Tigers, black swans, and unicorns: The need for feedback and oversight

Max M. Houck

Forensic & Intelligence Services, Saint Petersburg, FL, 33701, United States

**Abstract**

Humans have a decision-making system which is biased to avoid costly false negatives while the criminal justice system is designed to be biased in the opposite way, avoiding costly false positives. But systems fail, people do not; a badly out of kilter system can lead even the most expert to bad outcomes. Perverse incentives, driven by the fetishizing of DNA, put pressure on an already-stressed forensic system. Every system needs feedback, both positive and negative, to correct itself and stay stable, forensic science is only one of those in a criminal justice system. Recognizing false positives, false negatives, and how they happen is critical to stabilizing and calibrating a criminal justice system. Oversight, review, and addressing wrongful convictions is a necessary form of feedback to forensic science and any balanced and fair criminal justice system.

© 2019 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

**Keywords:** False positives, False negatives, DNA, Misconduct, Oversight

1. Tigers and wrongful convictions

Anyone who studies criminal justice is familiar with the notion “that it is better 100 guilty Persons should escape than that one innocent Person should suffer” [2]. This “long and generally approved” statement is the inflated version of the original sentiment by Voltaire that “tis much more Prudence to acquit two Persons, tho’ actually guilty, than to pass Sentence of Condemnation on one that is virtuous and innocent” [3; p. 53] which was then increased to ten guilty persons by Sir William Blackstone (“For the law holds, that it is better that ten guilty persons escape, than that one innocent suffer” [4; p. 358]), which must have been eventually read by Franklin who expanded it by a factor of 10. Forensic science, as a part of the criminal justice system and the overarching idea of the rule of law, also holds to the ideals of accuracy in justice. Science labels these as errors of Type I, false positives, and Type II, false negatives. Therefore, it is common to consider Type II errors (letting the guilty slip through the justice system unscathed) as having a lower moral cost for society and upholding the public’s concepts of justice than Type I errors (convicting and incarcerating someone who is innocent of the crime they are accused of) (Table 1.

For the roughly two million years of human evolution, however, it has worked the other way. Hearing the grass rustling in the savanna should force one to conclude that the noise comes from a predator and not a friend. The cost of mistaking a tiger sneaking up on you for a friend playing a practical joke (false negative) is fatally expensive; the reverse (false positive) likely results only in laughter and a good story later on (see Table 2).

If the cost or significance of a false positive greatly outweighs that of a false negative, as in the criminal justice system, then it makes sense to design the system to reduce the number of false positive outcomes. Likewise, if a false negative is much more costly, as with being eaten by a tiger, the decision system should reduce the number of very (or fatally) costly outcomes. This is true even if biasing the system to reduce costly false negatives results in a relatively low-cost event being detected when it is not happening. A modern example of biasing a decision system to prevent high-cost false negatives is airport security screening, where the system is designed to prevent high-cost events (bombings, shootings, crashes, etc.) even if it means routine low-cost inconveniences for innocent travelers. Humans, in fact, developed cognitive biases for just these reasons: the solutions have had a recurring, significant effect on survival and evolutionary fitness (tigers) and the costs a false positive or false negative greatly outweighs the cost of the alternative error (tigers or wrongful convictions) [5]. And our systems, be they social or algorithmic [6], reflect these biases.
Table 1
The societal costs of false positives are seen as higher than those of false negatives.

| The person is determined to be … | But they actually are … |
|----------------------------------|------------------------|
| Not Guilty                      | Guilty                 |
| True positive                   | False negative (II) (Low cost) |
| False positive (I)              | True negative (High cost) |

Table 2
The evolutionary costs of false negatives are much higher than false positives.

| You think the source of the rustling is a … | But it is actually a … |
|---------------------------------------------|------------------------|
| Tiger                                       | Friend                 |
| True positive                               | False positive (I)     |
| False negative (II)                         | True negative (High cost) (Low cost) |

The hitch comes when you have a decision-making system which is biased to avoid costly false negatives (the human brain) working in a decision-making system which is biased in the opposite way to avoid costly false positives (criminal justice). Systems are methods used to solve problems and achieve desired results. If those results are not achieved, the failure is due to inadequacies in the method or the system, but people are usually targeted as the problem:

“People are not the same as organizational systems. They work in systems, but the systems existed before most of the people were hired and will continue after the current employees are gone … When we don’t understand systems, we equate improving our people with improving our systems. [1; page 23]

Desired results are not achieved, people are blamed, and typically training ensues (after removing the individuals at “fault”) rather than repairing the deficiencies in the system. Training often does not work, if for no other reason than it is not possible to have a course or module for every possible adverse event. And one course cannot fix systemic problems that have a long operational history; it could even make it worse by giving the trainee the illusion of being “repaired” and thinking their behavior is now superior to what it actually is. Low-ability people cannot recognize their lack of ability (the Dunning-Kruger effect) and, consequently, cannot objectively evaluate their actual competence or incompetence [7]. On the high end of competence, increased information does not contribute to increased accuracy, just increased confidence in the answer already decided [8]. Training of alleged poor performers and the assumed competence of confident experts cannot fix an out-of-kilter system.

