kahneman seems to implicitly recognize this. In response to criticisms of surprising priming effects, and the "ease with which these effects can be undermined," he notes that "priming effects are subtle and their design requires high-level skills" (2012: 1). However, if the effects indeed are this subtle and fragile—and crafted like magic—then do they actually tell us anything *fundamental* about perception and the mind? The same holds for gorilla-type studies, where any number of small changes to (or mistakes in) the experimental design and execution would, in effect, "give away" the trick.[1]

Of course, scientists performing these studies argue that they are saying something important about the mind. They claim to be providing evidence of the key determinants of perception, judgment and behavior, and even providing evidence of the illusory nature of human agency and will (Bargh, 2008; Bargh and Chartrand, 1999). They claim to show how respondents can be subliminally manipulated and influenced—by using various perceptual tools: competing cues, masked primes, irrelevant stimuli—to see things in a certain way or to behave irrationally. Like magic, people appear to be controlled "through external means"—"and thereby [scientists] 'bypass the will' entirely" (Bargh and Chartrand, 1999: 465; Bargh, 2008). As put by Bargh and Chartrand, "the entire environment-perception-behavior sequence is automatic, with no role played by conscious choice in producing the behavior" (1999: 466).

A simpler explanation of these findings is that they capture a mix of demand characteristics and experimenter bias (cf. Orne, 1962). In natural environments, humans might attend to any number of different stimuli or cues, and thus aren't merely "passive receptacles of stimuli" (cf. Klein et al., 2012; also see Felin, Koenderink and Krueger, 2017).

The tools of magic might also be creating an illusion for scientists themselves, the illusion of saying something important. Kahneman, for example, suggests that "the gorilla study illustrates two important facts about our minds: we can be blind to the obvious, and we are also blind to our

[1] The practice of surprise-hacking can be seen as the upstream, theoretical cousin of p-hacking or data dredging. Surprise-hacking has to do with the set of *a priori* theoretical assumptions and precommitments that scientists have and make, which reveal themselves in the type of experiments they design and construct (in terms of what they are *looking for*). Surprise-hacking cannot be solved with more data or further replications. Rather, surprise-hacking will be solved with better theories that guide us toward more fundamental questions, rather than merely hacking for ephemera.

blindness" (2011: 23-24). But what exactly is obviousness (or salience) from the perspective of perception? Kahneman argues that obviousness is driven by the inherent nature of stimuli. That is, salience (or "accessibility") is driven by the "actual properties of the object of judgment" (2003: 700), actual stimulus properties like size. Thus the gorilla should be obvious.

But if a theory assumes that humans should see something (because it has certain characteristics), but then they don't, it is the theory that needs to be revised. An appeal to blindness is not a scientific explanation, but a mere re-labeling of the problem. Instead of giving these effects a name—x type of blindness (change blindness, inattentional blindness), y type of bias, or z type of visual illusion—a more productive approach is to provide the *reasons why* people see certain things in certain ways and the psychological processes that make it so.[2]

Like a rabbit pulled out of a hat, findings generated through surprise-hacking can be explained in non-magical terms. The very studies used to provide evidence of blindness can be discussed without resorting to magic-like surprise. Recall that Simons and Chabris (1999) argue that people should see the gorilla because it is a *large* object, but many don't see it because it is *unexpected*. This introduces two conflicting views of perception: one where perception is a function of object or stimulus characteristics and the environment, and one where perception is a function of expectations. Their focus is on blindness to the gorilla (or more generally: "large changes to objects and scenes"), but also on the fact that "observers fail to see an unexpected object" (1999: 1060). It's hard to have it both ways. Awareness remains a black box, and simply an artefact of the conjuring and experimental staging which ensures the sought-after effect. This is most evident from the fact that it is the *negative* version of the gorilla finding that gets the bulk of the attention. It is this version that is generalized and popularized by the authors in their bestselling book *The Invisible Gorilla*. They argue that the gorilla finding is a specific example of something far more general, namely, that "virtually no realm of human behavior is untouched by everyday illusions" (Simons and Chabris, 2010: 10). The book then goes on to discuss varied forms of blindness, illusion and—as the subtitle of the book says—"other ways our intuitions deceive us." The focus is rarely on the generativity, accuracy or even the usefulness of human perception. This illustrates how surprise-hacking has led cognitive psychology astray, in terms of our understanding of perception and the human mind.

