5P - A/B testing practices have legal obligations that are not being met

Background

Government digital services repeatedly point to the way digital services allow 'iterative and continuous improvements', claiming this as something that should benefit everyone, especially the most vulnerable. But as analysed by Richard Pope[1] and others, these benefits have largely accrued to government itself and not to the public, some of whom have found themselves disadvantaged and even harmed – possibly unlawfully, and quite probably in breach of both the Civil Service Code and the Nolan Principles.[2]

Annex 5 is a series of examples of what could *and should* happen in digital across government. Of necessity, due to its lack of transparency, not every example provided will relate to Universal Credit, but due to the pain on all sides that comes from having to work in the Monster Factory, or having to deal with it in any other way – especially for people who rely on UC, and for those who support people who do – this is another note about UC.

A/B testing

Research by the Nudge Unit[3] found that people tend to answer forms more honestly if they are asked to sign a 'truthfulness statement' at the beginning of a form, rather than at the end.[4] BIT discovered this by sending different people different forms, and comparing the response rates. This process also works for sending people different letters,[5] and for a range of other interventions developed over the last decade. (We have picked examples from 2014-2015.)

The technical term for running such tests is "A/B testing", and the approach is often used when changing a digital service – a change being based on how a large number of users have actually used different variants. A/B tests are usually done on relatively small variations; Google is known to have tested *forty-one different shades* of blue on a page, to see which one got the most clicks.[6] And such tests don't necessarily need to make sense when seen from outside the organisation, they only need to make sense internally.

While DWP has stated on the one hand that "Universal Credit is an online digital service which has a release every two weeks",[7] DWP has also publicly confirmed that it does "zero-downtime" releases multiple times a week.[8]

---

[1] https://pt2.works/reports/universal-credit-digital-welfare
[2] Statements about AI and Public Standards may be replicated for A/B testing: https://www.gov.uk/government/publications/artificial-intelligence-and-public-standards-report
[3] https://www.gov.uk/government/organisations/behavioural-insights-team
[4] page 14, https://www.bi.team/wp-content/uploads/2015/07/BIT_FraudErrorDebt_accessible.pdf
[5] page 3, https://www.bi.team/wp-content/uploads/2015/07/BIT_FraudErrorDebt_accessible.pdf
[6] http://www.zeldman.com/2009/03/20/41-shades-of-blue/
[7] https://www.whatdotheyknow.com/request/601989/response/1447720/attach/html/2/IR2019%2035392%20Reply.pdf.html
[8] https://diginomica.com/how-dwp-managed-surge-demand-universal-credit-during-covid-19

We infer that the difference between the two is simple: ‘fortnightly’ ‘releases’ contain changes that DWP Digital sees a need to tell other DWP staff about. By implication, DWP does not deem it necessary to tell DWP’s own staff (or others) about these more frequent “minor releases”, which we presume include A/B tests (since DWP says they do them on a sub-fortnight schedule)[9]. Such changes *should* be below the level where DWP staff need any training or notification…

However, just because the DWP Digital team does not think it has to tell other DWP staff, does not mean that a change is not something that will have an adverse effect on at least some DWP claimants – for example, people who are on the Asperger’s spectrum. DWP pays charities and support organisations to provide materials to help claimants through the UC process, and then it changes that process in subtle ways underneath them – in ways which may indeed undermine the very support for which DWP (and the taxpayer) pays.[10]

If DWP conducts an A/B test that swaps the order of some questions, that test – even though only a small proportion of UC users will see it – will be different to the documentation and guidance provided, and will become a barrier to some users in ways which may be entirely invisible to DWP.

DWP’s initial instinct, when asked, was to claim “prevention of crime” as a reason not to talk about the tests that it runs.[11] This is worthy of scrutiny. Some tests may indeed be explicitly about the prevention of crime, and it seems superficially reasonable that information about such tests should be restricted. But, quite aside from the fact that not all tests will be for such purposes, DWP’s insistence that no-one can know what it does for ‘prevention of crime’ purposes does have another obvious flaw: refusing to answer questions hampers only those who respect the law.

Criminals attempting to defraud DWP will be able to see exactly what the questions are, because they will see all of the questions as they fill in the application forms – giving them a distinct informational advantage; DWP’s policy-based ‘comfort blanket’ having blinded everyone else...