2. Black swans and DNA typing

No one who has paid any attention to forensic science since the late 1980s could argue that DNA typing did not completely change the face of the profession and its attendant systems. The application of DNA typing to forensic science was a surprise; people laughed at Sir Alec Jeffreys when he first suggested its use in criminal cases [9]. And, in hindsight, it has been treated as a magical creature, a unicorn, that will cure all forensic problems [10], when, in fact, it is no better than any other method employed in the discipline. Given these three factors—it came as a surprise, had a major effect, and was rationalized to unicorn status—forensic DNA could be considered what Nicholas Taleb has called a black swan [11]. Taleb’s black swan is not the problem of induction discussed in beginning philosophy classes but rather describes unexpected events of significant magnitude and effect that play an outsized role in history. Black swans often have unpredictable, unforeseen consequences and, because of their unpredictability, are distinct and memorable, making them a strong anchor for numerous cognitive biases, both individually and collectively [12]. People remember the strange and exceptional in favor of the mundane and normal, and this skews how they perceive current and future events. If DNA helped to convict in a particularly important or heinous case or exonerated a person long incarcerated, then it was the savior of that case and will be relied upon in the future. The number of cases where it didn’t work or produce desired results (“misses”) will be forgotten in favor of the “wins” (a function of confirmation bias [12]). Prosecutors are wary of bringing cases to trial without DNA; they worry that the juries expect it in every case [13]. Although the U.S. National Academies of Science [NAS] lauded forensic DNA typing as the “gold standard” in forensic science in 2009 [14], the method had been promoted long before from an unexpected black swan to an exceptional, magical unicorn.

3. Unicorns, poor incentives

Forensic DNA typing has helped to exonerate the innocent (in fact, its first application was an exclusion of an innocent person) and to help bring the guilty to justice. The power of DNA to help convict and acquit, as well as its origins outside the historical confines of police-run forensic science, lead to the “gold standard” appellation. DNA went from being just another type of evidence, one that only answers the question “who,” to being seen as a “truth machine” surrounded by “the rhetoric of infallibility” [15]. From black swan to a magical unicorn, DNA became fetishized as the pinnacle of evidence, taking on powers and meaning beyond its storage and replicatory capacities. DNA went from a simple molecule to the maker or breaker of cases, its meaning and relationship to cases and their success changing much as if a pawn had suddenly become a queen on a chessboard [16, page 88]. The molecule had not changed but its place in the forensic pantheon of evidence surely had, altering the forensic enterprise forever. DNA The Molecule had been replaced by DNA The Unicorn, fetishized through the “common process whereby inanimate, material things get drawn into social relations between people and take on a life of their own” [11, page 140].

Success breeds success. Police and prosecutors pushed for more and more DNA samples to be analyzed, fueling and fueled by the aforementioned CSI Effect. But increased pressure to produce means more chances for a false positive. With increased pressure and because each case is different, cases are rarely re-worked and a lack of replication means any mistakes likely will not be detected; “false discoveries are usually not identifiable at a glance, which is why they are problematic. In some cases, poor or absent theory amounts to hypotheses being generated almost at random, which substantially lowers the probability that a positive result represents a real effect” [17].

Prosecutorial pressure can produce perverse priorities. Systemic and persistent problems with scientific conduct are more caused by non-scientific incentives than simple mistakes or explainable misunderstandings. Pressure to produce, like operational pressures to “get the bad guy,” can lead to an environment where methods prioritize output over scientific rigor. As Smaldino and Edreath discuss about the “publish/perish” pressures in academic science:

An incentive structure that rewards publication quantity will, in the absence of countervailing forces, select for methods that produce the greatest number of publishable results. This, in
turn, will lead to the natural selection of poor methods and increasingly high false discovery rates [17].

The evidential and statistical power of DNA has been pushed close to its limits, with smaller sample sizes and striving to make the most of often low-quality forensic samples. This carries an additional danger, however, of shifting the power (1−β) of the analysis:

It is important to understand why increasing power tends to also increase false positives without the application of effort. It is quite easy to produce a method with very high power: simply declare support for every proposed association, and you are guaranteed to never mistakenly mark a true hypothesis as false. Of course, this method will also yield many false positives. One can decrease the rate of false positives by requiring stronger evidence to posit the existence of an effect. However, doing so will also decrease the power—because even true effects will sometimes generate weak or noisy signals—unless effort is exerted to increase the size and quality of one’s dataset.