An alternative, simpler, and biologically more compelling, interpretation of the surprising, magic-like findings of blindness is that perception is active, rather than passive, and shaped by problems, expectations, and questions (cf. Felin, Koenderink and Krueger, 2017). Whether researchers or research participants, we largely see what we are looking for. In staged experimental settings, scientists can easily shape (or misdirect) this *looking for* activity. But this is no basis for concluding that humans are blind. A radically different view of perception is possible, one that is mind-to-world rather than world-to-mind oriented. This model suggests that there is no perceptual obviousness—as determined by the nature of stimuli or objects (cf. Kahneman, 2003)—without an understanding of organism-specific factors, and questions and problems. An organism's *Suchbild* (search or seek image) directs awareness to certain features or things in its environment (von Uexküll, 2010), at the exclusion of an indefinite number of other things that also might be obvious in a visual scene. This exclusion of stimuli isn't a problem to be solved. After all, we can't and wouldn't want to see everything! Thus staged experimental studies of perception are essentially just fiddling with the human *Suchbild* and its perceptual interface (for further discussion see Chater et al., 2018; cf. Hoffman et al., 2015). When they succeed, the audience's surprise is hacked. But this surprise-hacking merely deceives scientists and audiences into believing that something fundamental about perception and the mind has been revealed.

[2] This problem is similar to current efforts to identify, categorize and name various types of visual illusions (Shapiro and Todorovic, 2017). Often such studies are nothing more than attempts to "catch" participants unawares. But upon closer examination, most of these purported illusions turn out not to be illusory at all (Rogers, 2014). Like magic, they can be explained. Of course, the narrative of blindness and illusion sells, and therefore continues to be the central thesis of popular books written by psychologists and cognitive scientists (e.g., Chater, 2018).

REFERENCES

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, *54*(7), 462.

Bargh, J. A., (2008). Free will is un-natural. In J. Baer, J. C. Kaufman, & R. F. Baumeister (Eds.), Are we free? Psychology and free will. Oxford, UK: Oxford University Press.

Chabris, C., & Simons, D. (2010). *The Invisible Gorilla: And Other Ways Our Intuitions Deceive Us*. Harmony.

Chater, N., et al. (2018). Mind, rationality and cognition: An interdisciplinary debate. *Psychonomic Bulletin & Review*, *25*(2), 793-826.

Chater, N. (2018). *The Mind is Flat: The Illusion of Mental Depth and the Improvised Mind*. Penguin UK.

Felin, T., J. Koenderink & J. I. Krueger. (2017). Rationality, perception, and the all-seeing eye. *Psychonomic Bulletin & Review 24*(4) 1040–1059.

Forgas, J. P. & Baumeister, R. F. (2019). *Homo Credulus: The Social Psychology of Gullibility*. New York, NY: Psychology Press.

Hoffman, D. D., Singh, M., & Prakash, C. (2015). The interface theory of perception. *Psychonomic Bulletin & Review*, *22*(6), 1480-1506.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*(9), 697.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kahneman, D. (2012). Open letter: A proposal to deal with questions about priming effects. *Nature*.

Klein, O., Doyen, S., Leys, C., Magalhães de Saldanha da Gama, P. A., Miller, S., Questienne, L., & Cleeremans, A. (2012). Low hopes, high expectations: Expectancy effects and the replicability of behavioral experiments. *Perspectives on Psychological Science*, *7*(6), 572-584.

Krueger, J. I., & Funder, D. C. (2004). Social psychology: A field in search of a center. *Behavioral and Brain Sciences*, *27*(3), 361-367.

Kuhn, G., Amlani, A. A., & Rensink, R. A. (2008). Towards a science of magic. *Trends in Cognitive Sciences*, *12*(9), 349-354.

Kuhn, G., Caffaratti, H. A., Teszka, R., & Rensink, R. A. (2014). A psychologically-based taxonomy of misdirection. *Frontiers in Psychology*, *5*, 1392.

Lamont, P., & Wiseman, R. (2005). *Magic in theory: An introduction to the theoretical and psychological elements of conjuring*. University of Hertfordshire Press.

Macknik, S. L., King, M., Randi, J., Robbins, A., Thompson, J., & Martinez-Conde, S. (2008). Attention and awareness in stage magic: turning tricks into research. *Nature Reviews Neuroscience*, *9*(11), 871.

O'Regan, J. K., Rensink, R. A., & Clark, J. J. (1999). Change-blindness as a result of 'mudsplashes'. *Nature*, *398*(6722), 34.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*(11), 776.

Rensink, R. A., & Kuhn, G. (2015). A framework for using magic to study the mind. *Frontiers in Psychology*, *5*, 1508.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, *8*(5), 368-373.

Rogers, B. (2014). Delusions about illusions. *Perception*, *43*(9), 840-845.

Sharpe, S.H. (1988). *Conjurer's Psychological Secrets*. Calgary: Hades Publications.

Shapiro, A. G., & Todorovic, D. (Eds.). (2017). *The Oxford Compendium of Visual Illusions*. Oxford University Press.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, *28*(9), 1059-1074.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences*, *1*(7), 261-267.

Von Uexküll, J. (2010). *A foray into the worlds of animals and humans*. University of Minnesota Press.