Considering the example of being required to sign at the beginning of a form rather than at the end; such a change would have the effect of excluding support services from being able to fill in a form without submitting it, so that they can see and explain to the person they are supporting the specific questions they will be asked (and in which order, which can be relevant to some). Legally binding ‘statements of truth’ at the top of forms exclude users including those who DWP pays to offer support to some of the most vulnerable!

Discriminative effects can be subtle, which is why a policy of proper consideration – not a blanket assumption of no harm – is required.

To take an example from the financial industry: it may support the goals of a bank’s ‘growth team’[12] to tell people they will start earning interest ‘as soon as they press the start button’,

---

[9] DWP’s explanation is here https://youtu.be/0LDfMXP4Te0?t=4467
[10] https://medconfidential.org/wp-content/uploads/2021/03/5Q-AB-as-used.pdf
[11] https://www.whatdotheyknow.com/request/ab_testing_results_from_universa#incoming-1697604
[12] https://www.ft.com/content/be723754-501c-11e9-9c76-bf4a0ce37d49

and only later in the process ask them whether they actually *want* to do so. Such an approach is likely to increase response rates from the majority of people who like a return for seemingly nothing, but it would also reduce the response rate from people whose faiths prohibit receiving financial interest. The resulting raw data of any overall increase in response rate would not be sufficient to discover that the response rate amongst those with a protected characteristic had dropped to zero.

If an A/B test on the application process resulted in some people 'falling out' of the application process who were legally eligible for UC, DWP would not necessarily know enough about these people to know how to tell them that a failure on DWP's part meant they had not made it through the process. And if an A/B test resulted in a claim being closed, would DWP have the ability to reopen it – or is that a burden that would fall on the claimant?

Returning to the Google example from earlier, if a particular shade of blue causes a 2% improvement according to some measure, that could arise from a combined 3% increase from the general public and a 1% decrease. If that 1% share a particular (protected) characteristic, while it might be lawful for Google to decide it does not wish to service those people, it would not be lawful for DWP to do so. And to the point of this annex, *no-one would have any idea that such an exclusion had happened* – because DWP does not tell anyone about these types of experiment on UC applicants, 'successful' or otherwise.

DWP's past and current practice of secrecy is unlikely to be resilient to legal challenge. It appears inherently discriminatory in ways DWP will not detect and, by design, *cannot* detect. Even with the collection of the 13 data items identified by Dr Byrom for monitoring equality within the Justice system,[13] it would require greater care than DWP offers for every A/B test, and greater transparency than that of which UC seems capable.

A/B testing can be useful, but it must be used *lawfully* – which requires transparency of tests,[14] results, and consequences.

Facebook and other malign corporate actors use dark anti-patterns and A/B testing to measure how they manipulate users; User Research teams across Government claim that this is not how they do things. We might believe them, but the law requires evidence of one's actions.

---

[13] https://www.gov.uk/government/news/hmcts-publishes-response-to-report-on-use-of-data

[14] https://www.whatdotheyknow.com/request/ab_testing_results_from_universa and https://www.whatdotheyknow.com/request/digital_ab_testing_of_uc_in_2020 The original document from DWP was an .eml file mistakenly uploaded as a .doc file. We have decrypted it and made a PDF of the contents of the original document including a link to the original file here: https://medconfidential.org/wp-content/uploads/2021/01/2021-01-18-UC-FOI-2FA-analysis.pdf https://medconfidential.org/wp-content/uploads/2021/01/2021-01-18-UC-FOI-Confirm-Your-Identity-Phase-2.pdf https://medconfidential.org/wp-content/uploads/2021/01/2021-01-18-UC-FOI-Implementing-an-A_B-test.pdf

The Nolan principles entirely aside,[15] one GDS slogan is: “Make things open, it makes things *better*”. In the context of A/B testing, that slogan should rather be: “Make things open, it makes things *lawful*”. Were a court to ban A/B testing for, e.g. its discriminative effects, this could have a far wider impact on the evolution of digital services.

Towards legality: immediately available next steps

1) Regularly publish a list of (forthcoming[16]) tests
    - this will enable consideration of the proposed interactions on those who rely on support services (e.g. moving certain fields earlier in the process)

This will also allow some understanding, and preparation of:

2) Contingency plans for adverse events, including measures of exclusion
    - Are subgroups adversely affected by increased failures, even if the success rate nominally went up in a way which disguised those failures?

Additional steps will be required.

[15] The statements about AI may be replicated for A/B testing, https://www.gov.uk/government/publications/artificial-intelligence-and-public-standards-report

[16] Historic tests should be published too, but these are likely to be of less ongoing legal interest.