Forensic science hasn’t had much choice in increasing the size and quality of its DNA dataset: The size and quality of biological samples for DNA analysis are adventitious. Detection and analysis methods have been driven by politically asymmetric demands—and scientists’ insatiable curiosity—to try to work with smaller and worse samples “to get the bad guy,” to make better and tastier chicken salad out of worse and fewer chicken droppings, to paraphrase a colloquialism. The danger was realized with the recent publication of a NIST study which demonstrated that DNA laboratories around the U.S. reached different conclusions for mixture interpretation from the sample data sets [18]. Even scientists from the same laboratory using the same mixture interpretation protocol calculated different outcomes. The study was designed to see what variation existed in forensic DNA interpretation. Not surprisingly, it was quite a bit. Most of the methods used were manual binary methods, like Combined Probability of Inclusion (CPI). Without additional danger, however, of shifting the power (1−β) of the analysis:

Rational discourse is predicated on having data to use that supports any assertions made; data that shows variation invites more vigorous discourse. Discourse without facts leads to speculation, guesses, and opinion. “Where opinion prevails, whoever has power is king,” notes Scholte, and “[t]he ultimate correlation, therefore, is more likely between assertiveness and clout, not assertiveness and objective truth. It is possible that managers who wish to hold on to the illusion of power may resist a statistical view of work.” [1, page 24] Resistance doesn’t need to be outright obstruction; it could take the form of influence, cajoling, or even pressure. Maintaining the status quo is generally beneficial for those in power and “good enough” typically means for the status quo’s needs. Phrases like “fit for purpose” begs the question of whose purpose something is fit for. The moral imperative to get the right answer, to reduce false positives and false negatives, runs through any criminal justice system: The goal of justice is to right wrongs and restore order. But holding science accountable to legal incentives, like “successful” convictions, deflects science’s contribution to that moral imperative. Removing the incentives has a better chance of fixing the problems than does adding more rules [1].

5. What price integrity?

DNA has become even more of a fetish for law enforcement with the application of commercial genealogy databases [21,22] and rapid DNA systems in police stations [23,24], despite concerns with both applications. DNA technology may be slipping entirely from science’s control and oversight. Science certainly wants to use valid methodology and correct any previous errors, especially the kind of errors that led to a wrongful conviction. Saying that discovered, proven errors are only “good for historical purposes” [18] ignores the moral imperative that drives science’s support of justice and justice itself. Learning of past mistakes and questionable or invalid methods should prompt at least a retrospective evaluation of contested cases, as some states have done with forensic science commissions. Several states 1 have also adopted “junk science” or “changed science” statutes or rules that waive restrictions on appeals if it can be shown that the science used to obtain a conviction has changed or been found to be invalid. Virginia considered such a bill but decided the costs—almost $440,000 plus funding to hire independent experts to address forensic science challenges in court—were too high [25]. Considering that only one wrongful conviction case in 1997 cost the state and its city of Norfolk $8.4 million [26], it seems a worthwhile investment. It’s the coldest comfort for those wrongfully convicted to know that methods used to convict them may have been invalid and will no longer be used going forward. Conviction integrity units, like the one recently created in the State Attorney’s Office in Jacksonville, Florida, whose first case saw a four-decade-old conviction vacated [27], also act as feedback to a politically asymmetrical system.

Confidence alone is not enough to assume accuracy [8] especially when justice is at stake. Blaming individual scientists or managers does not correct the systemic issues facing forensic

1 California, Connecticut, Michigan, Texas, and Wyoming.
science and its use. Systems fail, people do not; a badly out of kilter system can lead even the most expert to bad outcomes. Every system needs feedback, both positive and negative, to correct itself and stay stable. Recognizing false positives, false negatives, and how they happen is critical to stabilizing and calibrating a criminal justice system. Oversight, review, and addressing wrongful convictions is a necessary form of feedback to forensic science and any balanced and fair criminal justice system.

Conflicts of interest

The author has no financial or personal interests to report. The author is the Editor in Chief of Forensic Science International: Synergy and has previously published with Elsevier. The opinions are solely those of the author.

Acknowledgments

The author has no financial or personal interests to report. The author is the Editor in Chief of Forensic Science International: Synergy and has previously published with Elsevier. The opinions are solely those of the author. The author would like to thank the anonymous colleagues who helped with early drafts of this work.

References

[1] P. Scholte, The Leader’s Handbook, McGraw-Hill, New York, 1998.
[2] B. Franklin, A letter to Benjamin Vaughan, march 14, 1785, in: A.H. Smyth (Ed.), (1906) the Writings of Benjamin Franklin, vol. 9, Macmillan, New York, 1785, p. 293.
[3] Voltaire, Zadig, reprinted 1974), Nizet, Paris, 1749.
[4] W. Blackstone, reprinted 1978, Commentaries on the Laws of England, ninth ed., vol. 4), Springer, New York, 1783.
[5] M.G. Haselton, D.M. Buss, Biases in social judgment: design flaws or design features?, PDF in: J.P. Forgas, K.D. Williams, W. von Hippel (Eds.), Social Judgments: Implicit and Explicit Processes Cambridge University Press, New York, 2003, pp. 23–42.
[6] W. Knight, Biased Algorithms Are Everywhere, and No One Seems to Care, MIT Technology Review, 2017, July 12.
[7] J. Kruger, D. Dunning, Unskilled and unaware of it: how difficulties in recognizing one’s own incompetence lead to inflated self-assessments, J. Personality Soc. Psychol. 77 (6) (1999) 1121.
[8] C. Hall, L. Ariss, A. Todorov, The illusion of knowledge: when more information reduces accuracy and increases confidence, Organ. Behav. Hum. Decis. Process. 103 (2007) 277–290.
[9] J. Cobain, Killer Breakthrough—The Day DNA Evidence First Nailed a Murder, 2016. The Guardian, July 7.
[10] W. Langley, The Case against DNA. The Telegraph, 2012. March 6.
[11] N. Taleb, The Black Swan: the Impact of the Highly Improbable, Random House, New York, 2007.
[12] M.E. Oswald, S. Grosjean, Confirmation bias. Cognitive Illusions: A Handbook On Fallacies And Biases In Thinking 79, judgment and memory, 2004.
[13] M.M. Houck, CSI: reality, Sci. Am. 295 (1) (2006) 84–89.
[14] National Research Council of the National Academies, Strengthening Forensic Science in the United States: A Path Forward, National Academies of Science, Washington, D.C, 2009.
[15] W. Thompson, The myth of infallibility (Chapter 15), in: S. Krimsky, J. Gruber (Eds.), Genetic Explanations: Sense and Nonsense, Harvard University Press, Cambridge, 2013.
[16] F.D. Sauassure, Course in General Linguistics, Trans. R. Harris, Duckworth, London, 1983.
[17] P.E. Smaldino, R. McElreath, The natural selection of bad science, Roy. Soc. Open Sci. 3 (9) (2016) 160384.
[18] J.S. Buckleton, J.A. Bright, K. Cheng, B. Budowle, M.D. Cole, NIST interlaboratory studies involving DNA mixtures (MIX13): a modern analysis, Forensic Sci. Int.: Genetics 37 (2018) 172–179.
[19] S. Miller, JoAnn Parks Was Convicted of Killing Her Kids in a Fire. Now Journalist Edward Humes Is Making the Case for Her Innocence, The Los Angeles Times, 2019. January 29.
[20] M. Godsey, Blind Injustice: A Former Prosecutor Exposes the Psychology and Politics of Wrongful Convictions, University of California Press, 2019.
[21] D.S. Court, Forensic genealogy: some serious concerns, Forensic Sci. Int.: Genetics 35 (2018) 203–204.
[22] E. Murphy, Forensic DNA typing. Annu. Rev. Criminal. 1 (2018) 497–515.
[23] A. Bennett, New ‘Rapid DNA’ machine causes some concerns. KHOU-11, February 14, 2019, https://www.khou.com/article/news/new-rapid-dna-machine-causes-some-concerns/285-622e3057-4f52-4e71-b6f5-cd767c8371bf, 2019.
[24] S. Yost, Sheriff Scrapping DNA Machine Draws Sharp Criticism, Rhino Times, March 26, 2019, http://www.rhinotimes.com/news/sheriff-scrapping-dna-machine-draws-sharp-criticism/.
[25] F. Green, House Subcommittee Kills “Junk Science” Legislation, Richmond Times-Dispatch, 2019, February 14, 2019, https://www.richmond.com/news/house-subcommittee-kills-junk-science-legislation/article_174292f1-8437-5e8e-b3fc-3e360a29215d.html.
[26] R. Ballico, Virginia Legislators Claimed a Bill to Allow Appeals from People Convicted by Junk Science Was Too Expensive, The Washington Post, 2019, February 19, 2019, https://www.washingtonpost.com/opinions/2019/02/19/virginia-legislators-claimed-bill-allow-appeals-people-convicted-by-junk-science-was-too-expensive/?utm_term=.cea5fe3eeedc.
[27] A. Pantazi, Jacksonville Men Freed 43 Years after Wrongful Murder Conviction, a First for a Florida Conviction Review Unit, The Florida Times-Union, 2019. March 28, 2019, https://www.jacksonville.com/news/20190328/jacksonville-men-freed-43-years-after-wrongful-murder-conviction-first-for-florida-conviction-review-unit.